Reasoning Studies. From Single Norms to Individual Differences.

Dr. Dr. Niels Skovgaard-Olsen

November 2022

Kumulative Habilitationsschrift

zur Erlangung der *venia legendi*

im Fach Psychologie

Dr. Dr. Niels Skovgaard-Olsen,

Institut für Psychologie,

Abteilung Sozialpsychologie und Methodenlehre,

Wirtschafts- und Verhaltenswissenschaftliche Fakultät

Albert-Ludwigs-Universität Freiburg

Table of Contents

Acknowledgement

A number of people deserve thanks for their role in this habilitation. First, I want to thank Michael Waldmann and the members of his department of Cognitive and Decision Sciences for supporting my research at the University of Göttingen while a large part of this work was completed. I also want to thank student assistants who helped me implement the experiments, as well as students, who wrote their BSc or MSc theses under my supervision during this time. Thanks, Michael for making all of this possible, and thanks York Hagmayer, Simon Stephan, Alex Wiegmann, Juan Marulanda, Neele Engelmann, Birgit Bergmann-Bryant, Dominik Glandorf, Louisa Reins, and Maike Holland-Letz for your help and time.

Second, I want to thank Karl Christoph Klauer for making the habilitation possible by (again) welcoming me to his department of Social Psychology and Methodology at the University of Freiburg. I have always enjoyed our collaboration, Christoph, and I am happy to be here again in Freiburg. The German Research Council (DFG) also deserves a big thanks for supporting me with a research grant for my project.

Next, I thank Wolfgang Spohn and Markus Knauff for their continued support over the years and for their prudent advice that I do the habilitation in psychology. For the individual chapters, I want to thank my co-authors for their contributions. I also want to thank the editors and peer reviewers, who relentlessly forced my co-authors and I to be more precise, write in a more accessible manner, and to embed the research better in the psychological literature. Finally, I thank members of the audiences that I presented the individual chapters to throughout the years for their comments and questions. Individual chapters also contain further acknowledgements of contributions made by colleagues in different parts of the world.

This habilitation is dedicated to my parents – Niels Viggo Skovgaard Olsen and Anne Grete Skovgaard Olsen – and my brother Jens Skovgaard-Svane. Without your continued support and love, this udlandsdansker would not have made it thus far.

Part I
Introduction

Chapter 1: An Overview

Niels Skovgaard-Olsen

The present book consists of a collection of reasoning studies. The experimental investigations will take us from people's reasoning about probabilities, entailments, pragmatic factors, argumentation, and causality to morality. There are various common themes that unite the investigations, which are outlined below.

First, we will begin by highlighting a general problem in the recent Bayesian paradigm in the psychology of reasoning (see, e.g., the essays in Elqayam, Bonnefon, & Over, 2016), which is raised in Chapter 3, and which has not yet received the attention it deserves. Next, themes that unite the reasoning studies in Chapters 2-6 in the second part of the book are outlined. Finally, problems arising from the possibility of multiple, conflicting norms and individual variation investigated in Chapters 7-9 are presented.

1.1 Ideal Bayesian Rationality and Deductive Competence

In Elqayam and Over (2013, p. 259), it is said to be characteristic of the so-called New Paradigm in the psychology of reasoning that: "Studying probability judgements will tell us much more about the psychology of reasoning than trying to find out how far people conform to binary extensional logic in any deductive reasoning in which they engage". Given this popular view, it may be worthwhile to take a step back and review some of the developments in the psychology of reasoning, which led to this development.

In Rips (1994), the *Deduction-System Hypothesis* is advanced that as part of our native, cognitive architecture deductive principles underlie many of our other cognitive abilities (e.g., text comprehension, problem solving, categorization, and action planning). As a competence model, Rips employs a modified version of natural deduction, whereby logical connectives like 'and', 'or', 'not', and 'if-then' are characterized by inferential introduction and elimination rules (see Table 1).

Table 1. Introduction/Elimination Rules for Classical Sentential Logic

IF Elimination (modus ponens)

(a) If sentences of the form IF P THEN Q and P hold in a given domain,
(b) then the sentence Q can be added to that domain.

IF Introduction (Conditionalization)

(a) If a sentence Q holds in a subdomain whose supposition is P,
(b) then IF P THEN Q can be added to the immediate superdomain.

NOT Elimination (Reductio ad absurdum 1)

(a) If the sentences Q and NOT Q hold in a subdomain whose supposition is NOT P,
(b) then the sentence P can be added to the immediate superdomain.

NOT Introduction (Reductio ad absurdum 2)

(a) If the sentences Q and NOT Q hold in a subdomain whose supposition is P,
(b) then NOT P can be added to the immediate superdomain.

Double Negation Elimination

(a) If the sentence NOT NOT P holds in a given domain,
(b) then the sentence P can be added to that domain.

AND Elimination

(a) If the sentence P AND Q holds in a given domain,
(b) then the sentences P and Q can be added to the domain.

AND Introduction

(a) If the sentence P and the sentence Q hold in a given domain,
(b) then the sentence P AND Q can be added to that domain.

OR Elimination Introduction

(a) If the sentence P OR Q holds in a given domain D,
(b) and the sentence R holds in an immediate subdomain of D whose supposition is P,
(c) and the sentence R holds in an immediate subdomain of D whose supposition is Q,
(d) then R can be added to D.

OR Introduction

(a) If the sentence P holds in a given domain,
(b) then the sentences P OR Q and Q OR P can be added to that domain, where Q is an arbitrary sentence.

Note. Based on Table 2.3 in Rips (1994, p. 45). The rules that Rips implements in his computer program, PSYCOP, include modifications to the introduction and elimination rules listed in this table (see ibid., Tables 3.2, 4.1, and 4.2 for further details).

The theory postulates a notion of mental proof, whereby people untrained in logic reason from premises to conclusions by applying a series of mental inference rules. In this process, premises are stored in working memory and a goal for the deduction is set by working backwards from the conclusion. Some of the rules in Table 1 can be applied in a forward direction by adding further sentences to working memory as conclusions, once the premises are added (e.g., IF Elimination, AND Elimination). Others are applied in a backward direction to identify subgoals that contribute towards the final goal of deducing the conclusion. E.g., if the conclusion of the argument contains a disjunction (‘P or Q’), then OR Introduction can be used to identify the subgoal of deducing P, and if the conclusion contains a conjunction (‘P and Q’), then AND Introduction can be used to identify the subgoals of proving P and Q, separately.[1]

From the premises, the model then applies a combination of deductive rules that proceed either in said forward direction by generating new assertions or in the backward direction from a goal to a subgoal, until the conclusion has been proven. In addition, heuristics are used to make the search more efficient (e.g., by applying backward rules that involve simpler subgoals first; or rules that have been successfully applied recently). At several junctures, Rips (1994) modifies the model to make it more psychologically realistic[2] and has implemented it in the computer program PSYCOP. In addition to rules based on classical sentential logic, PSYCOP incorporates rules for dealing with quantifiers (∀, ∃) by first translating quantified sentences into sentences without quantifiers, which contain variables and temporary names, to model Aristotelian syllogisms, and rules for variable binding to permit the model to generalize over instances and handle examples from predicate logic.

PSYCOP is just one of several competing models based on natural deduction and the notion of a mental logic that were popular from the 70ies until the 90ies (Osherson, 1975; Braine, 1978; Macnamara, 1986). Today, theories with this architecture are more rarely advanced (Evans, 2002). Several events contributed to this state of affairs. Some are external to the psychology of reasoning. For instance, in the first half of this interval it was commonly assumed in cognitive science that there were central processes in higher cognition based on logic in an innate language of thought, which were supplemented by specialized modules (Fodor, 1975, 1983; Fodor & Pylyshyn, 1988; Pylyshyn, 1989), and research on symbolic reasoning and rule-based expert systems, e.g., based on logic programming, was the main paradigm in AI. Later, this paradigm encountered competition through connectivism (Bermúdez, 2010). Today,

[1] This backward use of deductive principles is invoked to deal with the problem of clutter-avoidance identified by Harman (1986) as one of the hindrances in applying deductive logic (without further bridge principles) as a theory of reasoning. The problem is that rules like OR Introduction and AND Introduction are capable of generating infinite sequences of sentences with no immediate practical use if applied recursively, which would clutter up working memory (for further discussion, see MarFarlane, draft; Skovgaard-Olsen, 2015).

[2] E.g., an upper limit is placed on the number of subgoals in a proof (p. 123), and additional inference rules that participants have applied in experiments were added to the model to compensate for inference rules, like OR-Introduction, that participants rarely use (112f.).

probabilistic models, e.g., based on Bayes nets or machine learning are increasingly used (Darwiche, 2009; Pearl, 2009; Miller & Forte, 2017; Lantz, 2019).

Other events were internal to the psychology of reasoning. In several chapters, we will encounter probabilistic theories and Mental Model Theory as the main contenders in the discussions of our experimental findings. Some of the developments internal to the psychology of reasoning that led to these developments were:

1) A common perception of poor performance of participants on tasks based on classical logic (Evans, 2002). For instance, Oaksford and Chater (2009, p. 69) write: "*Bayesian Rationality* (Oaksford & Chater, 2007, hereafter *BR*) aims to re-evaluate forty years of empirical research in the psychology of human reasoning, and cast human rationality in a new and more positive light. Rather than viewing people as flawed logicians, we focus instead on the spectacular success of human reasoning under uncertainty". On p. 72, they continue: "In Chapter 3, we argue that real-world, informal, everyday, reasoning is almost never deductive, that is, such reasoning is almost always logically *in*valid". Several of their commentators agree with their overall assessment. Evans (2009), for example, writes: "I agree strongly with them [i.e., Oaksford and Chater] that the Bayesian model is a far more appropriate reference for real world reasoning than one based on truth-functional logic, and that it is a standard much more likely to be approximated in the inferences that people actually make" (Evans, 2009, p. 89). Similarly, Khali (2009) concurs: "Let us agree with O&C that the probabilistic approach to human reasoning explains why humans, in laboratory settings, are bad at solving tasks formulated by classical logic" (p. 92).
2) A common perception that the monotonicity of classical logic is ill-suited to account for common-sense reasoning. In classical logic, a logical consequence of a set of premises remains valid when further premises are added. In nonmonotonic reasoning this property does not hold and an earlier drawn conclusion may later have to be retracted without giving up the original premises (Oaksford & Chater, 1991; Pfeifer & Kleiter, 2005). As Oaksford and Chater (2009, p. 73) argue: "there appears to be a fundamental mismatch between the nonmonotonic, uncertain character of everyday reasoning, and the monotonicity of logic; and this mismatch diagnoses the fundamental problem with logic-based theories of reasoning and logicist cognitive

science more broadly”. On p. 106, they add: “outside mathematics, deductive reasoning, which *guarantees* the truth of a conclusion given the premises, is, to a first approximation, never observed” (p. 106). On p. 4 in *BR* they state that: “We shall suggest that logic is simply not an appropriate foundation for understanding informal, everyday thought”. Oaksford and Chater (1991) recognize that classical logic could be replaced with a nonmonotonic logic developed in AI or formal epistemology but argue that the latter involves making computationally intractable consistency checks of the complete knowledge base each time default rules are applied.[3]

3) Discussions between Mental Logic and Mental Model Theory over whether reasoning is based on the employment of syntactical rules or semantic reasoning with mental models (see, e.g., Johnson-Laird, 1997; Rips, 1997; Elqayam, 2003; Knauff, 2007). Mental Model Theory denies the assumption that reasoning involves applying mental rules to logical formulas and holds instead that people reason using concrete representations in which the formulas are true, which they use to search for counterexamples to arguments. These representations can either be incomplete mental models that merely satisfy the premises of an argument or fully explicit mental models that cover all the cases in a truth table. It has, *inter alia*, been argued that the behavioural findings of illusory inferences and activation of brain areas involving the visual cortices and the parietal cortex support Mental Model Theory in this foundational debate over the nature of the mental representations used in deductive reasoning (Johnson-Laird, 1997; Knauff, 2007).[4]

4) Arguments against the material implication (‘⊃’) of classical logic as a theory of natural language conditionals (e.g., Evans, 2002; Oaksford & Chater, 2007). But while this objection is frequently raised against the deductive paradigm in the

[3] In contrast, Stenning and van Lambalgen (2008, 2009) use the nonmonotonicity of common-sense reasoning to motivate use of nonmonotonic logic in psychology and point out that improvements to the computational complexity of these systems have been made. Moreover, in Chap. 5, Oaksford and Chater (2007) point out that the problem of having to make consistency checks in a database of world knowledge re-emerges once the probabilistic paradigm applies the Ramsey test to evaluate natural language conditionals.

[4] But as Knauff (2007) argues, the evidence from brain research is mixed and is consistent with the use of different reasoning strategies; including ones that activate language-related areas in the temporal cortices with a left-hemispheric prevalence, as predicted by Mental Logic.

> psychology of reasoning, it is not fully accurate. It used to apply to Mental Model Theory before the revisions to Mental Model Theory that we will review in later chapters, but it lacks basis for Mental Logic (see also Evans & Over, 2004, Ch. 3). As O'Brien (2009) points out, the mental logic of Braine and O'Brien (1991) is not based on the material implication of classical logic and neither is Rips' (1994, Ch. 4).[5] In Osherson (1975, p. 55), it is said that "of all the conditional statements possible to express, truth-functional material implication is one of the least interesting. We might hope that a proposed model would not rely on that particular interpretation of "if…then…" locutions".

Usually, having deductive competence is translated to manifesting the ability to reason in accordance with classical logic in characterizations of the deductive paradigm (see, e.g., Evans, 2002; Oaksford & Chater, 2007, 2009, 2020a; Knauff & Gazzo Castañeda, 2021). Yet, this last point illustrates that the relationship between Mental Logic and classical logic is actually more complicated. For instance, in the case of Rips (1994), the correspondence between deductive competence and classical logic does not hold, because: a) Rips is fully aware of the problem posed by alternative logics (as we shall see in section 1.3.2), and b) the natural deduction system that Rips implements in PSYCOP is not complete with respect to classical logic by design (Chap. 4), because it rejects the material implication of classical logic, as noted above. Similarly, Osherson (1975) constructs a mental logic that is not logically complete, guided by a notion of psychological completeness. In contrast, Braine (1978) still tried to show that it was possible to recover classical logic from the modified rules of natural deduction used in his mental logic by further modifications, which leads him to conclude that "natural and standard propositional logic are the same system on different foundations" (p. 18).

In what follows, we focus on the shift towards Bayesian probabilistic theories of reasoning. In Chapters 2, 5, 6, and 8, we will return to Mental Model Theory. There is a debate about whether it is appropriate to characterize the Bayesian approach in psychology of reasoning

[5] Rips (1994, pp. 125-138) has designed PSYCOP such that the following inference rules are missing, which would render the conditional in PSYCOP logically equivalent to the material implication in classical logic (a move which Rips resists, p. 25ff.):

$$\frac{\text{NOT (IF P THEN Q)}}{\therefore \text{ P}} \qquad \frac{\text{NOT (IF P THEN Q)}}{\therefore \text{ NOT Q}} \qquad \frac{\text{IF P THEN Q}}{\therefore \text{ (NOT P) OR Q}}$$

as a new (Kuhnian) *paradigm*, when contrasting it with the deductive approach (see Knauff & Gazzo Castañeda, 2021). Here we continue to use these terms, although talk of competing frameworks may be more neutral and would fit better with the argument presented below on the role of classical logic in the probabilistic paradigm (see section 1.1.1).

If we paraphrase a bit, we might identify the following lines of argument as influential in the psychology of reasoning. The arguments are rarely (if ever) stated explicitly in this form, but arguments along these lines appear frequently as enthymemes with suppressed premises.

Deductive Paradigm Optimist	*Premise 1a:*	Participants are rational if and only if they have deductive competence.
	Premise 1b:	Participants have deductive competence.
	Conclusion:	Therefore, participants are rational.
Deductive Paradigm Pessimist	*Premise 2a:*	Participants are rational if and only if they have deductive competence.
	Premise 2b:	Participants do not have deductive competence.
	Conclusion:	Participants are not rational.
Probabilistic Paradigm Optimist	*Premise 3a:*	Premise 2b is true but Premise 1a/2a is false.
	Premise 3b:	If participants judge in accordance with probability theory, they are rational.
	Premise 3c:	Participants judge in accordance with probability theory.
	Conclusion:	Therefore, participants are rational.

In 1)-2) above we have already seen statements that appear to affirm Premises 3a and 3b. A further claim that points in this direction is: “According to this viewpoint, the apparent mismatch between normative theories and reasoning behaviour suggests that the wrong normative theories may have been chosen; or that the normative theories may have been misapplied.” (Oaksford & Chater, 2007, p. 30). Moreover, Oaksford and Chater (2009, p. 70) roughly summarize their book, *Bayesian Rationality* (*BR*), as attempting to motivate Premises 3a and 3b in the first part (Chapters 1-4), while attempting to establish Premise 3c in the second part (Chapters 5-7).

Oaksford and Chater (2009) are, of course, aware of the empirical problems relating to Premise 3c, given well-known biases such as base-rate neglect and the conjunction fallacy from

Judgment and Decision Making (see, e.g., Tversky & Kahneman, 1983; Gilovich, Griffin, & Kahneman, 2002). But they argue that while participants often make errors in precise numerical calculations, and may have difficulties expressing their degrees of beliefs in numerical values, they follow qualitative patterns of probabilistic reasoning (Oaksford & Chater, 2007, pp. 80-98). However, there is a further problem with Premise 3c. In Chapter 3 below, it is shown that the combination of Premises 3a and 3c is problematic. This problem casts doubt on claims to the effect that probabilistic approaches appear "to radically reduce the gap between rational norms and human behavior" (Oaksford & Chater, 2007, p. ix) and introduces complications for general statements such as "The logical mind should be replaced by the probabilistic mind" (p. 7). But since that argument is presented in a dense form as part of a published article, I will here try to restate the argument in a slightly different form.

1.1.1 Applying Probability Theory to Language

There are two standard routes for extending probability theory from its canonical application in assigning probabilities to the outcomes of random experiments (where the probabilities of events as subsets of the sample space are specified) to reasoning with language (where the probabilities of sentences or propositions being true are defined). One is to define an algebra of propositions representing the content of sentences and assign probabilities directly to propositions. Another is to define a formal language based on sentential logic and define a probability function for well-formed formulas of the language (see, e.g., Grandy & Osherson, 2010; Spohn, 2012; Huber, 2018; Peterson, 2017). We will here briefly consider these two routes to specify some of their implications for the deductive competence presupposed by ideal Bayesian agents, whose degrees of beliefs are given by a probability distribution.[6]

The first, propositional, route starts by defining a probability space, $\langle W, \mathcal{A}, P\rangle$, which consists of: a) a non-empty set of possible worlds, W, b) an algebra, $\mathcal{A},$ defined over W, and c) a probability measure, P, defined over $\mathcal{A}$.

[6] Leitgeb (2016) mentions a range of more advanced probabilistic logics that, e.g., include non-standard probabilities, probabilities defined for non-classical logics, and the possibility that certain subsets of the sample space W are not assigned probabilities at all. Moreover, Pfeifer and Kleiter (2005) use a coherence approach that does not start out with a complete algebra of propositions. The goal here is not to survey all available options but merely to illustrate some of the idealizations involved in the standard approaches.

To define probabilities for propositions (as the content or meaning of sentences), we let W be the set of all possible worlds, or possible states of affairs. Propositions are then subsets of W, or sets of possible worlds. Propositions are members of the algebra, $\mathcal{A}$. The algebra, $\mathcal{A}$, is a set of subsets of W such that for all subsets A and B of W:

1) $W \in \mathcal{A}$ (i.e., the set of all possible worlds is a proposition).
2) If $A \in \mathcal{A}$ then $A^C \in \mathcal{A}$ (i.e., for every proposition, its complement is also a proposition).
3) If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ (i.e., if A and B are propositions, then their union is also a proposition).

Probabilities are then assigned to the propositions that are members of $\mathcal{A}$. The function P, whose domain is $\mathcal{A}$ and whose co-domain is the set of real numbers, $\mathbb{R}$, $P: \mathcal{A} \rightarrow \mathbb{R}$, is a probability measure if, and only if, the following holds for all propositions, A and B, in $\mathcal{A}$:

4) $0 \leq P(\text{A}) \leq 1$
5) $P(W) = 1$
6) If $(\text{A} \cap \text{B}) = \emptyset$, then $P(\text{A} \cup \text{B}) = P(\text{A}) + P(\text{B})$

If A = B, then $P(\text{A}) = P(\text{B})$, which means that logically equivalent propositions are assigned the same probability (Spohn, 2012; Huber, 2018).

For the second, sentential route, a vocabulary with n sentential variables ($p, q, r\ldots$), the logical connectives $\{\neg, \vee, \wedge, \supset, \leftrightarrow\}$, and parantheses are used. A formal language, $\mathcal{L}$, over this vocabulary is then specified by stating a list of rules of how to apply the logical connectives to generate complex formulas from the atomic formulas given by the sentential variables. Thirdly, a function, P, is defined, which has $\mathcal{L}$ as its domain and the set of real numbers as its co-domain, $P: \mathcal{L} \rightarrow \mathbb{R}$. This function is a probability function if and only if the following holds for every well-formed formulas φ and ψ in $\mathcal{L}$:

7) $0 \leq P(\varphi) \leq 1$
8) $P(\varphi) = 1$ if φ is logically true
9) $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$ if φ and ψ are mutually exclusive

These axioms imply that sentences or well-formed formulas that are logically equivalent (and hence express the same proposition) are assigned the same probability (Grandy & Osherson, 2010; Huber, 2018). Thus, for all well-formed formulas φ and ψ in $\mathcal{L}$:

$$P(\varphi) = P(\psi) \text{ if } \varphi \text{ and } \psi \text{ are logically equivalent.}$$

These two routes are alternative ways of formulating probability theory, which assign probability to different entities. But they are closely related, as we shall see shortly. If the vocabulary has *n* sentential variables, there are 2^n possible truth-value assignments, which specify all the possible ways to distribute the values 'True' and 'False' to the *n* variables. These truth-value assignments play the role of possible worlds, if we only consider the possible states of affairs as specified to the resolution of the *n* sentential variables. Sets of truth-value assignments to formulas in $\mathcal{L}$ then correspond to propositions. Propositions in turn specify the meaning, or content, of formulas of the formal language by stating the sets of possible worlds in which the formulas are true. Accordingly, for any given formula of the language, $\varphi \in \mathcal{L}$, its propositional content, or meaning, $[\varphi]$, is given by a set of possible worlds in which the formula is true:

$$[\varphi] = \{\mathrm{w} \in W \,|\, \mathrm{w} \vDash \varphi\}$$

Before we denoted propositions as, e.g., A and B, and declared them to be members of the algebra, $\mathcal{A}$. Now we have a way of expressing that a proposition is a set of possible worlds in which a formula is true. Here φ is a meta-variable that can either be an atomic formula, given by a sentential variable, *p*, or some complex formula generated by applying logical connectives to atomic formulas (e.g., '$\neg\, (p \vee q)$'). So, while the first route assigned probabilities to propositions, we now see that these propositions express the content or meaning of the formulas of $\mathcal{L}$ that the second route directly assigns probabilities to. Moreover, just as it is possible to assign probabilities both directly to outcomes of the sample space and to sets of outcomes (i.e., events) in statistics via additivity, so probabilities can be assigned directly to possible worlds and summed up for probability assignments to either propositions or formulas. In the former case, $P([\varphi]) = \sum_{w \in [\varphi]} P(w)$, for $w \in W$. In the latter case, $P(\varphi) = \sum_{w \vDash \varphi} P(w)$, for $w \in W$ (Grandy & Osherson, 2010, Chap. 7).

The logical connectives correspond to set-operations applied to propositions (with complement, ∪, and ∩ fulfilling the roles of negation, ∨, and ∧ applied to formulas). Of the logical connectives, the set {¬, ∨} is a functionally complete set of logical connectives in terms of which the remaining connectives of {¬, ∨, ∧ , ⊃, ↔} can be defined (Wernick, 1942). Thus, complement and ∪ can be used to generate set-relations corresponding to every logically complex formular, which could be generated by applying the logical connectives, {¬, ∨, ∧ , ⊃, ↔}, to sentential variables. So, if A, B, A^C, B^C, and A ∪ B are members of the algebra (and assigned a probability via the first route), then any other set relation that we can express by logically complex combinations of sentential variables and the connectives, {¬, ∨, ∧ , ⊃, ↔}, can be defined in terms of members of the algebra and be assigned a probability. Similarly, since probabilities are defined for every well-formed formula of the language, $\mathcal{L}$, via the second route, and $\mathcal{L}$ includes logically complex formulas generated by repeated applications of the logical connectives to other formulas, each of these formulas is assigned a probability in accordance with the axioms of probability theory.

If either of these two probability functions represents the degrees of belief of an ideal Bayesian agent, then this agent should be able to assign probabilities to every logically complex formula or set-theoretic relation that can be formulated for the domain for which the probability function is defined. Let us briefly consider how formidable a task that is.

The set of all possible subsets of W, $\mathcal{P}(W)$, specifies how many unique propositions there are as an upper limit. For finite *W*, where we limit the number of possible worlds to correspond to the 2^n possible truth-value assignments to the *n* sentential variables in $\mathcal{L}$, there are finitely many propositions. Each possible world can either be included or excluded in a set of possible worlds, and so there are 2^n independent choices of inclusion or exclusion. Since each of these choices are independent, there are 2^{2^n} combinations of sets of possible worlds, or 2^{2^n} distinct propositions to consider (Grandy & Osherson, 2010, Chap. 4).

Table 2 illustrates how fast these numbers grow:

Table 2. Number of Propositions for *n* sentential variables

Number of sentential variables	*Number of truth-value assignments (or possible worlds)*	*Number of propositions (or sets of possible worlds)*
$n = 2$	4	16
$n = 3$	8	256
$n = 4$	16	65536
$n = 5$	32	4294967296
$n = 6$	64	18446744073709551616

Given our current restriction of W to possible worlds corresponding to truth assignments of the n sentential variables in $\mathcal{L}$, it holds that: for every proposition, A, there is a formula, $\psi \in \mathcal{L}$, in disjunctive normal form such that $[\psi] = \mathrm{A}$.[7] To illustrate, a formula like '$(p \wedge \neg q) \vee (r \wedge t)$' is written in disjunctive normal form, and so is '$p \wedge \neg q \wedge r \wedge t$', since the latter is considered as a disjunction with just one disjunct (ibid, p. 102).

Given that the set $\{\neg, \vee\}$ is functionally complete with respect to the set of logical connectives, $\{\neg, \vee, \wedge, \supset, \leftrightarrow\}$, every complex formula generated by applying connectives of the latter set to well-formed formulas in $\mathcal{L}$ are logically equivalent to a formula written in disjunctive normal form (Grandy & Osherson, 2010, Chap. 5). Of the 2^{2^n} unique propositions, each is then expressible in $\mathcal{L}$. But there are infinitely many expressions in $\mathcal{L}$ that express the same propositions. These infinitely many expressions can be generated by recursively applying the rules for generating well-formed formulas in $\mathcal{L}$ based on repeated applications of its logical connectives to formulas that are themselves produced by applications of the rules for constructing well-formed formulas. The 2^{2^n} propositions enter into subset relations which introduce entailments between the expressions (ibid, pp. 99-107). These entailments must be respected in the probability assignments of the ideal Bayesian agent, as we will see below.

Since we will return to Bayes nets in Chapter 6, it is useful to briefly comment on their use in the present context. By applying Bayes nets, a joint probability distribution over the n sentential variables can be specified via a small number of marginal probabilities and conditional probabilities by exploiting the conditional independencies encoded in the graphical structure of

[7] The formula, ψ, is written in disjunctive normal form, if it is either a simple conjunction, or a disjunction of simple conjunctions, according to the definition in Grandy and Osherson (2010, pp. 101-103). Here a single conjunct, p, or the negation of a single conjunct, $\neg p$, is included as a simple conjunction, along with conjunctions of variables and negation of variables, and disjunctions with just one disjunct are included as disjunctions of simple conjunctions.

Bayes nets (for examples of these graphical structures, see Chap. 6). This alleviates the need for assigning probabilities to each of the 2^n truth-value assignments in advance to specify a joint distribution (Darwiche, 2009). From the graph, we learn which variables are parent nodes, *pa*(A), and apply the chain rule from probability theory to factorize the joint probability over the set of variables, $A_1 \ldots A_n$, as follows (Hartmann, 2021):

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P\big(A_i \big| pa(A_i)\big)$$

$$= P\big(A_1 \big| pa(A_1)\big) \cdot P\big(A_2 \big| pa(A_2)\big) \cdots P\big(A_n \big| pa(A_n)\big)$$

By supplying a generative model for degrees of beliefs, Bayes nets make the representation of probability distributions of many sentential variables more compact and deal with the problem of the intractability of specifying a joint distribution based on probability assignments to all truth-value assignments and its associated storage problems. But the ideal Bayesian agent still needs to be able to produce all the correct probability assignments when queried, and so the requirements for the ideal Bayesian agent remain demanding even when reasoning with Bayes nets. Furthermore, this factorization of probability distributions does not undermine the point made shortly about the deductive competence presupposed by ideal Bayesian agents (see also Eva & Hartmann, 2018, for a discussion of the value of logic validity in the context of Bayesian updating).

In general, degrees of belief of rational Bayesian agents are constrained by the properties of logical truth, logical consequence, consistency, and logical equivalence as follows (Adams, 1998, pp. 21-24; Grandy & Osherson, 2010, p. 239ff.):

i) If φ is *logically true*, then their degree of belief in φ should be: $P(\varphi) = 1$,

ii) If φ *logically implies* ψ, then their degrees of belief in φ and ψ should conform to the inequality: $P(\varphi) \leq P(\psi)$

iii) If φ and ψ are *logically inconsistent*, then their degrees of belief in φ and ψ should conform to: $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$

iv) If φ and ψ are *logically equivalent*, then their degrees of belief in φ and ψ should conform to: $P(\varphi) = P(\psi)$

Since φ and ψ are variables that represent formulas of $\mathcal{L}$ of arbitrary complexity, these principles introduce the requirement that the ideal Bayesian agent should recognize arbitrarily complex, logical relations in the assignment of degrees of beliefs.

It is a familiar point that formal logic could be considered as dealing with the extreme case of complete certainty in the premises, where each premise is assigned a probability of 1, and that probability theory generalizes to cases of reasoning under uncertainty with probabilities less than 1 (see, e.g., Oaksford & Chater, 2009, p. 107). But the principles listed above also apply when the agent is reasoning about states of affairs, where one or several premises have probabilities less than 1. Indeed, a famous case where ii) is violated is in the conjunction fallacy (Tentori, et al., 2004). Accordingly, these principles illustrate the deductive competence built into probability theory, which is presupposed by the ideal Bayesian agent.[8] As Adams (1998, p. 22) says: “Pure logic is a prerequisite to probability logic because the probability axioms, and the theorems that follow from them, depend on concepts of logical truth, consequence, and consistency”.

While work on the relationship between logic and probability, like Adams (1998), is often cited in the psychology of reasoning, the consequences of these principles for frequent claims, such as the following, are rarely discussed:[9] “probability, rather than logic, provides an appropriate framework for providing a rational analysis of human reasoning; and (…) this undercuts existing logic-based theories of reasoning” (Oaksford & Chater, 2007, p. 16), and “We claim that almost no everyday human reasoning can be characterized deductively, or has any significant deductive component” (ibid, p. 43).

One instance of this is to leave out that the probabilistic semantics, e.g., of Adams (1998), is not suitable for an attempt to *replace* a logical semantics with a probabilistic one, as envisaged, e.g., in the following passage (Oaksford & Chater, 2009, pp. 106-107):

[8] In Grandy and Osherson (2010, pp. 240-41), it is shown that i), iv), and iii) generalized to n formulas, $\varphi_1 \ldots \varphi_n \in \mathcal{L}$, is necessary and sufficient for any function $F\colon \mathcal{L} \rightarrow [0,1]$ to represent a probability distribution. Since the generalized version of iii) follows from the law of total probability, they use it to state a representation theorem.

[9] Exceptions are when Pfeifer and Kleiter (2005), Oaksford (2014), and Schurz (2014) also point out that probability theory presupposes deductive logic. Yet, neither discuss the consequences that I here point to of principles i)-iv) for the probabilistic paradigm.

> The ubiquity of uncertain, knowledge-rich inference, argues for an alternative to invoking the semantics/pragmatics distinction to maintain a logical semantics for natural language [in the manner of Grice (1989)]: namely, that natural language semantics may be probabilistic "all the way down." Experiments in the psychology of reasoning, as reviewed in *BR,* find little support for the existence of a level of logic-based representation or inference. *BR* proposes a starting point for a probabilistic semantics: *If p then q* conditionals are assumed to express that the conditional probability *P(q|p)* is high (following Adams 1975; 1998; Bennett, 2003; and Edgington, 1995, among others); the quantifiers *Some, Few, Most, All* are similarly assumed to express constraints on probabilities (e.g., *Some A are B* is rendered as *P(A, B)* > 0; *Most A are B* claims that *P*(*B*|*A*) is high). Switching from a logical to a probabilistic semantics provides, we argue, a better fit with patterns of human reasoning. Of course, it remains possible that a logical core interpretation might be maintained – but it seems theoretically unparsimonious to do so (Edgington, 1995).

The issue is not whether one wants to make an unparsimonous choice of adding a logical part to the probabilistic semantics, because as Adams (1998, p. 154) points out: "presupposing classical logical relations among factual formulas means that our theory is an *extension* of classical logic, and not an alternative to it". As a special case: while the theory introduces a probabilistic semantics for natural language conditionals, it retains a logical semantics for negations, disjunctions, and conjunctions as applied to factual formulas. The same holds, e.g., for the probabilistic semantics of Grandy and Osherson (2010, Chapters 9-10), which, like Adams' (1998), explicates the meaning of natural language conditionals in terms of conditional probabilities.

However, the more general case concerns principles i)-iv) above. These principles illustrate the problem with, on the one hand, accepting that poor logical performance of participants in psychological experiments demonstrate that participants lack deductive competence, whilst, on the other, arguing that participants continue to be rational, because their performance is well-captured by a probabilistic model (cf. section 1.1). The problem is that the normative theory of Bayesian rationality already requires participants to adequately identify *logical tautologies*, *contradictions*, *logical equivalences*, and *entailments* in their probability assignments while requiring that they coherently produce numbers between 0 and 1 that sum up to 1. This is why it is argued in Chapter 3 that probability theory adds further requirements of rationality; not less. So, if participants lack the deductive competence to meet the demands of the deductive paradigm, it is difficult to see how turning to probability theory may help to restore their rationality.

One potential response would be to limit the scope to certain problematic expressions like the conditionals and quantifiers that Oaksford and Chater (2007) focus on. It could then be argued that the lack of deductive competence established by previous experimental results is restricted to their treatment in classical logic and that extensions of classical logic via probabilistic accounts improve participants' performance in psychological experiments. It would still have to be shown, however, that participants possess deductive competence with other well-formed formulas of classical logic, and general statements about participants' lack of deductive competence would have to be qualified.[10]

Yet, as Braine and O'Brien (1991, p. 185) note: "no psychological theory should predict perfect accuracy under conditions of complexity", and so some restriction on the complexity of the derived formulas covered by this conjectured deductive competence would be needed. In Braine (1978), this problem was handled by conjecturing that ordinary people lack explicit awareness of the inference rules introduced by Mental Logic and that ordinary people did not apply the inference rules to arbitrarily chosen statements, and so they are predicted to be unaware of the entailments of inference rules that logicians are able to derive. Similarly, Braine and O'Brien (1991) restrict application of the inference rules for derived theorems and conjecture that ordinary people lack meta-logical awareness of the rules of their mental logic. For the probabilistic paradigm, corresponding restrictions are needed for the complexity of the formulas that participants can be expected to assign probabilities to. Emphasizing as Oaksford and Chater

[10] In support of this line of argument, a case could be made that the claim of participants' lack of deductive competence is often based on results from Aristotelian syllogisms, conditional syllogisms, and the Wason selection task. Yet, Rips (1994) reports further experiments with other introduction and elimination rules from classical logic and predicate logic that showed a comparably better performance, provided that further processing assumptions are granted. Osherson (1975) also describes a number of valid arguments without quantifiers accepted by adolescents, which extend far beyond the few inference rules (e.g., MP, MT, AC, DA) normally tested. Moreover, even when reviewing the biases and pragmatic influences of content and context uncovered by the deductive paradigm, Evans (2002, p. 992) emphasizes that there is still a rudimentary deductive competence that participants show, which needs to be accounted for empirically.

(2007, pp. 80-98) do that participants do not comply with probability theory in their numerical reasoning but only in their qualitative reasoning may go some of the way, but it does not yet address the problem outlined above about whether participants' probabilistic reasoning conforms to the deductive competence presupposed by Bayesian rationality.

1.2 Part II: The Semantics and Pragmatics of Conditionals

The intention of the foregoing argument was not to downplay the tremendous impact that the turn to probability theory has had on the psychology of reasoning. Indeed, use of response formats that trigger probabilistic reasoning has led to much fruitful work in the psychology reasoning (reviewed, e.g., in Oaksford & Chater, 2020a), which many of the chapters of this book build on and try to contribute to. As we have seen, a central motivation for the shift to the probabilistic paradigm (e.g., in Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2005; Pfeifer, 2013)[11] has been a dissatisfaction with the material implication, '⊃', of classical logic as an account of natural language conditionals and the hope that a probabilistic account of conditionals may fare better, which draws on the works of Adams (1975, 1998). Part II of this book therefore contains 5 chapters that investigate various aspects of conditional reasoning through a series of experiments.

On Adams' account, the formal language of section 1.1.1 does not yet contain a connective that adequately accounts for the probability of indicative conditionals ('if *p*, then *q*') in natural language. To introduce it, Adams extends the language with conditional events, 'φ_s | ψ_s', and defines their probability to be conditional probabilities. As Grandy & Osherson (2010) note, one way to consider conditional events is not just as a new formula in the language, but as an ordered pair of formulas, (φ_s, ψ_s), whose meaning cannot be reconstructed in terms of truth-functional combinations of negations, conjunctions, and disjunctions, and where 'φ_s' and 'ψ_s'

[11] In contrast to the others, the coherence approach of Pfeifer and Kleiter (2005) and Pfeifer (2013) preserves a *deductive* relation between premise and conclusion of arguments and uses it to assign probability intervals to the conclusion based on the probabilities of the premises. Thus, instead of departing from the mental logic of Rips (1994) by turning to probability theory like, e.g., Oaksford and Chater (2007), Pfeifer and Kleiter propose a *mental probability logic* in its place, which requires people to reason deductively about probability intervals. Yet, unlike Rips (1994), who attempts to specify a mechanism whereby participants could reason deductively (via the construction of mental proofs), and who attempts to account for the logical errors of participants, these parts seem to lack counterparts in the mental probability logic account.

only range over sentential variables, like *p* and *q*. Due to the triviality results of Lewis (1976), it has been proven that there is no unconditional proposition expressible by a formula in $\mathcal{L}$ such that it has the same probability as the conditional probability of *q* given *p*, for all probability distributions. So, if the probability of conditionals are conditional probabilities, then their probability cannot equal the probability that a proposition is true, in the same way in which the probability of a factual statement, *p*, is normally taken to equal the probability that the factual statement is true.[12] Adams' (1975, 1998) work on formulating a probability logic that is compatible with these requirements has, through its extensions in Edgington (1995) and Bennett (2003), been one of the main sources of inspiration for the New Paradigm in the psychology of reasoning. Indeed, so much so that Vance and Oaksford (2002) directly characterize the new Bayesian paradigm via its acceptance of the thesis that $P(\text{if } p \text{ then } q) = P(q|p)$, which is commonly referred to as "the Equation", following Edgington (1995).[13]

1.2.1 Semantic and Pragmatic Processes in Reasoning

In Chapters 2, 3, 6, and 7, earlier results by Skovgaard-Olsen et al. (2016) that challenge "the Equation" by a finding that we will come to know as "the Relevance Effect", play a key role. Using so-called missing-link conditionals (e.g., 'If Niels Bohr read Kierkegaard, then Copenhell plays loud music') and negative relevance conditionals (e.g., 'If Niels Bohr read Kierkegaard, then Niels Bohr was ignorant of who Kierkegaard was'), it was found in Skovgaard-Olsen et al. (2016) that "the Equation" breaks down for the latter stimulus materials.

The dialectic in Chapter 2 takes the following form: there is a challenge of how to reconcile the Relevance Effect with existing theories of meaning of conditionals employed in psychology and neighbouring disciplines. The most conservatory response is not to revise the semantic theories but to extend them with auxiliary principles concerning pragmatic enrichment of linguistic content, whereby people take contextual factors into account (concerning, e.g.,

[12] Instead, Adams (1998) formulates *Ersatz* truth values for conditional events, which coincide with the de Finetti truth table that we will encounter in Chapter 5 in that the conditional event has the same truth evaluation as the consequent, iff the antecedent is true, and lacks a truth value, iff the antecedent is false. Concerning these *Ersatz* values, Adams (p. 65) notes that they are "truth-values in name only, and there is no suggestion that this kind of 'truth' is something that should be aimed at in reasoning, or that it is better than falsehood".

[13] A better name is the 'probability conditional theory', as suggested by Adams (1998, Ch. 6), but the "the Equation" is commonly used in the literature and so will be adopted below.

speaker intentions, situational knowledge, or the information state of the interlocutors) to interpret utterances. In fact, as we shall see, this is also the route that has been taken by mental model theorists and proponents of the Suppositional Theory of conditionals in psychology.

To advance the debate, Chapter 2 identifies a number of criteria from linguistics to decide whether a meaning component belongs to pragmatic or semantics and applies these diagnostic tests in a range of experiments. In particular, Chapter 2 focuses on designing experiments that are diagnostic for presuppositions, conversational and conventional implicatures, which are three well-known types of contents from philosophy of language and linguistics at the interface between semantics and pragmatics (Carston, 2002; Grice, 1989; Potts, 2007; Potts, 2015).

In Chapter 4, a related investigation is made for another central aspect of conditional reasoning; the distinction between indicative conditionals, "if *p* then *q*" (used, e.g., for prediction and argumentation), and counterfactuals, "if *p* had been the case, then *q* would have been the case" (used, e.g., in explanation or causal inference). Again, here we have a situation where there on the surface appears to be obvious meaning differences between a set of sentences, and the question is whether uniform semantic theories are to be favoured, which attempt to account for them by assigning the same truth-conditions to the sentences in question, while using auxiliary pragmatic principles to account for meaning differences, or whether the sentences also differ in their semantic core. Through two experiments, we probe two salient hypotheses concerning the meaning differences between indicative conditionals and counterfactual conditionals and find that only one of them is supported by the results. The two hypotheses differ on whether the implied falsity of the antecedent in counterfactuals is to be explained by conversational implicatures or presuppositions. Our empirical results and theoretical arguments indicate that the former is the more promising candidate. If this pragmatic account of the difference between indicative and counterfactual conditionals can be upheld, then one of the stumbling blocks for a uniform semantics for these two forms of conditionals has been removed.

The above suggests that one method for investigating semantic content is to experimentally probe which truth conditions ordinary speakers assign to them. This idea has indeed been a driving force behind use of the truth-table task in psychology (see, e.g., Evans & Over, 2004; Manktelow, 2012), where participants are presented with truth-table cells and asked to assign truth values to conditionals and connate sentences. Chapter 5 extends this line of research by presenting a new truth-table task to empirically investigate the possible-worlds semantics of Lewis (1973) and Stalnaker (1968). Although this theory has been highly influential in philosophy and linguistics through its semantics for modal expressions and counterfactuals (Portner, 2009; Kratzer, 2012; Lassiter, 2017; Starr, 2019), possible-worlds semantics is rarely investigated empirically in psychology. Chapter 5 reviews some of the reasons for why this is so and attempts to develop an experimental paradigm, where we can sidestep some of the controversial issues relating to the measurement of the distance between possible worlds. Through this task, we probe the theory's ability to account for participants' truth evaluations of indicative conditionals and counterfactuals in direct competition with six other theories. Since the possible-worlds semantic of Stalnaker (1968) specifies truth conditions that apply to both indicative conditionals and counterfactuals, the results supporting the latter theory in Chapter 5 are complemented by the results from Chapter 4 that some of the differences in interpretation of these two types of sentences can be accounted for pragmatically via conversational implicatures.

Aside from direct investigations of truth-value assignments, a method for investigating semantic content favoured by linguists consists in examining whether an aspect of the meaning of a linguistic expression affects its entailments (see, e.g., Winter, 2016; Grandy & Osherson, 2010, p. 287). Consequently, there is a need in psychology to design experimental tasks for examining entailments directly to tests its semantic theories. Chapter 3 takes up this challenge by its attempt to develop a new entailment task.

1.2.2 From Reasoning to Argumentation

Both the entailment task presented in Chapter 3 and the cancellation task in Chapters 2 and 4 place the evaluations of the linguistic content in an argumentative context. This focus on reasoning in an argumentative setting runs throughout this book and becomes a central theme in the individual profiling of participants' norm adherence in Part III of the book.

The underlying motivation for investigating reasoning competence in an argumentative setting in these experiments comes from Skovgaard-Olsen (2015). In this paper, it was argued that the types of idealizations in rationality theories that we encountered in section 1.1.1 are better handled, if the requirements of rationality are reinterpreted as requirements on argumentation. For its normative theories, the psychology of reasoning tends to import theories from formal epistemology (see, e.g., Pfeifer & Douven, 2014). Yet, both theories based on deductive logic and Bayesian models from formal epistemology assume the following minimal requirements of rational beliefs (Spohn, 2012; Huber, 2013):

(I) Rational beliefs are deductively closed.

(II) Rational beliefs are completely consistent.

(III) Every logically equivalent sentence is always believed to the same degree by the rational agent.

That these principles are too demanding in that they presuppose that rational agents are logically omniscient is what has been known as the *problem of logical omniscience* (Stalnaker, 1999; Levi, 1991, 1997). In this context, it is useful to return to principles i) - iii) from section 1.1.1, which Adams (1998, p. 21) uses to state the Kolmogorov Axioms of probability theory (here formulated for formulas of a formal language rather than for sets):

K1. $0 \leq p(\varphi) \leq 1$

K2. If φ is logically true then $p(\varphi) = 1$.

K3. If φ logically implies ψ then $p(\varphi) \leq p(\psi)$

K4. If φ and ψ are logically inconsistent then $P(\varphi \vee \psi) = p(\varphi) + p(\psi)$.

Concerning these, Adams (ibid., p. 22) makes the following remark:

> But the Kolmogorov axioms only concern 'static probabilities', that apply to propositions at one time, and beyond that they idealize by presupposing that all of the purely logical properties and relations among the propositions they apply to are *known*. Thus, if φ is a logical truth that is assumed to be *known*, and in those circumstances $p(\varphi)$ should equal 1. Similarly, if φ logically implies ψ that is assumed to be *known*, and therefore $p(\varphi) \leq p(\psi)$. [emphasis added, NSO]

Borrowing a familiar tactics from formal epistemology (see, e.g., Spohn, 2012, Chap. 2-4), the adherent of the probabilistic paradigm need not require that the ideal Bayesian agent already has assigned a probability representing a degree of belief to every formula that can be expressed by the formal language. It is sufficient if these degrees of belief are dispositional degrees of belief consisting of dispositions to assign coherent probabilities, if presented with arbitrarily complex formula generated by the language. For this purpose, Bayes nets may be utilized, as noted earlier. Another route is to follow Skovgaard-Olsen (2015) in reinterpreting the requirements of rationality as requirements on public commitments in argumentative discourse rather than as principles that participants follow in their individual reasoning, which then leads to designing argumentative tasks for reasoning studies.

A good example is when the Dialogical Entailment Task in Chapter 3 presents participants with an argumentative exchange between two persons, where the first person asserts the premise of a supposed entailment and then denies its conclusion. The other person then points out that the speaker has said two things that cannot both be true, and the task of the participants is to decide whether they agree or disagree with this criticism. In addition to placing the evaluation of entailments in an argumentative context, one of the innovations of this task is to present the inference problem in a format that specifically targets entailment relations (for which no models exist that satisfy all the premises, but which fail to satisfy the conclusion). If, on the contrary, we had merely asked whether the conclusion "follows from" the premises, as sometimes suggested, then the outcome would be ambiguous between logical entailment, probabilistic inference, plausible inference, degree of confirmation, and pragmatic implicature.[14]

The investigation of deductive inferences is by no means new to the psychology of reasoning. Yet, the topic of deductive reasoning has grown controversial due to the developments in the psychology of reasoning reviewed in section 1.1, which are elaborated further in Chapter 3. The experimental paradigm developed in Chapter 3 is a direct response to the problem of the underexplored deductive competence presupposed by the probabilistic paradigm identified above (section 1.1.1). Chapter 3 takes up this challenge by its attempt to develop an entailment task to test participants' acceptance of deductive relations and their probabilistic coherence vis-à-vis the accepted entailments.

1.2.3 Conditionals and Causality

In Chapter 6, the large issue of the relationship between conditionals and causality is investigated. The background is several recent attempts to interpret the Relevance Effect as an effect of causal reasoning (van Rooij & Schulz, 2019; Oaksford & Chater, 2020a, 2020b).

[14] Evans (2002, p. 992) suggests that 'what follows from' is preferable to 'what necessarily follows', if researchers are mainly interested in pragmatic and probabilistic reasoning rather than deductive reasoning. But the fine distinctions between probabilistic inference, plausible inference, degree of confirmation, and pragmatic implicature are not targeted by this dependent variable. For this reason, it would be preferable if psychologists developed separate tasks for each of the five inference categories listed above.

In van Rooij and Schulz (2019), this takes the form of applying a measure of causal power by Cheng (1997) and others, which has been influential in the psychology of causality (Waldmann, 2017). Under certain boundary conditions, causal power is able to predict the Relevance Effect (van Rooij & Schulz, 2019) and Chapter 6 therefore investigates through several experiments whether this prediction and its auxiliary assumptions hold.

In a constructive endeavour, the theory of a hierarchy of causal queries by Pearl (2009) is applied to conditional reasoning in Chapter 6. On Pearl's theory, there is a fundamental difference between the computational models needed to answer predictive queries, interventionalist queries, and counterfactual queries. Since indicative and counterfactual conditionals are often used to formulate such queries in natural language, it is investigated through 6 experiments which levels in this causal hierarchy maps onto indicative and counterfactual conditionals. In Gerstenberg (2022), further results are presented that directly build on this investigation.

1.3 Part III Norm Conflicts and Individual Differences

Whereas detailed investigations into the psychology of reasoning based on contrasting single norms has mainly been the focus of the chapters in Part II (with one slight exception in Chap. 5), the chapters in Part III differ by exploring the hypothesis that participants could be following different norms.

1.3.1 The Symphony of Many Voices

Chapters 3, 5, 7, and 8 employ Bayesian mixture modelling to investigate individual variation. One way to illustrate the need for mixture modelling is to consider a case, where one is modelling the signal not just from a single instrument but from a whole collection of instruments (or even an orchestra). In this case, the inferential problem is not just one of separating the signal from the noise but separating different signals from each other while factoring out the noise. Each signal may have its own Gaussian distribution and so the population of signals can be considered a mixture distribution with several components that receive different weights.[15]

[15] This example was used by Serrano G. Luis in his online lectures on statistical models: https://www.youtube.com/c/LuisSerrano

In the chapters listed above, mixture models are applied to competing semantic theories of the same linguistic expressions. This is done to empirically investigate the following commonly held assumption:

> (U) There is a uniform interpretation of any given linguistic expression, φ. If several semantic theories for φ exist and they are incompatible, then at most one of them can be descriptively adequate.

(U) dates back (at least) to the beginning of formal semantics in the writings of Gottleb Frege (1848-1925) and is the default assumption in psychology of reasoning, formal semantics in linguistics, and philosophy of language.

On Frege's (1892, 1918) semantic analysis, the different parts of a sentence contribute compositionally to compose a thought expressed by the sentence. 'Thoughts' is here a technical term for the objective, propositional content expressed by a sentence, which has a truth-value (true vs. false) independently of whether the sentence is used or the thought grasped by a person. While the mental representations of a person are subjective, they are capable of having a truth-evaluable *content* that is objective and the basis of logic. On Frege's model of communication, intersubjective communication is made possible by the speaker grasping a thought, which is expressed by the utterance of the sentence, and the listener grasping the thought by hearing the utterance. The model assumes that each sentence expresses one and only one thought and that it is the same thought that is grasped by both the hearer(s) and the speaker (Newen, 2001).

Since different semantic theories of φ usually specify different truth conditions for φ, the different theories imply that different propositional contents are expressed by the sentences in which φ occurs. But then it is no longer possible for the speaker and hearer to grasp the same thought, or content, if they apply different semantic interpretations to the same sentence. Hence, communication is rendered impossible, on this Fregean model.[16] For this reason, (U) is a natural assumption to make.

[16] Incidentally, Newen (2001) argues that Frege himself would only apply this model of communication to ideal languages and that he realized that natural languages fall short of the ideal. Nevertheless, this model of communication has become associated with Frege by other authors who applied it to natural language and so we will continue to use the label here.

That (U) is the standard assumption is, e.g., seen by the way that competing semantic theories are discussed in textbooks (e.g., Bennett, 2003; Kadmon, 2001; McCawley, 1993; Portner, 2009), whereby examples covered by one theory and unaccounted for by a second theory is taken as evidence for the hypothesis that the first theory characterizes our linguistic competence and as evidence against the second theory. In disputes among proponents of the contrasting theories, theoreticians on each side will appeal to intuitive examples explained by their own theory, which are unaccounted for by the theories of their competitors.

The alternative is to view the fact that there are longstanding disputes about the meaning of the same terms as itself *prima farcie* evidence that there can be individual variation in how sentences are interpreted. In several of the chapters that follow, we will return to the empirical assessment of (U) for different domains. Examples are conflicting theories of indicative conditionals (Chap. 5 and 7) and epistemic modals (Chap. 8). In each case, patterns of individual variation are investigated experimentally.

In the context of epistemic modals, we will return to the Fregean question of how communication can be possible, if interlocutors apply different semantic interpretations to the same linguistic expressions. In Newen (2001), it is already shown in the context of sentences containing indexicals (e.g., '*I* was wounded') that it is problematic to assume that communication is only possible, if the speaker and listener grasp the same thought. Instead, Newen argues that the dogma that one unambiguous, utterance expresses exactly one semantic content should be overturned. In Chap. 2, further work in linguistics is introduced (Bach, 1999; Potts, 2005), which likewise challenges the notion that each sentence only expresses one proposition (e.g., by considering examples with appositives like 'Mozart, *the famous composer*, used to live here'). But the challenge we will consider in Chap. 8 is different. It concerns how different speakers can engage in argumentation, if they interpret common expressions of epistemic uncertainty, like 'might', differently, as our empirical findings appear to show.

1.3.2 The Problem of Arbitration

The empirical investigations in part III aim to investigate which norms ordinary subjects enforce on their peers in argumentative settings. Previous work has investigated rationality and individual variation (Stanovich, 1999; Stanovich & West, 2000) and probabilistic renderings of informal arguments (see, e.g., Oaksford & Hahn, 2004; Hahn & Oaksford, 2007; Fenton et al., 2013). But whereas much rationality research has proceeded by designing experiments based on one preconceived norm, the goal of the chapters that follow is to put ordinary subjects as normative agents on the agenda to address the question of which norms they recognize and impose on their peers to study individual variation in norm adherence.

The fundamental problem is well-stated in the following passage, which is worth quoting in full length to show that its formulation is older than its most recent discussion:

> In early research on the psychology of reasoning, experimenters counted as error any deviation from a standard logical system—generally Scholastic doctrine in the case of syllogisms or classical sentential logic in the case of sentential arguments. However, if an "error" of this sort is something for which we can actually criticize a subject, rather than just a convenient label for classifying the subject's responses, then the logical systems themselves must be correct norms of appraisal. Some early researchers may have accepted the standard systems as normatively appropriate simply out of ignorance of rival ones, but the variety of contemporary logics raises questions about the correctness of Scholastic and classical logic. Inferences that follow the pattern of Double Negation Elimination, for examples, are valid in classical logic but not in intuitionistic logic (…). Similarly, the syllogism <All(G, H), All(G, F), ∴ Some(F,H)> is valid according to Scholastic logic but not according to classical logic (…). This means that attributing errors to subjects can be a delicate matter that depends on how we justify the logical systems themselves. (Rips, 1994, p. 378)

This problem of arbitration is a central problem in the application of formal, normative theories in the psychology of reasoning, which is receiving increasing emphasis (see, e.g., Stenning & van Lambalgen, 2008; Schurz & Hertwig, 2019; Knauff & Spohn, 2021).

Elqayam and Evans (2011) use this problem to argue that the psychology of reasoning should eschew normative theorizing and make a fresh start as a purely descriptive discipline. However, this is a view that has been met with heavy resistance by commentators (see, e.g., the commentaries in the volume Elqayam & Over, 2016). Underlying this resistance is the insight that normative theories have been a major impetus for theory development in psychology spanning the areas from decision theory (Wakker, 2010) and causal judgment (Rottman & Hastie, 2014) to the psychology of reasoning (Oaksford & Chater, 2007).

Chapter 7 features a discussion of the various arguments used by Elqayam and Evans (2011) to motivate a descriptive turn in psychology. The constructive contribution of Chapter 7 is to present a new, experimental approach to the problem of arbitration.

Aside from his clear statement of the problem, there is another reason to cite Rips (1994) in the present context. Rips considers the possibility of building non-classical logic into PSYCOP (p. 124), has deliberately omitted inference rules for PSYCOP that would make its conditional equivalent to the material implication of classical logic, and considers the possibility of including rules that would make its conditional the intensional conditional of Lewis and Stalnaker (p. 48), which we investigate in Chapter 5. To guide the choice of which rules PSYCOP should have, Rips adopts the criterion of psychological completeness from Osherson (1975) and posits that the model should capture the inferences that "untrained people can accept in ideal circumstances" (p. 124).

Entangled in these discussions are different views on the normativity of logic. In the discussion of the function of logic in the study of human reasoning, Steinberger (2019) argues that at least three normative roles have to be distinguished – directive, evaluative, and appraising. A chess analogy can be used to bring this distinction into focus. A *directive* use of chess theory consists in the formulation of guidelines of how a chess player can improve herself given her present level of understanding. An *evaluative* use of chess theory consists in the investigation of optimal play in a given position, which disregards constraints set by the player's limited understanding. An *appraising* use of chess theory departs from the players' expected level of competence, given slight idealizations of her actual play (to exclude minor performance errors that she understands were mistakes), and criticizes or praises her play based on what chess theory would recommend at her given level of competence.

Under this framing, one pressing issue is why the psychology of reasoning should occupy itself with appraising uses of normative systems, as opposed to mere evaluative uses, and if it does, then which formal systems should be employed. Key to this question is that appraisals play a role in attributing reasoning errors. In this context, it is worth noticing the following: when Rips (1994, p. 378) talks about the difficulty in attributing errors to participants given the possibility of multiple conflicting norms, he makes explicit in the quote above that the norms he has in mind are "correct norms of appraisal", where an error "is something for which we can actually criticize a subject". Correspondingly, in their classical work on the conjunction fallacy, Tversky and Kahneman (1983) treated judgments as fallacies only when participants were disposed to accept (after suitable explanation) that they had made a non-trivial, conceptual error; an error which the participants had the competence to avoid. In this notion of fallacy, the participants' own understanding of their performance plays a central role in the assessment. This is characteristic of appraising uses of normative systems according to Steinberger's taxonomy.

Although criticism and praise in appraisals are value-laden terms, we can view them as highlighting distinctions in performance that an agent with a given level of competence would make. Psychology of reasoning need not pursue value-laden interests for this to have utility. Empirically identifying aspects of an agent's performance which could give rise to criticism or praise by assessing her performance based on a given level of assumed competence serves a descriptive function in disclosing what level of competence the agent has. Being sensitive to distinctions that a competent agent would make involves psychological mechanisms and is as such amenable to scientific research.

For the evaluative comparison of which formal system is the most optimal way of reasoning, different methods have been invoked including *a priori* arguments and comparing formal systems based on a theory-neutral notion of cognitive success involving truth conduciveness and computational costs (Schurz, 2014). To illustrate, Table 3 outlines means-end relations for formal systems as they are presented in Huber (2014).

Table 3. Means-End Relations for Formal Systems

Means	Cognitive End
Probability Theory	Inaccuracy minimization in the organization of credences at a given moment in time. (*Or*: Avoiding sure loss in the combination of bets described by a Dutch Book.)
Ranking Theory	Holding beliefs that are jointly consistent and deductively closed synchronically and diachronically after updating beliefs on evidence.
Classical Logic	Truth preservation for inferences in all logically possible worlds.
Non-monotonic Logic	Truth preservation for inferences in the subset of all possible worlds that are normal.

Note. Illustration of objective means-end relations for four formal, normative theories, which justify hypothetical imperatives that prescribe which means to select given the participants have a particular cognitive aim.

Achourioti, Fugard, and Stenning (2014) argue that the formal system for attributing reasoning errors should be selected instrumentally, such that error attribution is to be based on the system that best suits the goals of participants in the specific reasoning task they are engaged in. Achourioti *et al.* (2014) therefore advocate for reasoning researchers to corroborate their attributions of reasoning errors based on independent evidence concerning the participants' individual goals, and their implicit understanding of the logical concepts required to meet those goals. Instead of appraising the participants' performance based on a single norm established *a priori*, researchers should accordingly try to construct a charitable interpretation based on the participants' understanding of the task (see also Elqayam, 2012; C. J. Lee, 2006)

In an attempt to develop an experimental approach for dealing with the problem of arbitration, Chapter 7 presents a task, "the Scorekeeping Task", for implicitly eliciting participants' recognition that they adhere to one of several conflicting norms. This paradigm also applies to norms of appraisal and tries to elicit participants' reflective attitudes about which norms they would apply, instead of committing them based on either explicit avowals or their immediate reactions only. In Chapter 7, this approach is applied to participants' reasoning with conditionals ('if *p* then *q*') under uncertainty, in continuation with Part II of this book. In this case, the two norms amount to two different interpretations of indicative conditionals. On the one hand, participants are classified as following the Suppositional Theory of conditionals building on Adams' (1998) work. On the other, participants are classified as following Inferentialism, which regards the Relevance Effect as an effect of semantics rather than pragmatics.

In Chapter 8, the approach is applied to participants' reasoning with epistemic modals. Epistemic modals (e.g., might, must) is another important class of linguistic expressions to express states of uncertainty. In addition, they also make up a suitable test case for the experimental approach to the problem of arbitration developed in Chapter 7, since the recent literature in linguistics and philosophy of language has developed a collection of competing theories (see, e.g., Egan & Weatherson, 2011). For the experiments in Chapter 8, we focus on three such theories, Contextualism, Relativism, and Objectivism. As we show, a reanalysis of previously published findings indicate that the previous results, which neither favored Contextualism nor Relativism univocally, could be the result of individual variation. The goal of Chapter 8 is therefore to apply the Scorekeeping Task from Chapter 7 and investigate whether there is individual variation in the interpretations of epistemic modals that participants follow.

Given that Chapters 7 and 8 have shown how the problem of arbitration can be dealt with experimentally in two core applications to reasoning under uncertainty with conditionals and epistemic modals, respectively, a natural question is whether the approach could be applied in other domains as well. In Chapter 9 some first steps towards such an application is taken, when we finally turn to the opposition between Utilitarianism and Deontology in the psychology of morality. Before individual variation in participants' adherence to these two moral outlooks can be further examined, a measurement model of participants' moral judgments first needs to be validated. As part of such an endeavor, Chapter 9 examines an invariance assumption in the CNIS model of Gawronski et al. (2017). The CNIS model is a process-dissociation model, which attempts to remove certain confounds in the traditional trolley-based investigations of Utilitarianism and Deontology. However, like other process-dissociation models applied in psychology, the model makes an invariance assumption which has been found to be problematic for similar models in social psychology and cognitive psychology (Klauer et al., 2015). The goal of Chapter 9 is therefore to empirically investigate whether the invariance assumption is violated in the CNIS model of moral judgment, and if so, develop an improved model, which can fulfill some of the original aims of the CNIS model without enforcing the invariance assumption.

As indicated, this investigation in Chapter 9 is preparatory for further investigations into individual variation in participants' adherence to conflicting norms in moral psychology, along the same lines as Chapters 7 and 8 in the domain of reasoning.

References

Achourioti, T. Fugard, A. J. B., and Stenning, K. (2014). The empirical study of norms is just what we are missing. *Frontiers in Psychology, 5*, 1159.

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

-- (1998). *A Primer of Probability Logic*. Stanford, CA: CLSI publications.

Bach, K. (1999). The Myth of Conventional Implicature. *Linguistics and Philosophy, 22*(4), 327-366.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bermúdez, J. L. (2010). *Cognitive Science*. New York: Cambridge University Press.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1-21.

Braine, M. D. S., and O'Brien, D. P. (1991). A Theory of *If*: A Lexical Entry, Reasoning Program, and Pragmatic Principles. *Psychological Review, 98*(2), 182-203.

Brandom, R. (1994). *Making it Explicit*. Cambridge, MA: Harvard University Press.

Carston, R. (2002). *Thoughts and utterances.* Oxford: Blackwell Publishers.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, *104*(2), 367–405.

Darwiche, A. (2009). *Modelling and reasoning with Bayesian networks*. New York: Cambridge University Press.

Edgington, D. (1995). On Conditionals. *Mind*, *104*, 235-327.

Egan, A. and Weatherson, B. (Eds.) (2011). *Epistemic Modality*. Oxford: Oxford University Press.

Elqayam, S. (2003). Norm, error, and the structure of rationality: The case study of the knight-knave paradigm. *Semiotica*, *147*, 1/4, 265–289.

-- (2012). Grounded Rationality: Descriptivism in epistemic context. *Synthese*, *189*, 39-49.

Elqayam, S. and Evans, J. St. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*, 233-90.

Elqayam, S. and Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning*, *19* (3-4), 249-265.

-- (Eds.) (2016). *From is to ought: The place of normative models in the study of human thought*. Lausanne: Frontiers Media.

Elqayam, S., Bonnefon, J.-F., Over, D. E. (2016) (Eds.). *New Paradigm Psychology of Reasoning. Basic and applied perspectives.* New York: Routledge.

Eva, B., and S. Hartmann (2018b). Bayesian Argumentation and the Value of Logical Validity. *Psychological Review*, *125*(5), 806–821.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978-996.

-- (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.

-- (2009). Does rational analysis stand up to rational analysis? *Behavioral and Brain Sciences*, *32*(1), 88–89.

Evans, J. St. B. T. and Over, D. (2004). *If*. Oxford: Oxford University Press.

Fenton, N., Neil, M., and Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*, 61-102.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

-- (1983). *The Modulairty of Mind*. Cambridge, MA: The MIT Press.

Fodor, J. A., and Z.W. Pylyshyn. 1988. Connectionism and Cognitive Architecture: A Critical Analysis, *Cognition*, *28*, 3-71.

Frege, G. (1892), 'Über Sinn und Bedeutung'. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50.

--- (1918), 'Der Gedanke. Eine Logische Untersuchung'. In: *Beiträge zur Philosophie des deutschen Idealismus*, I (1918–1919) (pp. 58–77).

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hutter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113,* 343–376.

Gerstenberg, T. (2022). What would have happened? Counterfactuals, hypotheticals and causal judgements. *Phil. Trans. R. Soc. B., 377*, 20210339. https://doi.org/10.1098/rstb.2021.0339

Gilovich, T., Griffin, D. W., and Kahneman, D. (2002) (Eds.). *Heuristics and Biases. The Psychology of Intuitive Judgments*. New York: Cambridge University Press.

Grandy, R. and Osherson, D. (2010). *Sentential Logic for Psychologists*, accessed 10 Obtober 2022, <http://www.princeton.edu/~osherson/primer.pdf>.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Hahn, U. and Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review*, *114*(3), 704-732.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: The MIT Press.

Hartmann, S. (2021). Bayes Nets and Rationality. In Knauff, M. & Spohn, W. (eds.), *Handbook of Rationality* (pp. 253-264). London: The MIT Press.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271.

Huber, F. (2013). Formal representations of belief. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2013 Ed.). http://plato.stanford.edu/archives/sum2013/entries/formal-belief/.

-- (2014). New Foundations for Counterfactuals. *Synthese*, 191, 2167-93.

-- (2018). *Logical Introduction to Probability and Induction*. New York: Oxford University Press.

Johnson-Laird, J. (1997). Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*. *Minds and Machines, 7*, 387-407.

Kadmon, N. (2001). *Formal Pragmatics*. Malden, MA: Blackwell Publishers.

Khali, E. L. (2009). Are stomachs rational? *Behavioral and Brain Sciences*, *32*(1), 91-92.

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221.

Knauff, M. (2007). How our brains reason logically. *Topoi*, *26*, 19-36.

Knauff, M. and Gazzo Castañeda, L. E. (2021). When nomenclature matters: is the "new paradigm" really a new paradigm for the psychology of reasoning? *Thinking & Reasonng*, DOI: 10.1080/13546783.2021.1990126

Knauff, M. and Spohn, W. (2021). Psychological and Philosophical Frameworks of Rationality—A Systematic Introduction. In Knauff, M. & Spohn, W. (eds.), *Handbook of Rationality* (pp. 1-65). London: The MIT Press.

Kratzer, A. (2012). *Modals and Conditionals: New and Revised Perspectives*. New York: Oxford University Press.

Lantz, B. (2019). *Machine Learning with R.* Birmingham: Packt Publishing.

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives.* Oxford: Oxford University Press.

Lee, C. J. (2006). Gricean Charity: The Gricean Turn in Psychology. *Philosophy of the Social Sciences*, 36, 193-218.

Leitgeb, Hannes (2016): Probability in Logic. In Hájek, Alan, Hitchcock, Christopher (Eds.), *The Oxford handbook of probability and philosophy* (pp. 681-704). Oxford: Oxford University Press.

Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.

-- (1997). *The covenant of reason*. Cambridge: Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

-- (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*, 297-315.

MacFarlane, J. (draft). In what sense (if any) is logic normative for thought? http://johnmacfarlane.net/work.html.

Macnamara, J. (1986). *A Border Dispute. The Place of Logic in Psychology*. Cambridge, MA: The MIT Press.

Manktelow, K. (2012). *Thinking and Reasoning: An Introduction to the Psychology of Reason, Judgment and Decision Making*. Sussex: Psychology Press.

Miller, J. D. & Forte, R. M. (2017). *Mastering Predictive Analytics with R. Machine learning techniques for advanced models* (2nd Edition). Birmingham: Packt Publishing.

Newen, A. (2001). Fregean Senses and the Semantics of Singular Terms. In: A. Newen & U. Nortmann & R. Stuhlmann-Laeisz (Eds.), *Building on Frege. New Essays on Sense, Content, and Concept* (pp. 113-140). CSLI.

Oaksford, M. and Hahn, U. (2004). A Bayesian Approach to the Argument from Ignorance. *Candadian Journal of Experimental Psychology*, *58*(2), 75-85.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian Models. *Frontiers in Psychology, 5* (332), 1-5.

Oaksford, M., and Chater, N. (1991). Against Logicist Cognitive Science. *Mind & Langauge, 6*(1), 1-38.

-- (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

-- (2009). Précis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Behavioral and Brain Sciences, 32*(1), 69-119.

-- (2020a). New Paradigms in the Psychology of Reasoning. *Annual Review of Psychology*, *71*(1), 1–26.

-- (2020b). Integrating Causal Bayes Nets and Inferentialism in Conditional Inference. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 116-132). London: Routledge.

O'Brien, D. P. (2009). Human reasoning includes a mental logic. *Behavioral and Brain Sciences, 32*(1), 96–97.

Osherson, D. N. (1975). *Logical Abilities in Children*, volume 3. Erlbaum.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Peterson, M. (2017). *An Introduction to Decision Theory (Second Edition)*. Cambridge: Cambridge University Press.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective, *Thinking & Reasoning*, *19*(3-4), 329-345.

Pfeifer, N., and Douven, O. (2014). Formal epistemology and the new paradigm psychology of reasoning. *Review of Philosophy and Psychology*, *5*(2), 199–221.

Pfeifer, N., and Kleiter, G. D. (2005). Towards a mental probability logic. *Psychologica Belgica*, *45*, 71–99.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

-- (2007). Conventional implicatures, a distinguished class of meanings. In Gillian Ramchand and Charles Reiss (Eds.). *The Oxford handbook of linguistic interfaces* (pp. 475–501). Oxford: Oxford University Press.

-- (2015). Presupposition and implicature. In Shalom Lappin, & Chris Fox (Eds.). *The handbook of contemporary semantic theory (2nd Ed.)* (pp. 168–202). Oxford: Wiley-Blackwell.

Pylyshyn, Z. W. (1989). Computing and cognitive science. In Michael I. Posner (Eds.). *Foundations of Cognitive Science* (pp. 51–91). Cambridge: The MIT Press.

Rips, L. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: The MIT Press.

-- (1997). Goals for a Theory of Deduction: Reply to Johnson-Laird. *Minds and Machines, 7,* 409–424.

Rottman, B. M., and Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*, 109–139.

Schurz, G. (2014). Cognitive success: instrumental justifications of normative systems of reasoning. *Frontiers in Psychology*, 5, 1-16.

Schurz, G. and Hertwig, R. (2019). Cognitive Success: A Consequentialist Account of Rationality in Cognition. *Topics in Cognitive Science*, 1-30.

Skovgaard-Olsen, N. (2015). The problem of logical Omniscience, the preface paradox, and doxastic commitments. *Synthese*, *194*(3), 917-939.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, 150, 26-36.

Spohn, W. (2012). *The Laws of Beliefs*. Oxford: Oxford University Press.

Stalnaker, R. C. (1968). A Theory of Conditionals. In: Rescher, N. (Eds.), *Studies in Logical Theory* (pp. 98-112). Oxford: Basil Blackwell.

-- (1999). *Context and content*. Oxford: Oxford University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* New York: Psychology Press.

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665.

Steinberger, F. (2019). Three ways logic may be normative. *The Journal of Philosophy, 116*(1), 5-31.

Stenning, K., and van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

-- (2009). ”Nonmonotonic” does not mean “probabilistic”. *Behavioral and Brain Sciences*, *32*(1), 102–103.

Starr, W. (2019). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). Retrieved from forthcoming URL = <https://plato.stanford.edu/archives/fall2019/entries/counterfactuals/>.

Tentori, K., Bonini, N., Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science, 28,* 467–477.

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.

van Rooij, R., and Schulz, K. (2019). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

Vance, J., and Oaksford, M. (2020). Explaining the implicit negations effect in conditional inference: Experience, probabilities, and contrast sets. *Journal of Experimental Psychology: General, 150*(2), 354-384.

Wakker, P. P. (2010). *Prospect Theory*. *For Risk and Ambiguity*. Cambridge: Cambridge University Press.

Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. Oxford: Oxford University Press.

Wernick, William (1942). Complete Sets of Logical Functions. *Transactions of the American Mathematical Society*. *51*, 117–32.

Winter, Y. (2016). *Elements of Formal Semantics*. Edinburgh: Edinburgh University Press.

Part II
Psychology of Reasoning: The Semantics and Pragmatics of Conditionals

Chapter 2:
Cancellation, Negation, and Rejection[17]

Niels Skovgaard-Olsen,
Peter Collins,
Karolina Krzyżanowska,
Ulrike Hahn, Karl Christoph Klauer

In this paper, new evidence is presented for the assumption that the reason-relation reading of indicative conditionals ('if A, then C') reflects a conventional implicature. In four experiments, it is investigated whether relevance effects found for the probability assessment of indicative conditionals (Skovgaard-Olsen, Singmann, and Klauer, 2016a) can be classified as being produced by a) a conversational implicature, b) a (probabilistic) presupposition failure, or c) a conventional implicature. After considering several alternative hypotheses and the accumulating evidence from other studies as well, we conclude that the evidence is most consistent with the Relevance Effect being the outcome of a conventional implicature. This finding indicates that the reason-relation reading is part of the semantic content of indicative conditionals, albeit not part of their primary truth-conditional content.

[17] This chapter has been published as:

Skovgaard-Olsen, N., Collins, P., Krzyzanowska, K., Hahn, U., and Klauer, C. K. (2019). Cancellation, Negation, and Rejection. *Cognitive Psychology*, *108*, 42-71.

Authors' Note: This work was supported by grants to Wolfgang Spohn and Karl Christoph Klauer from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).
Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).
Supplementary Materials: https://osf.io/hz4k6/

2.1 Introduction[18]

Very few linguistic expressions are as important to our reasoning, argumentation, and decision making as indicative conditionals, that is, a class of sentences typically of the form: "If A, (then) C," where A, the *antecedent*, and C, the *consequent*, stand for arbitrary sentences. It is not surprising then that conditionals have been a subject of extensive research in philosophy, linguistics, computer science, and psychology. What is more surprising though is that, despite decades of multidisciplinary efforts to understand how people interpret indicative conditionals, many aspects of their meaning remain the matter of ongoing controversy.

Intuitively, whenever someone asserts a conditional they communicate that there is some sort of a relation between the content of that conditional's antecedent and consequent; that, for instance, the antecedent is a reason for the consequent, or that the consequent can be inferred from the antecedent. Take, for example, the following sentence:

(1) If more parents refuse to vaccinate their children, diseases such as measles and whooping cough will make a comeback.

Clearly, someone who asserts (1) seems to be expressing their belief about the connection between the anti-vaccination movement and the possible outbreak of infectious diseases. That a conditional conveys such a relationship is not controversial. But the status of this connection is one of the most contentious issues in the current debate on the meaning of indicative conditionals. At issue is whether the connection is part of the *semantics* or the *pragmatics* of the conditional. Roughly, the semantics of the conditional – sometimes referred to as the *core meaning* – is its literal, conventional, context-independent meaning. The pragmatics of the conditional is its non-literal, inferred, context-dependent meaning (we will elaborate on these definitions later).

The question of whether the connection is semantic or pragmatic has attracted such interest in large part because the issue forms a dividing line between theories of the

[18] **Acknowledgement:** We would like to thank our Editor, Art Markman, and reviewers for detailed comments that helped improve the paper. In particular, David Over, Florian Schwarz, and Sunny Khemlani deserve credit. Furthermore, we would like to thank audiences to talks at New Perspectives on Conditionals and Reasoning (Regensburg, 2018), Annual Meeting of New Frameworks of Rationality (Etelsen, 2018), What if, Scientific Retreat (Konstanz, 2018), ECAP9: European Congress of Analytic Philosophy (Munich, 2017), as well as Robyn Carston, John MarFarlane, Maria Biezma, and others, who were so helpful as to engage in discussions with us over these matters. Thanks also go to students helping us carrying out the experiments like Dennis Heiler.

conditional in psychology and philosophy. On the one hand, there are historically established theories in psychology, such as the Suppositional Theory and the Mental Models Theory, which take the connection to be pragmatic. On the other, there are recently revived 'inferentialist' accounts, which take the connection to be semantic. The opposition of these theories in psychology is an echo of similar debates in philosophy (for references, see Douven, 2015 and Skovgaard-Olsen, 2016).

In comparison, linguistic debates about conditionals have been more influenced by the possible-worlds semantics of Stalnaker (1968), Lewis (1973), and Kratzer (1986). In none of these theories is the connection between the antecedent and the consequent taken to be semantic. Rather what we get is roughly a description of the antecedent worlds in the context set, which are most similar to the actual world, stating that they are worlds in which the consequent is true (Biezma, 2014). As such, formal semantics sides with the psychological theories in denying that the connection between the antecedent and consequent is part of the core semantic meaning of conditionals. But in this paper, our focus is on the former division within theories of conditionals in psychology. See, however, Skovgaard-Olsen (*in review*) for more on the connections between the present discussion and related work in linguistics.

The Mental Model Theory is one of the most influential theories in the psychology of reasoning. It postulates that compound sentences, such as conditionals, refer to conjunctions of possibilities, where possibilities are understood epistemically as situations that are compatible with what is known (Khemlani, Byrne, & Johnson-Laird, 2018). Consequently, on the Mental Model Theory, interpreting a sentence amounts to constructing mental models that represent possible states of affairs that are compatible with that sentence, while what is impossible tends to be omitted (Johnson-Laird & Byrne 1991, 2002; Johnson-Laird, Khemlani, & Goodwin 2015). A fully fleshed out, explicit model of a conditional, if A then C, can be depicted then in the following way, where each row denotes a mental model of one possibility:[19]

A C
¬A ¬C
¬A C

Importantly, already Johnson-Laird and Byrne (1991, 2002), but also Khemlani et al. (2018), emphasized that many people do not immediately construct fully explicit models of a conditional. Instead they stop their model construction at the initial, abbreviated, *implicit*

[19] For consistency we use 'A' and 'C' throughout to refer to the antecedent and consequent of the conditional respectively.

model consisting of the representation of the possibility that both the antecedent and the consequent are true:

A C

. . .

The ellipsis signals that there are other possibilities, which could be evoked, if necessary.

It is important to note that such conjunctions of possibilities constitute the meaning of a compound sentence on the Mental Models Theory. More precisely, the Mental Model Theory holds that a sentence is true if all the corresponding models are possible. This makes the mental models of a conditional different than a material conditional, even though the explicit model bears a resemblance to the truth table for the material implication. As Khemlani et al. (2018) say:

> In the model theory, a conditional's meaning is not a material implication, not a conditional probability, not a set of possible worlds, and not an inferential relation. It is instead a conjunction of possibilities, each of which is assumed in default of information to the contrary. (p. 31)

Accordingly, Johnson-Laird et al. (2015) hold that "a basic conditional, 'if A then C', is true only if all three situations in its fully explicit models are possible: "possibly(A & C) & possibly(not-A & not-C) & possibly(not-A & C) and A & not-C is impossible" (p. 206).[20]

But what matters for present purposes is just that although Mental Model Theory does not treat the relation between the antecedent and consequent of a conditional as a part of its core meaning, proponents of Mental Model Theory acknowledge that conditionals are often interpreted as conveying that there is such a relation due to "modulating effects of semantics and pragmatics" (Johnson-Laird & Byrne, 2002, p. 651). More specifically, "the meaning of words, knowledge, and the conversational context can block the construction of models of possibilities, and they can add causal, spatiotemporal, and other relations between elements in models (Khemlani et al., 2018, p. 12-13). They argue that the context of an utterance (pragmatics), or semantic relationship between the content of the antecedent and the content of the consequent, may block the construction of a model that normally belongs to the core meaning of a sentence, or trigger the construction of a model that is not part of the core meaning of a sentence. As we will argue below, semantic and pragmatic modulation is not

[20] See Baratgin, Douven, Evans, Oaksford, and Politzer (2015) for a discussion of some challenges for the revised version of the theory, and Khemlani et al. (2018) for a response. See further the discussion in Hinterecker, Knauff, and Johnson-Laird (2016) and Oaksford, Over, and Cruz (2018).

sufficient to account for the complex data pattern that emerges out of the present study when taken together with other recent published results, however.

An alternative approach in the psychological study of conditionals stems from the so-called New Paradigm Psychology of Reasoning (Oaksford & Chater, 2007; Over, 2007), which emphasizes the role of uncertainty in human reasoning. Indicative conditionals on the New Paradigm are interpreted as probabilistic or *suppositional*. Although the term *Suppositional Theory* (ST) refers to a whole family of related views, they can all be construed as the formalizations of the Ramsey Test, which provides a procedure for fixing one's degree of belief in a conditional, and, by the same token, for determining whether an indicative conditional is acceptable (Ramsey 1929/1990, p. 155):

> If two people are arguing 'if [A] will [C]?' and are both in doubt as to [A], they are adding [A] hypothetically to their stock of knowledge and arguing on that basis about [C]: so that in a sense 'If [A], [C]' and 'if [A], [¬C]' are contradictories. We can say that they are fixing their degrees of belief in [C] given [A]. (Editorial changes preserve the consistency of notation)

Consequently, what the different versions of the Suppositional Theory have in common is their commitment to *The Equation,* according to which the probability of a conditional equals the conditional probability of that conditional's consequent given its antecedent (where 'A' and 'C' are restricted to atomic sentences, that is, they are not conditionals themselves): P(If A, then C) = P(C|A). The development of the Suppositional Theory that became particularly influential in the psychology of reasoning resulted from combining the Ramsey Test, and thus the Equation, with three-valued de Finetti's semantics. De Finetti treated conditionals as true when both of its clauses are true, and false when the antecedent is true but the consequent is false. When the antecedent is false, the truth value of a conditional is undetermined, "void." Conditionals with false antecedents can be compared to called-off bets: a bet that if you throw a fair coin it will land heads is neither won, nor lost – it is called off – when the coin is not thrown at all (see, e.g., Politzer, Over, & Baratgin, 2010). A more refined version of the de Finetti's system can be obtained by replacing the third, "void" value with the conditional probability itself (Jeffrey, 1991; see also Baratgin, Politzer, Over, & Takahashi, 2018; Kleiter, Fugard, & Pfeifer, 2018; Over & Cruz, 2018).

The Equation has received strong empirical support. Participants in reasoning experiments tend to evaluate the probability of a conditional by estimating the corresponding conditional probability (Evans, Handley, & Over, 2003; Fugard, Pfeifer, Mayerhofer, &

Kleiter, 2011; Oberauer & Wilhelm, 2003; Politzer et al., 2010). Yet, a recent study has challenged the generality of these results (Skovgaard-Olsen, Singmann, and Klauer, 2016a), as we shall see in more detail below.

Although the original phrasing of the Ramsey Test, with its focus on arguing about C *on the basis of* A, seems to capture the intuition that the antecedent of a conditional needs to be somehow relevant for the consequent, this is not true of the suppositional accounts. As long as the antecedent is possible, a true or even highly probable consequent will render the conditional acceptable. If we believe that Brexit is inevitable and that it is quite possible that there are at least some microorganisms living on some planets outside of our Solar System, we are committed to accepting the following missing-link conditional:

(2) If there is life on some extra-solar planet, then the UK will leave the European Union.

This is because "the UK will leave the European Union" was already part of our stock of beliefs, and it remains so upon expanding it by "there is life on some extra-solar planet." By contrast, for a person who deems it rather unlikely that there are any advanced alien civilizations, and that this likelihood will not increase were we to learn that there are planets that host some form of life, (3) does not appear acceptable after performing the Ramsey Test:

(3) If there is life on some extra-solar planet, then somewhere in the Universe there exists an advanced alien civilization.

Yet it would be easier to make sense of a speaker who asserts (3), even if we strongly disagree with it, than of someone who asserts (2) (Krzyżanowska, 2015, p. 9). If we disagree with (3), it would be because learning that there is life on some extra-solar planet would not be a good enough reason for us to believe in the existence of an advanced alien civilization. Nevertheless, since the truth of the antecedent of (3) slightly increases the probability of its consequent, we can imagine someone who would find such an argument convincing enough to accept the conditional.

Psychologists of reasoning who claim that the Equation captures a central part of the meaning of an indicative conditional do not deny that these sentences often seem to suggest stronger (e.g., causal or inferential) relations between their antecedents and consequents. What they do deny is that these relations belong to the semantics, that is, to the core meaning of the conditional. It seems that when researchers wish to account for our intuitions about the relation between a conditional's antecedent and consequent, and for data on the actual use of conditionals, they most commonly invoke pragmatics (Johnson-Laird & Byrne, 2002; Over et al., 2007; for discussion, see, also Krzyżanowska, Collins, & Hahn, 2017a, 2017b, and below). But pragmatics is not the only option. The alternative approach is to treat that relation, however it is defined, as the starting point for developing an account of the meaning of a conditional. Douven and Verbrugge (2010) argued that one can distinguish between different types of inferences, and classified inferential conditionals as *deductive*, *inductive*, or *abductive*, inspired by classifications of conditionals in empirical linguistics (Declerck and Reed, 2001; Dancygier and Sweetser, 2005). This typology motivated a new, "inferential," truth-conditional semantics for indicative conditionals (Krzyżanowska, Wenmackers, & Douven, 2014; Douven, 2015). Independently, a related account has been proposed by Spohn (2013) and Olsen (2014), who analyzed indicative conditionals as expressing *reason relations* between their antecedents and consequents. This relationship can be operationalized probabilistically in terms of the ΔP rule (ΔP = P(C|A) - P(C|¬A)). A conditional's antecedent, A, is said to be a reason for the consequent, C, if A raises the probability of C, that is, if A is positively relevant for C. Since ΔP is defined as a difference between P(C|A) and P(C|¬A), ΔP must be positive for A to be a reason for C, and, consequently, for a conditional, "If A, then C," to be acceptable. Positive Relevance can be seen in example (1) above: parents' refusing to vaccinate their children increases the probability of measles or whooping cough outbreaks. By contrast, probabilistic irrelevance can be seen in example (2) above: the probability of UK leaving the EU given the existence of life on some extra-solar planet is, to the best of our current knowledge, exactly the same as the probability of UK leaving the EU given that there is no life outside of the Solar system at all. That is, ΔP = 0, or the antecedent is probabilistically irrelevant for the consequent in this case.

The Relevance Effect

Results by Skovgaard-Olsen et al. (2016a) recently raised an explanatory challenge for the Suppositional Theory of conditionals and Mental Model Theory. Both theories postulate that indicative conditionals have a core meaning which does not include relevance relations between the antecedent and the consequent. However, when investigating the probability and acceptability of indicative conditionals, Skovgaard-Olsen et al. (2016) found that relevance strongly moderated the evaluations of indicative conditionals. For cases of Positive Relevance (P(C|A) - P(C|¬A) > 0 <=> ΔP > 0), the conditional probability remained a good predictor of both the acceptance and probability of the conditional. For cases of Negative Relevance (P(C|A) - P(C|¬A) < 0 <=> ΔP < 0) and Irrelevance (P(C|A) - P(C|¬A) = 0 <=> ΔP = 0), this relationship was disrupted because the participants tended to view the indicative conditional as defective under those conditions.

In what sense does the Relevance Effect constitute a challenge to the Mental Model theory and the Suppositional Theory? The extent to which it does depends on whether the Relevance Effect belongs to the core meaning of the conditional - its semantics - or arises, instead, from the context of utterance of a conditional - its pragmatics. If the Relevance Effect belongs to pragmatics, then the main theories can just claim to be theories about the core content of indicative conditionals and hold that they need to be supplemented with auxiliary hypotheses concerning the pragmatic mechanisms involved in communication.

To address this question, we focus on a set of well-known phenomena at the interface between semantics and pragmatics: namely conversational implicature, presupposition, and conventional implicature (see later sections for definitions of each). We do so because these are phenomena for which there are reasonably well-established diagnostic tests. If we can explain the Relevance Effect by one of these phenomena, we are a step closer to adjudicating on the semantics/pragmatics issue. A final judgment will depend both on how we define semantics and pragmatics and on how we subsequently classify conversational implicature, presupposition, and conventional implicature. Both the definition and subsequent classification are live issues. But instead of resolving those issues here, our focus will be on classifying relevance effects with respect to these three established linguistic phenomena at the interface between semantics/pragmatics. For present purposes, we follow Birner (2014) in adopting the following typical characteristics of semantics and pragmatics (Table 1):

Table 1. Pragmatic/Semantic Distinction

Semantics	Pragmatics
literal	non-literal
context-independent	context-dependent
non-inferential	inferential
truth-conditional	non-truth-conditional[21]

To this we might add that semantics typically concerns the conventional meaning of words and sentences, while pragmatics typically concerns non-conventional meaning. While these characteristics might define the prototypical semantic and pragmatic phenomena, the characteristics can come apart. For instance, a phrase such as 'the foot of the mountain' may strike us as non-literal, mountains not having body parts, but it will also likely strike us as non-inferential, truth-conditional, and conventional. Unsurprisingly, then, it can prove

[21] Some would argue that some pragmatic phenomena are, in fact, truth-conditional. For discussion, see Carston (2002), Recanati (2011), and Birner (2014).

controversial to categorize any given phenomenon as semantic or pragmatic. Of the phenomena we consider, only conversational implicatures are regarded as uncontroversially pragmatic. Conventional implicatures, in contrast, are commonly thought of as a secondary layer of semantic meaning which is auxiliary to the primary truth-conditional semantic layer (Potts, 2007, 2015). Presuppositions, on the other hand, have both semantic and pragmatic interpretations (Beaver & Geurts, 2014), with influential proponents on either side—with, for instance, Stalnaker (2016) defending a pragmatic approach and von Fintel (2008) adopting a semantic one. The distinction between these various linguistic categories is discussed in further details below.

We test among the linguistic categories in four experiments. Experiment 1 tests whether the Relevance Effect arises because of conversational implicature. Experiment 2 tests whether it arises because of a presupposition failure. Experiment 3 tests whether it arises because of a conventional implicature.

2.2 Experiment 1: Conversational Implicatures

Conversational Implicatures

We start with the paradigm-case of pragmatics: the conversational implicature. Conversational implicatures arise when a speaker means something different from the conventional meaning of the sentence they utter. For instance:

> Alan: Are you going to Paul's party?
> Barb: I have to work. (Davis, 2014)

Here, Barb utters a sentence with a clear conventional meaning—that she has to work—but also conversationally implicates that she will not be attending Paul's party (because she has to work). To take another familiar example:

> Angry Parent: Did you eat all of the chocolate cake?
> Guilty Child: I ate *some* of it.

Here, the child utters a sentence that conventionally means something like 'I ate at least one morsel of cake' and is quite compatible with 'I ate all of the cake'. But the child, perhaps hoping to spread the blame, also implicates that he/she did not eat all of the cake: that there is another culprit. Inferences of this latter type are known as scalar implicatures.

Grice (1989) set out to explain how conversational implicatures arise, formulating a general principle of cooperative discourse: that speakers 'make [their] contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which [they] are engaged' (Grice, 1975, p. 45). He fleshed this principle out into a set of conversational maxims, or descriptive norms. On his account, speakers should give enough, but not too much, information (Maxim of Quantity); should avoid saying falsehoods or things for which they lack evidence (Maxim of Quality); should be relevant (Maxim of Relation); and should avoid obscurity and ambiguity and be brief and orderly (Maxim of Manner).

According to Grice, implicatures can only arise at all because hearers assume that speakers are generally cooperative: that they follow the maxims. But, as he pointed out, speakers can, in fact, behave in different ways towards the maxims: they can observe, violate, flout, or opt out of a maxim. Most important for present purposes are the observing or flouting of maxims, either of which generates an implicature.[22] For instance, other things being equal, when a speaker says, 'I had two bagels for breakfast', a hearer will assume that the speaker is observing the Maxim of Quantity (is providing sufficient information) and implicating 'I did not have three bagels for breakfast' (Birner, 2013). For an example of flouting, consider a professor who is writing a recommendation letter for a student, and is expected, in the normal run of things, to comment favorably on the student's academic ability, diligence, and so on. A professor who comments, instead, on the student's handwriting is flouting—openly disregarding—the Maxim of Quantity and implicating that the student's academic ability (and so on) is not worthy of praise (Grice, 1989). In cases of flouting, we can think of an implicature as being necessary to preserve the assumption that the speaker is being cooperative.

Conversational implications constitute a paradigmatic example of pragmatic modulation, an interpretational process whereby semantic content is extended by pragmatic mechanisms that take contextual factors into account. It is not unusual to find references to processes of pragmatic modulation as extending semantic theories in the psychology of reasoning (e.g., Johnson-Laird & Byrne, 2002). But their status tends to be that of underspecified and untested, auxiliary hypotheses (Douven, 2017).

Since conversational implicatures are attempts to reconstruct the speaker's intended meaning, which goes beyond what is literally said, they are defeasible inferences, which can be explicitly blocked by the speaker. For instance, imagine a conversation between John and

[22] A speaker violates a maxim when they inconspicuously disregard it, not intending the hearer to notice—as when a speaker lies or misleads a hearer; a speaker opts out of a maxim when, say, they simply disengage from a conversation and 'do not play the game at all' (Birner, 2013).

Sophia at a party they are hosting. John, who is in the kitchen, asks Sophia 'Where are our guests?' If Sophia replies 'Some are in the garden' she might well be taken to mean that not all of the guests are in the garden. But Sophia can straightforwardly cancel this scalar implicature by adding, 'In fact, they all are.' We will call this type of hedging 'a cancellation speech act'. One of the main characteristics of conversational implicatures is that a commitment to them can be blocked by performing cancellation speech acts without producing a contradiction (Blome-Tillmann, 2013).

Conversational implicatures have already featured in the debate on relevance. For instance, Over et al. (2007) found a modest effect of $P(C|\neg A)$ as a predictor of P(if A, then C), for conditionals which were positively relevant (where $\Delta P > 0$). They offered the following explanation (p. 92):

> An Adams conditional is not equivalent to an explicit statement that [A] raises the probability of [C] (...), nor that [A] causes [C] (...). A conditional probability [$P(C|A)$] can be high when [A] does not raise the probability of [C] and when [A] does not cause [C]. For example, [$P(C|A)$] can be high simply because [$P(C)$] is high. Does this mean that supporters of the view that these conditionals are Adams conditionals cannot account for the weak negative effect of [$P(C| \neg A)$] in the current studies? Not necessarily, for they can argue that the use of a conditional pragmatically suggests, in certain ordinary contexts, that [A] raises the probability of [C] or that [A] causes [C]. (Editorial changes preserve the consistency of notation)

Over et al. (2007) then go on to suggest that probability raising, and the causal reading of indicative conditionals may be produced by a conversational implicature. It may be misleading to assert conditional sentences in the absence of probabilistic dependencies, but the reason for this does not reside in the core, semantic content of indicative conditionals.

Similarly, Johnson-Laird and Byrne (2002) have considered whether to make a connection between the antecedent and the consequent part of the core meaning of conditionals, only to reject it:

> We do not deny that many conditionals are interpreted as conveying a relation between their antecedents and consequents. However, the core meaning alone does not signify any such relation. If it did, then to deny the relation while asserting the conditional would be to contradict oneself. Yet, the next example is not a contradiction:

> If there was a circle on the board, then there was a triangle on the board, though there was no relation, connection, or constraint, between the two—they merely happened to co-occur. (p. 651)

Their argument is that one can cancel a commitment to there being a relation between the antecedent and the consequent without contradicting oneself. If so, then this commitment bears the mark of a conversational implicature. In Experiment 1 we will test this hypothesis.

The Cancellation Task

The purpose of Experiment 1 is to test whether the reason-relation reading of conditionals can be attributed to the presence of a conversational implicature. To test this hypothesis, we investigated whether the reason-relation reading of conditionals can be cancelled without contradiction. More specifically, Experiment 1 uses the perceived degree of contradiction in cancelling a scalar implicature as the lower baseline. We have already seen two examples of scalar implicatures, where speakers used (or could be mistaken for using) the weaker term 'some' to implicate 'not all'. The 'some' to 'not all' inference is the most famous case, but scalar implicatures can arise with various scales, such as scales of possibility ('It's possible he will come' can implicate 'It's not definite that he'll come'). The implicature is that the speaker has some reason for not using the more informative, stronger term. Scalar implicatures can be cancelled. In this respect they contrast markedly with our upper baseline, entailment. If sentence A entails sentence B, then whenever A is true, B is also true. For instance, 'John is a bachelor' entails 'John is unmarried'. By definition, entailments cannot be cancelled without contradiction. The test then consists in measuring whether attempts to cancel the reason-relation reading of conditionals are viewed as more like cancelling a scalar implicature than like cancelling an entailment relation. The rationale is that while scalar implicatures can be cancelled without contradiction, entailments cannot.

In addition, Experiment 1 contrasts attempts to cancel the reason-relation reading of conditionals with attempts to cancel the reason-relation reading of conjunctions, which is another connective featuring a prominent reason-relation reading (Carston, 1993). Finally, comparisons are made with two control items that do not involve conditionals. The first is an attempt to cancel a scalar implicature. The second is an attempt to cancel the entailment of a categorical assertion.

2.2.1 Method

Participants

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 100 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk from the USA, UK, Canada, and Australia. They were paid a small amount of money for their participation.

The following exclusion criteria were used: not having English as native language (zero participants), completing the task in less than 240 seconds or in more than 3600 seconds (15 participants), failing to answer two simple SAT comprehension questions correctly in a warm-up phase (44 participants), and answering ‘not serious at all’ to the question 'how serious do you take your participation' at the beginning of the study (1 participant). Since some of these exclusion criteria were overlapping, the final sample consisted of 65 participants. Mean age was 41.66 years, ranging from 24 to 65, 38.5 % of the participants were male; 58.5 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. Applying the exclusion criteria had a minimal effect on the demographic variables.

Design

The experiment had a within-subject design with three factors: Relevance (with two levels: Positive Relevance (PO), Irrelevance (IR)), Priors (with four levels: HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH), and Sentence Type (with two levels: Conditional (if), Conjunction (and)). Since the And_{PO} cell of our design was empty,[23] the participants were presented with 12 within-subject conditions.

Materials and Procedure for All the Experiments

Each of the 12 within-subjects conditions was randomly assigned one of 12 scenarios. Random assignment was performed without replacement such that each participant saw a different scenario for each condition. This ensured that the mapping of condition to scenario was counterbalanced across participants preventing confounds of condition and content. The list of the 12 scenarios used can be found in the Supplemental Materials.

To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-

[23] To avoid prolonging the experiment too much for the participants, we chose to focus on the IF_{IR} and AND_{IR} comparison in this paper.

up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002).

The experiment was split into twelve blocks of four pages, one block for each within-subjects condition. The order of the blocks was randomized anew for each participant and there were no breaks between the blocks. On the first page of each block, the participants were presented with a randomly chosen scenario text (which was repeated on the three following pages in a brighter grey color for reference). These scenario texts have been found in previous studies (Skovgaard-Olsen, Singmann, and Klauer, 2016b) to reliably induce assumptions about relevance and prior probabilities of the antecedent and the consequent that implement our experimental conditions. Table 2 displays sample conditions for the Mark scenario for Positive Relevance and Irrelevance.

Table 2. Stimulus Materials, Mark Scenario

Scenario	Mark has just arrived home from work and there will shortly be a great movie on television, which he has been looking forward to. Mark is quite excited because he recently bought a new TV with a large screen. He has a longing for popcorn, but his wife has probably eaten the last they had while he was gone.	
	Positive Relevance	**Irrelevance**
HH	If Mark presses the on switch on his TV, then his TV will be turned on.	If Mark is wearing socks, then his TV will work.
HL	If Mark looks for popcorn, then he will be having popcorn.	If Mark is wearing socks, then his TV will malfunction.
LH	If the sales clerk in the local supermarket presses the on switch on Mark’s TV, then his TV will be turned on.	If Mark is wearing a dress, then his TV will work.
LL	If Mark pulls the plug on his TV, then his TV will be turned off.	If Mark is wearing a dress, then his TV will malfunction.
	Positive Relevance (PO): mean ΔP = .32 Irrelevance (IR) mean ΔP = -.01	High antecedent: mean P(A) = .70 Low antecedent: mean P(A) = .15 High consequent: mean P(C) = .77 Low consequent: mean P(C) = .27

Note. HL: P(A) = High, P(C) = low; LH: P(A) = low, P(C) = high. The bottom rows display the mean values for all 12 scenarios pretested in Skovgaard-Olsen et al. (2016b).

For the Mark scenario text in Table 2, participants assume that Mark pressing the on switch raises the probability of his TV turning on, and that both of these sentences have a high prior probability given the scenario (Positive Relevance, HH). Conversely, the participants tend to assume that Mark’s wearing socks is irrelevant for whether the TV will work, and that both have a high prior probability (Irrelevance, HH).

Materials and Procedure specific to Experiment 1

For Experiment 1, the participants were given the following instruction:

> In the following you are going to see a short conversation, where Louis accuses Samuel of contradicting himself. Whether you agree with Samuel's assertions is beside the point. What we are interested in is just the extent to which you agree with Louis that Samuel is contradicting himself. When you read the sentences please pay attention to small differences in their content, so that we don't unfairly accuse Samuel of contradicting himself.

The participants were then presented with two control items:

> Samuel: John is a bachelor [/Some of the employees are invited to the party]
> ...but I am not suggesting that John is unmarried. [/that they're not all invited]
> Louis: Wait, you're contradicting yourself.

The task of the participants was to indicate the extent to which they agreed or disagreed with Louis's statement on a five-point Likert scale {strongly disagree, disagree, neutral, agree, strongly agree}. Before beginning the experiment proper, the participants moreover saw one practice trial, where we emphasized that attention was needed to notice the subtle differences between the wordings of the various types of cancellations used in the experiment.

On the following three pages, the participants were presented with one of the three dependent variables in random order (perceived contradiction of cancellation of entailment, of scalar implicature, and of the reason-relation reading). The task of the participants was always to assess the extent to which they agreed with Louis' claim that Samuel contradicted himself. Using the HH conditions from above, the three types of cancellation were implemented as follows:

Table 3. Cancellation Types in Experiment 1

Entailment	Scalar Implicature	Reason Relation
	Conditionals, Positive Relevance	
Samuel: Mark presses the on switch on his TV. And IF Mark presses the on switch on his TV, THEN his TV will be turned on. ...but I am not suggesting that Mark’s TV will be turned on.	**Samuel:** Mark presses the on switch on his TV. And IF Mark presses the on switch on his TV, THEN it is possible that his TV will be turned on. ...but I am not suggesting that if so, it isn't highly likely that Mark’s TV will be turned on.	**Samuel:** Mark presses the on switch on his TV. And IF Mark presses the on switch on his TV, THEN his TV will be turned on. ...but I am not suggesting that these two things are related.

Conditionals, Irrelevance		
Samuel: Mark is wearing socks. And IF Mark is wearing socks, THEN his TV will work. ...but I am not suggesting that Mark's TV will work.	**Samuel:** Mark is wearing socks. And IF Mark is wearing socks, THEN it is possible that his TV will work. ...but I am not suggesting that if so, it isn't highly likely that Mark's TV will work.	**Samuel:** Mark is wearing socks. And IF Mark is wearing socks, THEN his TV will work. ...but I am not suggesting that these two things are related.
Conjunctions, Irrelevance		
Samuel: Mark is wearing socks AND his TV will work. ...but I am not suggesting that Mark's TV will work.	**Samuel:** Mark is wearing socks AND it is possible that his TV will work. ...but I am not suggesting that it isn't highly likely that Mark's TV will work.	**Samuel:** Mark is wearing socks AND his TV will work. ...but I am not suggesting that these two things are related.

Note. For conditionals, the entailment of the conclusion of Modus Ponens was cancelled. For conjunctions, the entailment of the conclusion of conjunction elimination was cancelled.

2.2.2 Results

Control Items

As Figure 1 suggests, the degree to which the participants viewed the cancellation speech act as giving rise to a contradiction was found to be significantly higher in the entailment control (*Mdn* = 5.00) item than in the scalar control item (*Mdn* = 3.00), V = 1260.5, *p* < .0001, *r* = -.65, for the Asymptotic Wilcoxon signed-rank test.

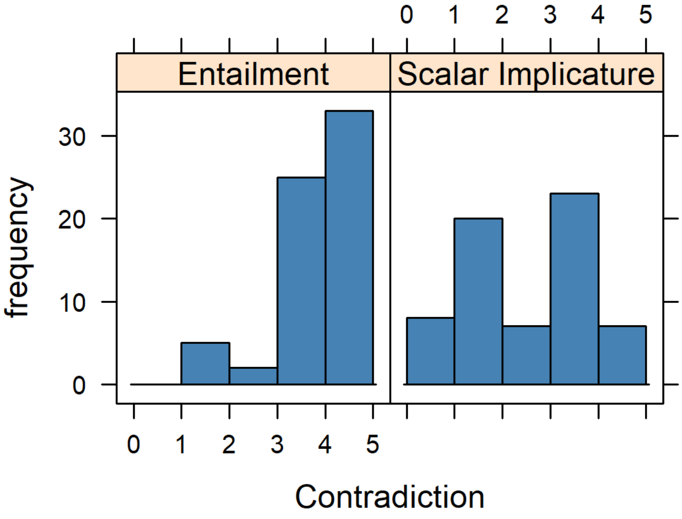


Figure 1. Histogram for Control Items. Note. The width of the bins is 1, so the bin from 0-1 on the histogram = 'strongly disagree' (or = '1', on the original response scale), the bin from 4-5 on the histogram = 'strongly agree' (or = '5', on the original response scale).

Comparing Cancellation Types for And$_{\text{Irrelevance}}$ and IF$_{\text{Positive}}$, IF$_{\text{Irrelevance}}$

Given the design, there were replicates for each participant and item. Hence, it was not appropriate to assume that the data were independently and identically distributed. Accordingly, the appropriate analysis was to use linear mixed-effects models, with crossed random effects for intercepts and slopes by participants and by scenarios (Baayen, Davidson, & Bates, 2008). This analysis was conducted using the statistical programming language R (R Core Team, 2013), and the package brms for mixed-effects models in Bayesian statistics was used (Bürkner, 2017). On the project page on the Open Science Framework, previous analyses of all the experiments reported in this paper are reported for classical statistics: https://osf.io/hz4k6/. As seen, the classical and Bayesian analyses converge on qualitatively similar results for all the studies.

Separate analyses were run for the Irrelevance (IR) and Positive Relevance (PO) items because there was no AND$_{\text{Positive}}$ cell of the design. For the Irrelevance items (AND$_{\text{Irrelevance}}$, IF$_{\text{irrelevance}}$), the following models were fit to the data:

(M1) a model that treats the participants' ratings of perceived contradiction as a function of the Cancellation Type factor (scalar implicature vs. entailment vs. reason relation), the Sentence factor ('if, then' vs. 'and'), and their interaction.
(M2) a model that builds on M1 but without the two-ways interaction.
(M3) a model that builds on M2 but without a main effect for the Sentence factor.

As indicated, these models were implemented in a Bayesian framework with weakly informative priors, using the R package brms (Bürkner, 2017). Since the responses obtained from the five-point Likert scale are ordinal responses, the responses were modelled as generated by thresholds set on a latent continuous scale with a cumulative likelihood function and a logit link function (Bürkner & Vuorre, 2018). Table 4 reports the performance of the models as quantified by the leave-one-out cross validation criterion and the WAIC information criterion.

Table 4. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1	3323.73	0	--	3313.2	0.9992
M2	3341.44	17.71	7.17	3329.2	0.0003
M3	3340.82	17.09	7.30	3328.6	0.0004

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC.

The information criteria clearly favour M1. Figure 2 plots its posterior predictions.

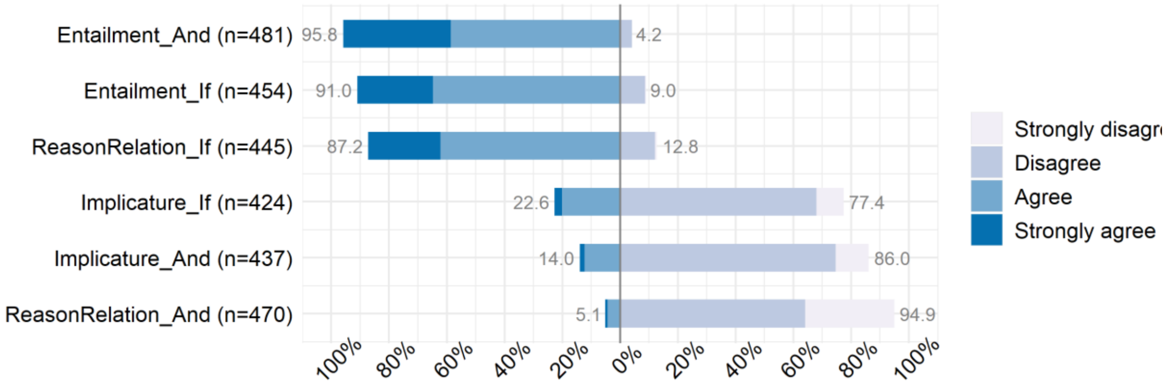


Figure 2. Posterior Predictions Based on M1. Note. The perceived degree of contradiction of the cancellation speech act was measured on a scale from ‘strongly disagree’ (1), ‘disagree’ (2), ‘neutral’ (3), ‘agree’ (4), and ‘strongly agree’ for Louis’ attribution of a contradiction to Samuel. The plot represents the predicted posterior probability of new participants selecting one of the displayed categories, given that they do not select ‘Neutral’. For each cancellation-sentence type, the number of samples drawn from the posterior distribution is shown.

Moreover, as shown in Figure 2, there is an interesting interaction in the data whereby attempting to cancel a commitment to the reason relation is viewed as just as contradictory as attempting to cancel a commitment to an entailment, for conditionals. In contrast, for conjunctions attempts to cancel a commitment to the reason relation is viewed as less contradictory than attempts to cancel a commitment to a conversational implicature.

As a manipulation check, it can be observed across sentences that the participants clearly distinguish between attempts to cancel a commitment to entailments and conversational implicatures. While the participants agree that Samuel is contradicting himself, when attempting to cancel a commitment to an entailment, they disagree when he is attempting to cancel a commitment to a conversational implicature ($b_{\text{Implicature}}$ = -4.12, 95%-CI [-5.04, -3.23], BF_{H0H1} = -1.44 * $10^{-38} \approx 0$).[24] The clear preference in favour of M1 (see Table 4) reflects the fact that while there is no main effect for the Sentence factor (b_{If} = -0.44, 95%-CI [-0.90, 0.05], BF_{H0H1} = 3.93), the Sentence factor is involved in an interaction with the Cancellation Type factor. Indeed, the evidence in favour of higher ratings for attempts to cancel reason relations with conditionals than with conjunctions is very strong on conventional standards for interpreting Bayes Factors (b_$_{\text{ReasonRelation:If}}$ = 5.02, 95%-CI [3.91, 6.17], BF_{H0H1} = 1.28 * $10^{-16} \approx 0$).

For the positive relevance conditionals, a similar mixed effects ordinal regression model was fitted to the data, and the same data pattern was found as for irrelevance

[24] Note that the slightly negative BF is probably due to a minor imprecision when estimating extremely small numbers around zero in R.

conditionals: while the participants tended to agree that Samuel was contradicting himself when attempting to cancel a commitment to an entailment, they disagreed when he attempted to cancel a commitment to a conversational implicature ($b_{\text{Implicature}}$ = -3.35, 95%-CI [-4.41, -2.34], $\text{BF}_{\text{H0H1}} = -1.89 * 10^{-18} \approx 0$). In contrast, moderate evidence in favour of the H_0 that entailment and reason relations did not differ could be obtained ($b_{\text{ReasonRelation}}$ = 0.40, 95%-CI [-0.19, 0.99], $\text{BF}_{\text{H0H1}} = 6.83$), with the reason relation cancellations, in fact, being rated slightly more contradicting than the entailment cancellations.

2.2.3 Discussion

For the control items it was found that the cancellation of an entailment was seen as more contradictory than the cancellation of a scalar implicature. While the control items concerned atomic sentences, our results indicate that this effect generalizes to the cancellation of entailments and scalar implicates occurring in conjunctions and conditionals. Consequently, we are able to use the perceived degree of contradiction of cancellation of scalar implicatures and entailments as two baselines that allow us to diagnose whether the cancellation of a reason relation suggested by conjunctions and conditionals is more like the cancellation of a conversational implicature or the cancellation of an entailment. Across conditions, it was found that the cancellation of a reason relation of conditionals is viewed as more like cancelling an entailment than cancelling a commitment to a conversational implicature. In contrast, the reverse pattern was found for conjunctions, where it was seen as *less* contradictory to cancel a reason-relation reading than to cancel a conversational implicature. The evidence thus indicates that while both conditionals and conjunctions have reason-relation readings, the content components that contribute to them are substantially different. For conjunctions, our results indicate that the reason-relation reading has the fingerprints of a conversational implicature. This finding goes against Carston (1993), who holds in relation to the enrichment of conjunctions beyond logical conjunction that they "are instances of pragmatically derived content which contributes to the proposition expressed by the utterance, to 'what is said' in Gricean terms. That is, they are not implicatures" (p. 30).[25] And it, moreover, directly contradicts the statement on conditionals by Johnson-Laird and Byrne (2002) quoted above. In contrast, for conditionals the results suggest that the reason-relation reading does not originate in a conversational implicature, since the content component that

[25] However, Robyn Carston (p.c.) points out that our data for conjunctions would be compatible with other types of cancellable pragmatic inferences like those discussed in Relevance Theory (Carston, 2002). If such an explanation is viable, predictions and alternative diagnostic tests would need to be formulated for future empirical work.

contributes to the reason-relation reading of conditionals is not cancellable without giving rise to a contradiction.

Now it may be argued that in Johnson-Laird and Byrne’s (2002) example, it seems perfectly acceptable to say “If there was a circle on the board, then there was a triangle on the board, though there was no relation, connection, or constraint, between the two—they merely happened to co-occur”. Similarly, a reviewer points out that in an example like the following, we have a conditional without a connection between the antecedent and the consequent:

Detective interviewing shopkeeper:

D: We need to know what Mr. Smith bought today, can you help us out?

S: I’m sorry, I didn’t find out about any customers’ names today.

D: Well, he was carrying a large polka-dotted umbrella.

S: If he carried a polka-dotted umbrella, then he bought a gold watch.

We acknowledge that there may be no relation between < there was a circle on the board; there was a triangle on the board >, nor between < he carried a polka-dotted umbrella; he bought a gold watch > at the type-level, because in general, propositions like the ones listed in these pairs are not probabilistically related. Still, at the token level, for the specific contexts in which these conditionals are used, there is a relation. If indeed ‘there was a circle on the board’ happens to co-occur with ‘there was a triangle on the board’, then there is a correlation, or a probabilistic dependency, which in fact can be used to make predictions for that specific context. What this tells us is that we need to carefully distinguish between whether the cancellation of the commitment to a connection is performed at the type or the token level. Moreover, these examples suggest that sometimes further contextual information may be needed for identifying the reason relation conveyed by indicative conditionals. For in this particular context, the proposition “Mr. Smith carried a polka-dotted umbrella” is indeed a reason for believing that Mr. Smith bought a gold watch, since it raises its probability.

Alternative Explanations

One of our reviewers suggests that implicatures vary in strength, with some being more cancellable than others (and potentially having to be cancelled in different ways), and cites the data of van Tiel et al. (2016) as evidence. Van Tiel et al. explored whether different lexical scales give rise to scalar implicatures at different rates, and found a considerable range, from 100% of participants taking 'cheap' to implicate 'not free' to only 4% taking 'content' to implicate 'not happy'. The reviewer, we take it, uses the rate at which implicatures arise as a proxy for their strength and, inversely, cancellability.

It is an intriguing possibility that, although implicatures have classically been taken to be cancellable, some are strong enough to resist cancellation, or may require other wordings for cancellation to be effective. However, it should be kept in mind that our wording effectively cancelled the reason-relation reading of conjunctions. At first sight, it is not clear why this wording would work with conjunctions and not with conditionals that otherwise express the same content, unless there is some difference in the status of the reason-relation reading between conjunctions and conditionals. That difference, we suggest, is precisely that the reason-relation reading is a conversational implicature for conjunctions but not for conditionals.

Yet, we welcome future research comparing the reason-relation reading with a broader class of scalar implicatures that implements various wordings for cancellation. We question, though, whether van Tiel et al.'s (2016) data are sufficient to suggest that implicatures can resist cancellation. Even when scalar implicatures arise at a rate of 100%, they are intuitively cancellable. For instance, the following sentence does not seem a contradiction: 'The course is certainly cheap and may even be free'. And although 96% of van Tiel et al.'s participants took 'some' to implicate 'not all', this implicature seems readily cancellable, as in 'some - in fact, all', as, indeed, is shown by the data on the control items for Experiment 1. The rate with which scalar implicatures arise may then not be a good proxy for degree of cancellability. In the absence of countervailing evidence, we therefore retain our conclusion that whereas the reason relation reading of conjunction could be produced by a conversational implicature, such a hypothesis is not viable for the reason-relation reading of conditionals.

A final objection is that, in Experiment 1, the contexts establish the truth of the antecedent. For instance, one item specifies that Mark is wearing socks, and continues "And if Mark is wearing socks, then it is possible that his TV will work". It is widely assumed that, for a conditional to be assertable, the truth of its antecedent must not already be settled. On a Gricean (1989) view, for instance, asserting a conditional 'if A, C' is infelicitous if the speaker knows something stronger - for instance, that A and C are both true. The infelicity supposedly arises because the speaker should respect the Maxim of Quantity and asserting something stronger - for instance 'A and C'. We might question, then, whether the present data will generalize to cases when the antecedent is not a matter of fact.

In response, we point out that existing data cast doubt on this standard assumption. In a series of studies, Krzyżanowska, Collins, and Hahn (2017b, 2018) have explored the assertability of conditionals in different contexts. It was found that conditionals can, in fact, be assertable, and not reliably less assertable than conjunctions, even when the component clauses (i.e., the antecedent and consequent) are known to be true.[26] Krzyżanowska, Collins, and Hahn's data challenge, then, the argument for why the present results might not generalize to cases when the antecedent is not known to be true. Nonetheless, empirical tests of the generalizability of our results are desirable, and desirable, too, would be further studies that directly address the range of alternative accounts suggested by the reviewer.

2.3 Experiment 2: Presuppositions

Having considered and rejected conversational implicatures, we turn, now, to presuppositions. Conceptually, we can think of presupposition as the marking of information that speakers take for granted when performing a speech act (Beaver & Geurts, 2014). That is, by making an assertion, the speaker acts as if the presuppositions are already an uncontroversial part of the common ground that the speaker shares with his or her interlocutors (Potts, 2015). For example, in the sentence 'Peter has stopped smoking' the word 'stopped' triggers the presupposition that Peter previously smoked. In the sentence, 'The fête was opened by the Duke of Oxford' the phrase 'the Duke of Oxford' triggers the presupposition that there is a (unique) Duke of Oxford.

The following example from Over et al. (2007) shows that the reason-relation reading of conditionals could be considered a presupposition as well:

> Consider for example:
>
> (6) If you take extra vitamin C, then your cold will be gone in three days.
>
> In most contexts, asserting (6) would be misleading, and very bad advice, if extra vitamin C was not a causal factor raising the probability that the cold will be gone in 3 days. The argument would be that there is often a pragmatic implicature when a conditional like (6) is asserted: that not taking extra vitamin C will make it probable that the cold will last longer than 3 days (p. 92)

To be sure, Over et al. (2007) here treat the example as indicating a conversational implicature, as noted above. However, one could argue that the speaker makes the assumption of a causal relation between vitamin C and getting rid of a cold as a presupposition of his or her assertion being a meaningful utterance. In addition, a presupposition failure hypothesis could be motivated by the intuition that indicative conditionals take a reason relation for granted. As Kadmon (2001: 14) says: "There is no better proof that a sentence S presupposes

[26] A reliable difference in assertability emerged only when there was no connection between clauses (i.e. when they were irrelevant).

a proposition B than our intuition that B is 'taken for granted' and is a precondition for felicitous use of S". For these reasons, we will consider an explanation in terms of presuppositions.[27]

A sentence containing a presupposition failure has traditionally been treated as either introducing a truth-value gap or being uniformly false (von Fintel, 2004). Since the traditional semantic framework was formulated in terms of truth conditions, presupposition failures are usually conceptualized in terms of their influence on truth evaluations. However, much has happened in the field of formal semantics since Russell (1905) and Strawson (1950) had their famous debates over presupposition failures. Since many contemporary developments explicate semantic content in terms of probability distributions (Yalcin, 2012; Lassiter, 2012; Moss, 2015), it seems natural to generalize the notion of a presupposition failure to a probabilistic context as well. That is, just as a sentence may carry presuppositions for a sensible truth value assignment (such as that the entity talked about actually exists), so sentences could carry presuppositions for a coherent probability evaluation. In particular, it could be conjectured that what the Relevance Effect really shows is that indicative conditionals have the condition of Positive Relevance as a presupposition for a coherent probability assignment.

The goal of Experiment 2 is to find out whether such a linguistic phenomenon plays a role in the participants' probability assignments underlying the Relevance Effect. One of the most characteristic properties of presuppositions is their projection behavior (Karttunen, 1973). Projection occurs when (1) expressions are embedded under operators to form more complex expressions and (2) the presuppositions of the simpler expression are inherited by the complex expression. To see projection in action, compare the following sentences:

The Danish pope is blue-eyed.
The Danish pope is not blue-eyed.

Both sentences presuppose that there is a Danish pope. Here, we say that the presupposition projects - is constant - under negation. Presuppositions are not affected by embedding under a range of logical operators (e.g., negation, modal operators) which alter the semantic entailments

[27] However, while the reason relations expressed by indicative conditionals will often be taken for granted as part of the common ground, there are argumentative uses of conditionals to introduce new reason relations in discussions, where which reason relations to accept itself becomes the content at-issue. This is, however, an issue that we will return to in the General Discussion, when considering what bearing our experimental results have for argumentation with indicative conditionals.

of the sentences in which they occur. Accordingly, 'the family of sentences test' is one of the main diagnostics for presuppositions (Chierchia & McConnell-Ginet, 1990). In this test, it is probed whether a conjectured presupposition survives under embedding in negation, interrogation operators, as the antecedent of a conditional, and when it is placed under a possibility modal (Kadmon, 2001). For some candidate content to be a presupposition, it is a necessary condition that the candidate survives embedding under these semantic operators, but it is not a sufficient condition, since other types of content also exist which can project across them (Potts, 2015).

Note, however, that presuppositions do not always project (Gazdar, 1979; Heim, 1983). For instance, while "Peter didn't stop smoking" carries the presupposition that Peter smoked in the past, this presupposition is blocked in "Peter didn't stop smoking. He never smoked!" (Xue & Onea, 2011). But it is possible that this shows not so much that presuppositions are not characterized by their projection behavior, but rather that they are defeasible and can be cancelled, when they are embedded under other operators (Beaver and Geurts, 2011).

Consequently, to test whether the Relevance Effect is an indicator of a presupposition failure, we propose in Experiment 2 to investigate its projection behavior by embedding with irrelevant clauses under negation operators.

The Negation Task

The main purpose of the Negation Task is to test whether the relevance effect is due to a presupposition failure. One of the central characteristics of presuppositions is that they project under negation (and other embeddings). The notion of a presupposition was introduced within a truth-conditional framework, but the idea can be generalized to probabilistic content and used to account for the Relevance Effect, if the positive relevance constraint ($\Delta P > 0$) is a presupposition of a coherent probability assignment to 'if A then C'. Since $\neg\varphi$ shares the same presuppositions as φ, the Relevance Effect is conjectured to be a probabilistic presupposition failure if the Relevance Effect applies equally to P(if A, then C) and P(¬(if A, then C)). That is, the same low probability assignments to P(if A, then C) in the Irrelevance condition should then be seen for the probability assignments of P(¬(if A, then C)) in the Irrelevance condition, whereas P(if A, then C) and P(¬(if A, then C)) should approximately sum to one for the Positive Relevance condition.

Alternatively, the negation operator might interact with the reason-relation reading in a way that one would expect of semantic content (although in this case it would be a case of probabilistic, semantic content). In the Irrelevance condition, both P(if A, then C) and P(if A, then ¬C) would receive low probability assignments, but P(¬(if A, then C)) may receive a high probability assignment, because [¬(if A, then C)] denies that A is a reason for C in the Irrelevance condition, whereas [if A, then C] says that A is a reason for C, and [if A, then ¬C] says that A is a reason against C. If one takes such a pattern of results together with the dissociations between the strong effect of relevance on probability assignments and the lack of effect of relevance on truth value assignments that was found in Skovgaard-Olsen et al. (2017), one might conclude that the reason-relation reading may not affect truth conditions, but is part of the probabilistic semantic content of conditionals. For instance, one could view the reason-relation reading as a conventional implicature, which is only tapped into through probability and acceptability evaluations.

It is possible that conventional implicatures also project across embeddings under logical operators when the participants are asked for truth evaluations (Potts, 2007). But if conventional implicatures are directly tapped into by probability assignments, then they should interact with these logical operators for probability evaluations. In contrast, if probabilistic presuppositions are conditions for a coherent probability assignment to φ, then they will also be conditions for a coherent probability assignment to ¬φ, and thus project past the negation operator.

Previous Work on Negated Conditionals

Although our main interest in negations is due to their diagnostic power for determining whether the Relevance Effect is due to a presupposition failure, it is necessary to consider briefly how the Suppositional Theory and Mental Model Theory have dealt with negations of conditionals, in the interest of investigating whether these semantic theories can predict our findings.

The literature on conditionals and negation has focused on distinguishing between suppositional and mental models theories of the conditional. The Suppositional Theory makes straightforward predictions. As we have seen, the theory predicts that people judge the probability of the affirmative conditional ‘If A, then C’ to be P(C|A). It also predicts that people judge the probability of a conditional wide-scope negation to be P(¬C|A), based on the Negation Principle (see below), and that both probabilities sum to unity. These predictions are taken to be part of the core of the Suppositional Theory. Indeed, the Negation Principle has been called a litmus test of suppositional theories as semantic theories of the conditional (Handley, Evans, & Thompson, 2006).

WIDE-SCOPE NEGATION: ¬ (if A, then C)

NARROW-SCOPE NEGATION: if A, then ¬ C
NEGATION PRINCIPLE: ¬ (if A, then C) <=> if A, then ¬ C
P(¬ (if A, then C)) = P(if A, then ¬ C)

Mental Model Theory makes more complex predictions (see, e.g., Espino and Byrne, 2012; Khemlani, Orenes, & Johnson-Laird, 2012). On the Mental Models account, an affirmative utterance is represented by a set of one or more possibilities represented by mental models; a negative utterance is represented by the complement of that set. The negation of the conditional depends on how people interpret the affirmative conditional. One possible interpretation is the so-called initial model (*A*, *C*), the negation of which amounts to the negation of the conjunction. But another interpretation is the fully explicit model, which contains all possibilities other than *A*, ¬*C*. In this case, the negation amounts to the conjunctive conclusion ‘*A* and ¬*C*’ (Espino & Byrne, 2012). Yet, following the latest developments in Khemlani, Byrne, and Johnson-Laird (2016), where ‘if A, then C’ is thought of as expressing the conjunction [‘A, C is possible’ ∧ ‘¬A, C is possible’ ∧ ‘¬A, ¬C is possible’], the negation of the conditional would have to express a disjunction: [¬(A, C is possible) ∨ ¬(¬A, C is possible) ∨ ¬(¬A, ¬C is possible)].

Experimental data show a complex picture. To support their account, suppositional theorists can point to evidence that people are reluctant to draw decisive conclusions from the negation of a conditional (Handley, Evans, & Thompson, 2006). This evidence seems to suggest that people eschew the conjunctive conclusion ‘*A* and ¬*C*’ (Handley, Evans, & Thompson, 2006). However, participants frequently endorse the conclusion ‘If *A*, then ¬*C*’ (Espino & Byrne, 2012; Khemlani, Orenes, & Johnson-Laird, 2014).[28] More problematically, when participants are given more options, they also endorse ‘If ¬*A*, then *C*’ responses, which do not follow from the Suppositional Theory, as well as the conjunctive responses ‘*A* and ¬*C*’ and ‘¬A and *C*’ (Espino & Byrne, 2012).

Since the Negation Principle has been used as a litmus test for the Suppositional Theory, an additional aim of Experiment 2 is to test whether it holds across relevance levels.

2.3.1 Method

Participants

[28] Egré and Politzer (2013) reconciled these data sets somewhat by arguing that the negation of a conditional is, at base, understood as ‘If A then possibly not C’. They then explicated contextual factors that modify this reading to recover stronger conjunctive and conditional responses.

Like Experiment 1, Experiment 2 was conducted over the Internet using Mechanical Turk and sampling from USA, UK, Canada, and Australia. 105 people participated in the experiment in exchange for a small payment. The same exclusion criteria were applied as in Experiment 1. The final sample consisted of 67 participants. Mean age was 41.3 years, ranging from 23 to 71 years; 41.8 % of the participants were male; 68.7 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The sample differed only minimally on the demographic variables above before and after applying the exclusion criteria.

Design

Experiment 2 implemented a within-subject design with the following factors: Relevance (with two levels: Positive Relevance, Irrelevance), and Priors (with four levels: HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH).

Materials and Procedure

Before beginning the study, the participants saw two control items in random order. For each, they were asked to assign a probability on a scale from 0 to 100% to a categorical sentence and to its negation, with an existential presupposition failure, using a slider. For each, the participants were randomly assigned to narrow and wide scope negations. An example of the categorical control items and their negations is as follows:

Affirmative: The queen of the USA is in her mid-thirties.

Narrow scope: The queen of the USA is NOT in her mid-thirties.

Wide scope: It is NOT the case that the queen of the USA is in her mid-thirties.

For the main study, the 8 within-subject conditions were randomly assigned to 8 different scenarios from to the same pool of 12 scenarios from Experiment 1, for each participant anew. Similarly, to Experiment 1, Experiment 2 was split into 8 blocks, one for each within-subject condition, with the same type of randomization structure as in Experiment 1.

In the context of the Mark scenario from Table 2, the participants might then be asked to make the following conditional probability judgments in the Positive Relevance HH condition:

Suppose that Mark presses the on switch on his TV.

Under this assumption, how probable is the following statement on a scale from 0 to 100%:

Mark’s TV will be turned on [/Mark’s TV will NOT be turned on].

In addition, the participants were asked to assign probabilities to conditional statements and their narrow and wide scope negations:

IF Mark presses the on switch on his TV, THEN his TV will be turned on.
IF Mark presses the on switch on his TV, THEN his TV will NOT be turned on.
It is NOT the case that IF Mark presses the on switch on his TV, THEN his TV will be turned on.

2.3.2 Results

As with Experiment 1, the within-subject design required analysis with linear mixed-effects models. As before, three models were fitted to the data using the package brms with weakly informative priors for mixed-effects models in Bayesian statistics with crossed random effects for intercepts and slopes by participants and by scenarios (Bürkner, 2017). Since the dependent variable consisted of continuous proportions containing zeros and ones, a zero-or-one inflated beta likelihood function was used (Ospina & Ferrari, 2012). In addition to Relevance (Positive Relevance, Irrelevance), the Type of Dependent Variable was used as a factor in the models. The types were as follows:

Affirm	=	P(If A, then C)
Wide	=	P(¬ (If A, then C))
Narrow	=	P(If A, then ¬ C)

Of the models fitted, M4 included the ‘DV Type’ factor, Relevance, and their interaction. M5 was like M4 but removed the interaction. M6 was like M5 but without a main effect for the Relevance factor. Table 5 reports the performance of these models as quantified by the leave-one-out cross validation criterion and the WAIC information criterion.

Table 5. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M4	1575.43	0	--	1572.8	0.922
M5	1583.80	8.36	6.14	1580.1	0.024
M6	1582.00	6.57	7.94	1578.4	0.054

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC.

The information criteria clearly converge on treating M4 as the winning model. This preference for M4 reflects the fact that there is a strong interaction effect in the data, which is displayed in Figure 3 through the cross-over of the lines for Affirm and Wide.

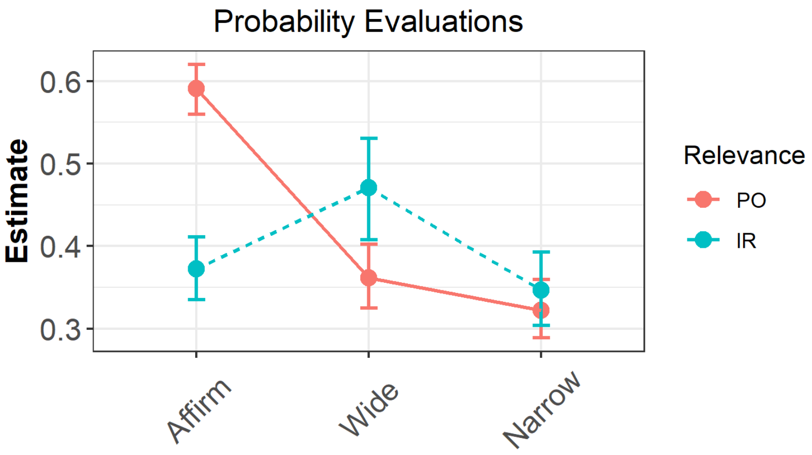


Figure 3. Posterior Mean Estimates for M4. *Note.* 'PO' = Positive Relevance; 'IR' = Irrelevance.

When averaging across the levels of the Relevance factor, Narrow was evaluated below Affirm (b_{Narrow} = -1.43, 95%-CI [-1.68, -1.18], $\text{BF}_{\text{H0H1}} = -1.43*10^{-38} \approx 0$), and Wide was rated below Affirm (b_{Wide} = -1.21, 95%-CI [-1.46, -0.96], $\text{BF}_{\text{H0H1}} = 3.14*10^{-182} \approx 0$). And when averaging across the levels of the DV Type factor, the effect of Irrelevance was to suppress the ratings ($b_{\text{Irrelevance}}$ = -1.15, 95%-CI [-1.40, -0.91], $\text{BF}_{\text{H0H1}} = 8.17*10^{-23} \approx 0$), which is in particular visible in the large drop in the ratings of Affirm in the Irrelevance condition. However, the effect of the interaction is to substantially raise the ratings of Wide in the Irrelevance condition ($b_{\text{Irrelevance:Wide}}$ = 1.72, 95%-CI [1.29, 2.14], $\text{BF}_{\text{H0H1}} = 2.06*10^{-16} \approx 0$).

Control Items

The control items served the function of securing construct validity. When attempting to draw conclusions about whether probability assignments to conditionals across relevance levels involving negation operators are diagnostic of a presupposition failure, we need to ensure that our interpretation of a data pattern in probability ratings is actually representative for presupposition failures. To do this, we employ items that represent classical cases of presupposition failures and determine empirically how their probability ratings depend on the presence and absence of negation operators.

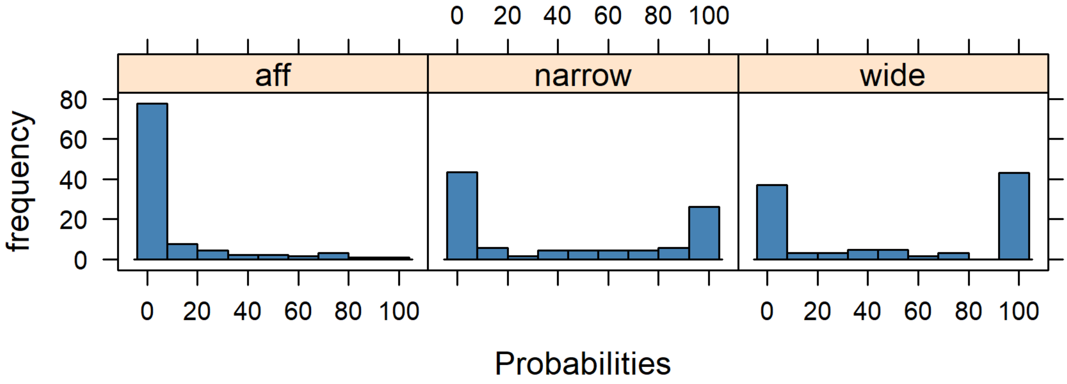


Figure 4. Histogram for Control Items. Note. 'wide' = wide scope negation; 'aff' = affirmativ (no negation); 'narrow' = narrow scope negation. Probabilities on a scale from 0 to 100%.

Since the participants differed as extremely as they could in reaction to the control items in the presence of negations, we ran a separate analysis to investigate whether the group-level performance of the participants with respect to conditionals likewise disguises such a marked individual variability.

Individual Differences in Response to the Control Items

A group of 20 participants with the lowest responses (*Mdn* = 2) to the wide scope control item, and a group of 20 participants with the highest responses (*Mdn* = 82) to the wide scope control items, were formed. It was found that the High group also assigns significantly higher probabilities to the narrow negation control item than the Low group, W = 122, *p* = .017, *r* = -.53, exact Wilcoxon rank sum test.

However, for both groups the same data pattern reported on the group level in Figure 3 was found (see Figure 5 below). This in turn indicates that whatever individual differences in the probability assignments to presupposition failures are found in the control items, they are not matched by the participants’ probability assignments to conditionals in the Irrelevance condition. For both groups, there is little overlap between their behaviour with respect to the control items and the conditionals in the Irrelevance condition, as shown in Figure 5 based on a mixed linear model like the previous, which, however, included the Group factor along with its interactions:

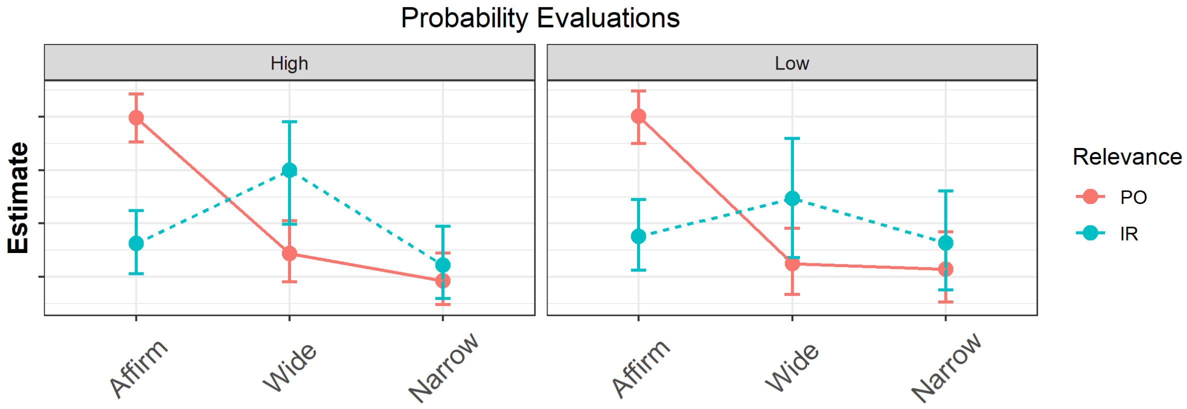


Figure 5. Group Low and High in Comparison. Note. Plots show the posterior means for the Low group and the High group. The groups were formed based on the participants' responses to the control items with wide scope negation. 'PO' = Positive Relevance; 'IR' = Irrelevance.

Probabilistic Coherence

As part of Experiment 2, we investigated the participants' probabilistic coherence, that is, whether complementary pairs of participants' probability assignments sum to 100%. We did so on the grounds that, if reason relations are supposed to be made part of the semantic, probabilistic content of indicative conditionals, then probabilistic coherence makes up a natural requirement. In Appendix 1, we present analyses based on linear mixed-effects models in classical statistics which indicate violations of probabilistic coherence in the Irrelevance condition due to the impact of an influential group of outliers, which violates additively strongly in both directions (either by having values of complementary pairs summing to 0 or 200 in the Irrelevance condition). For this reason, we here follow Kruschke's (2014) recommendation of conducting a robust regression analysis with a t-distribution as the likelihood function, which is less sensitive to the influence of outliners. As before, we did the analyses using mixed-effects models in Bayesian statistics with crossed random effects for intercepts and slopes by participants and by scenarios (Bürkner, 2017). The linear mixed-effects models treated the participants' probability judgments as a function of Relevance (Positive Relevance, Irrelevance) and Constraint Type (see below), and their interaction, which was allowed to vary across participants and scenarios. Three such models were contrasted: (M7) modelled the rating as a function of the factor Relevance, the factor Constraint Type, and their interaction, (M8) like M7 but without the interaction, and (M9) like M8 but without a main effect for the Constraint Type factor. The Constraint Type factor had the following levels:

P(C|A) Consistency: $y = |100 - (P(C|A) + P(\neg C|A))|$

Narrow Consistency: $y = |100 - (P(\text{If A, then C}) + P(\text{If A, then } \neg C))|$

Wide Consistency: y = |100 – (P(If A, then C) + P(¬(If A, then C))|

That P(C|A) + P(¬C|A) = 100% across relevance conditions is a general requirement of probabilistic coherence when reasoning with conditional probabilities. That P(If A, then C) + P(¬(If A, then C)) = 100% across relevance conditions is a requirement of conditional consistency that theories of conditionals across the board should accept. In contrast, that P(If A, then C) + P(If A, then ¬C) = 100% is a requirement that theories of conditionals adopting the Negation Principle should accept.

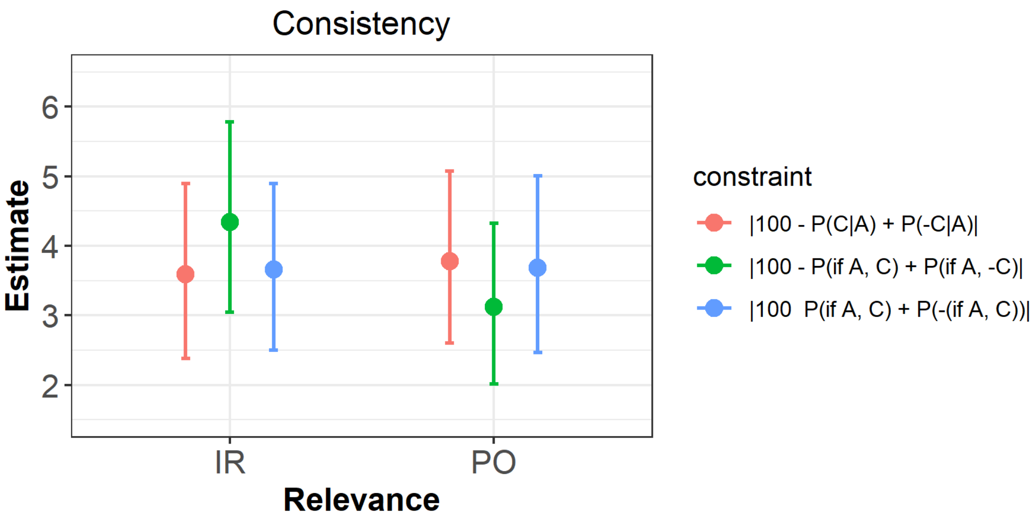


Figure 6. Probabilistic Coherence. Note: The plot displays departures from probabilistic coherence on the three investigated measures using robust regression analysis with a t-distribution as the likelihood function.

As Figure 6 indicates, the responses of the participants were shown to be remarkably consistent once the influence of outliers was controlled through robust regression analyses. This contrasts with the analysis reported in Appendix 1, where particularly large departures from probabilistic coherence were found for the Narrow Consistency constraint in the Irrelevance condition (see Figure 6, Appendix 1).

Table 6 reports the performance of these models as quantified by the leave-one-out cross validation criterion and the WAIC information criterion.

Table 6. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M7	17069.63	0.66	4.36	17068.8	0.4080
M8	17072.61	3.64	0.52	17072.1	0.0819
M9	17068.97	0	--	17068.4	0.5101

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC.

The model comparison in Table 6 indicates a tight run between M7 and M9 with a slight preference for M9. The preference for M9 is explained by the lack of a main effect for Constraint Type ($b_{\text{constraintP(C | A)}}$ = 0.05, 95%-CI [-0.81, 0.89], BF_{H0H1} = 7.00; $b_{\text{constraintNarrow}}$ = 0.76, 95%-CI [-0.20, 1.76], BF_{H0H1} = 1.84). Similarly, there was no main effect of Relevance ($b_{\text{PositiveRelevance}}$ = 0.18, 95%-CI [-0.62, 1.00], BF_{H0H1} = 6.63). The relative competitiveness of M7 indicates that the Constraint Type factor was involved in an interaction with the Relevance factor, however, which is also the reason why removing the interaction term in M8 results in the worst fitting model.

A closer look at the interaction indicates that while there is moderate evidence for H_0 when setting the coefficient for the P(C|A) Consistency constraint equal to zero ($b_{\text{PositiveRelevance:constraintP(C | A)}}$ = -0.15, 95%-CI [-1.23, 0.94], BF_{H0H1} = 5.23), there is anecdotal evidence against H_0 when setting the coefficient for the Narrow Consistency constraint equal to zero ($b_{\text{PositiveRelevance:constraintNarrow}}$ = -1.42, 95%-CI [-2.82, -0.16], BF_{H0H1} = 0.43).

2.3.3 Discussion

The results indicate that the Negation principle (¬(if A, then C) <=> if A, then ¬C) could only be maintained for the Positive Relevance condition; for the Irrelevance condition it was systematically violated.

For our diagnostic purposes, it is particularly interesting that the wide scope negation (¬(if A, then C)) was rated the highest for the Irrelevance condition. This finding indicates that, whatever content component is responsible for the low probability assignments in the Irrelevance condition (dubbed ‘the Relevance Effect’), it does not project under wide scope negation. Yet surviving embeddings under the semantic operators listed in the family of sentences test is treated as a necessary condition for the targeted content to count as a presupposition (Potts, 2015). This suggests that the Relevance Effect is probably not produced by a presupposition failure (with conditionals requiring that a Positive Relevance constraint is satisfied in order to obtain a high probability). To substantiate this interpretation, a comparison was made with the control items (discussed below). But note that we emphasize that passing the family of sentences test is a necessary, but not sufficient, condition for counting as a presupposition, because other types of content are known to have instances where they project as well. In particular, the conventional implicature 'but' expresses a contrastive relation that projects in examples like ‘If John is rich, but happy, then…’. However, unlike classical presupposition triggers like 'know' and 'again', conventional implicatures like 'but' and 'therefore' are only able to pass the family of sentences test for some of its instances, it seems (‘?’ suggesting possible failures):

? It is not the case that John is rich but happy
? It is possible that John is rich but happy
? Is John rich but happy?
If John is rich but happy, then ….
? It is not the case that John is rich therefore he is happy
? It is possible that John is rich therefore he is happy
? Is John rich therefore he is happy?
? If John is rich therefore he is happy, then …

Arguably, in order for the embedded therefore-sentences even to be grammatical, it seems that they would need to be rewritten as 'John is rich *and therefore* he is happy', but then 'and' takes the role of being the main connective (see also McCawley, 1993, pp. 318-319).

Control Items

In the present case, our control condition introduces the rather curious finding that the participants differ as extremely as they can on a probability scale on their reaction to the control item in the presence of negations (in particular with respect to wide scope negations).

For this reason, we did a separate analysis to investigate whether the group-level performance of the participants with respect to conditionals likewise disguises such a marked individual variability. As we saw, this turns out not to be the case (since the same qualitative differences with respect to the probability assignments of the conditional at the group level is found within each group who differed in their performance on the control item, Figure 5). This in turn can be interpreted as supererogatory evidence for the above conclusion that the participants’ performance with conditionals in the Irrelevance condition is not an expression of how they generally treat presupposition failures. Although both groups differed markedly in how they assign probabilities to a paradigmatic case of presupposition failures, the two groups showed the same qualitative pattern when assigning probabilities to conditionals. Furthermore, their common data pattern in relation to conditionals was not matched with respect to the control items for either group. Whatever is driving the participants’ responses for the conditionals in Irrelevance conditions is in other words unlikely to be the result of how the participants process presupposition failures in general.

Moreover, looking back to the results from Experiment 1, it is worth keeping in mind that conversational implicatures and presuppositions can be denied without contradiction, whereas conventional implicatures and entailments cannot be denied without contradiction (Potts, 2007). Yet in Experiment 1 it was found that denying the reason-relation reading of conditionals is viewed as just as contradictory as denying an entailment. It has been observed, though, that the examples with cancellable presuppositions seem to concern specifically embedded presuppositions (Beaver & Geurts, 2011), which was not what was investigated in Experiment 1.

Local Accommodation?

A reviewer suggested that the lack of projection behaviour found for embedding conditionals in the Irrelevance condition under the wide-scope negation operator could be explained by a process of local accommodation, which essentially turns presuppositional content into regular entailed content. To illustrate, von Fintel (2008) discusses the following example by Heim (1983):

There is no king of France. Therefore, the king of France is not hiding in this room.

The point is that accepting the first sentence updates the common ground so that it is taken for granted that there is no king of France. Against this utterance context, it is not possible for the second sentence to presuppose that there is a king of France. Instead, the presupposition of the second sentence is cancelled by a process of local accommodation. More specifically, the context is treated as updated vacuously by an empty set consisting of the contradictory context that would be generated had ‘there is a king of France’ been added to the context set.

The suggestion is then that the reason why we find the highest probability ratings for the wide negated conditional in the Irrelevance condition is because such a process of local accommodation cancels the presuppositional content and blocks its projection behaviour.

In response, we would like to point to a number of disanalogies between our experimental task and examples like the one above. First of all, there is no conversational context, no utterance of the conditionals as assertions, and no prior updates of the common ground in our experimental task. The participants merely see the conditional sentences and are asked to assess their probabilities. To be completely parallel with the example above, the participants would have to be presented with assertions of the following kind by a speaker:

[A] is irrelevant for [C]. Therefore, it is not the case that if [A], then [C].

Had the participants been presented with such materials, it could be argued that local accommodation would cancel the $\Delta P > 0$ presupposition of the conditional, since the first sentence already rules out this presupposition. But given that our experimental task was different, we would still need to see it be established that local accommodation can arise for sentences such as ”If Mark is wearing socks, then his TV will work”, for which it is obvious that the two things are unrelated and where one cannot rely on the participants to modify the context such that these sentences become accepted.

Negation Principle

The results indicate that the simple relationship by which P(C|A) predicts P(if A, then C) and P(¬C|A) predicts both P(¬(if A, then C)) and P(if A, then ¬C) across relevance levels does not hold. In particular, this relationship is generally less strong for Irrelevance than for Positive Relevance, and less strong for wide scope negation (where it is almost absent) than for narrow scope negation. This in turn violates the negation principle (P(¬(if A, then C)) = P(if A, then ¬C), which was found to hold only for the Positive Relevance condition.

Moreover, the results indicate that there is an interaction with the negation operator and the relevance factor, which makes '¬(if A, then C)' the most probable statement in the Irrelevance condition. That is, the low probability assignment to 'if A, then C' in Irrelevance is not matched by a low perceived probability of the wide scope negated conditional in Irrelevance. As already discussed, this is evidence against an interpretation in terms of a presupposition failure which projects under wide scope negation; evidence to which the pattern of results for the control items adds further weight.

How do the present data on the Negation Principle bear on the existing literature? Our data bear most straightforwardly on the Suppositional Theory of conditionals. Recall that suppositional theorists have themselves called the Negation Principle a litmus test of the Suppositional Theory (Handley et al., 2006). Our data suggest that the Negation Principle does not hold in general, and that an important qualification is needed. Moreover, Mental Model Theory is not straightforwardly able to predict systematic differences according to relevance condition. Recall that, for Byrne and Johnson-Laird (2002), reason relations best correspond to a conversational implicature, but Experiment 1 has already cast doubt on this possibility at least for affirmative conditionals.

Probabilistic Coherence

If reason relations are supposed to be made part of the semantic, probabilistic content of indicative conditionals, then probabilistic coherence makes up a natural requirement. In our analysis, we tested three requirements of probabilistic coherence: probabilistic coherence of conditional probabilities (P(C|A) + P(¬C|A) = 100% across relevance conditions), conditional consistency (P(If A, then C) + P(¬(If A, then C)) = 100% across relevance conditions), and a conditional consistency requirement based on the Negation Principle (P(If A, then C) + P(If A, then ¬C) = 100% across relevance conditions).

Divergences from all three requirements of probabilistic coherence were found for the Irrelevance condition, when outliers were included (see Appendix 1). The strongest divergences were found for the conditional consistency requirement based on the Negation Principle. In contrast, it was possible to find high levels of probabilistic coherence for all three constraints across relevance levels, when the influence of outliers was controlled through robust regression techniques within a Bayesian framework (see Figure 6).

A priori we had only expected to find violations of probabilistic coherence for the Irrelevance condition, if $\Delta P > 0$ were made a presupposition of coherent probability assignment (just as the existence of what one is talking about can be made a presupposition of a truth value assignment). However, given the weight of the evidence cited above against the presupposition failure account of the Relevance Effect, the violations of probabilistic coherence in the Irrelevance condition when outliers are included need to be accounted for differently. We offer the following *post hoc* explanation. With 24 combinations of negation operators (no negation, narrow scope, wide scope), prior (HH, HL, LH, LL), and relevance levels (Positive Relevance, Irrelevance) for conditionals, and 16 combinations of conditional probabilities, it can be challenging to both pay attention to fine details in the stimulus materials and maintain consistency internally in the responses given. Accordingly, limitations in the cognitive resources invested in maintaining internal consistency in the online study may account for the violations of probabilistic coherence found in the group of outliers.

2.4 Experiment 3: At-Issue Content

Given that Experiments 1 and 2 cast doubt on the hypotheses that the Relevance Effect arises from conversational implicature or presupposition failures respectively, Experiment 3 investigated whether it arises from conventional implicature. For this purpose, Experiment 3 investigates a much-discussed property of conventional implicatures: that conventional implicatures are content that is not-at-issue.

Conventional Implicatures

Grice (1989) noticed meanings which do not seem to contribute to the truth or falsity of a sentence, but which nevertheless seem conventional. He termed such meanings 'conventional implicatures'. Classic examples include words such as 'but', 'therefore' and 'even' (Potts, 2007; Valleé, 2008; Salmon, 2011; Blome-Tillmann, 2013). For instance,

'He is English but brave.'
'Even the English can be brave'

'He is English and, therefore, brave.'

In the first sentence, 'but' signals a contrast between Englishness and bravery: bravery is somehow unexpected from the English. In the second sentence, 'even' likewise signals unexpectedness, perhaps also implicating that bravery is not so exceptional after all if the English are capable of displaying it. In the third sentence, 'therefore' signals a consequence relation between Englishness and bravery.

Each example above makes a claim that is contentious. Many readers might want to deny that there is a relationship - positive or negative - between Englishness and bravery. But according to Grice (1989), a reader so minded would struggle to say that the relationship expressed makes the sentences false. This is not, of course, to say that the sentences cannot be true or false. Readers can easily deem the first and third examples false if they believe that the 'he' in question is a coward or not English (or both) and can deem the second example false if they believe all English people to be incapable of bravery. But according to Grice (1989), the relationship between Englishness and bravery does not contribute to the truth evaluation of the sentence. Accordingly, these meanings - these conveyed relationships - do not seem to be straightforwardly semantic on the typical understanding of the term. There is, nevertheless, something conventional about these meanings, since they attach to specific words. Equally, these meanings are not straightforwardly pragmatic: they are not calculable based on the Gricean maxims, and they cannot be cancelled without contradiction.

Conventional implicatures contrast markedly with both presuppositions and conversational implicatures. Presuppositions do affect the truth evaluation of the sentences in which they occur, either by creating a truth value gap - the sentence is neither true nor false - or making the sentence uniformly false. Conversational implicatures are calculable based on the Gricean maxims and can be cancelled without contradiction.

One clear indicator that the Relevance Effect might be produced by a conventional implicature is found in a truth-table task with relevance manipulations in Skovgaard-Olsen et al. (2017). Across two experiments, a strong dissociation was found indicating that while the reason-relation reading of conditionals clearly influences probability and acceptability evaluations, it has almost no influence on truth evaluations. These results might be interpreted as support for a conventional implicature hypothesis, according to which the reason-relation reading of indicative conditional is a conventional aspect of their meaning, which cannot be cancelled without contradiction, is not calculable by the Gricean maxims, and is not targeted by truth evaluations. According to this interpretation, the reason-relation reading of conditionals would be similar to the reason-relation reading of 'A but C' and 'A therefore C ' (which in turn suggest that A is a reason against or for C, respectively). Support for this interpretation can also be derived from the fact that Skovgaard-Olsen et al. (2017) found the same strong dissociations with respect to but- and therefore-sentences as a function of the reason-relation reading that they found for the indicative conditionals.

One property that seems to be attributed to both presuppositions and conventional implicatures in recent discussions is that they are content that is not-at-issue. It has long been observed in relation to presuppositions that they are backgrounded content which is taken for granted (Kadmon, 2001). Recently, a similar idea has come to play a major role in treatments of conventional implicatures. The backdrop is that Bach (1999) made an influential case that appositives (e.g., 'Mozart, *the famous composer*, used to live here') should be treated as a central instance of conventional implicatures, instead of the kind of examples Grice (1989) considered. This recommendation was later taken up and fully developed in Potts (2005), where a formal system was developed that treats conventional implicatures as logically and compositionally independent of the at-issue content.

However, whereas the not-at-issue content of presuppositions is thought of as backgrounded and known by the participants, Potts (2015) argues that conventional implicatures usually introduce new content that is not-at-issue. One way in which this difference shows up is that whereas it is usually not redundant to first explicitly state a presupposition and then make an utterance carrying this presupposition, it is considered redundant to explicitly assert conventional implicatures, because their content is lexically encoded (Potts, 2007). To illustrate, whereas it is redundant to say, 'Mozart is a famous composer' and 'Mozart, the famous composer, used to live here', it is not considered redundant for a speaker to explicitly state that in his or her view 'there is a Danish pope' and 'the Danish pope is blue-eyed'.

A further difference is that whereas the not-at-issue content of presuppositions is thought of as cancellable in compound sentences, one cannot cancel a lexically encoded conventional implicature without contradiction (Potts, 2015).

Tonhauser (2012) collects a series of diagnostic tests for deciding whether a content component makes up content-at-issue. Central among these is that content-at-issue is targeted by direct denials ("No, ...") and acceptances ("Yes, ..."), whereas content-not-at-issue can only be denied by more indirect measures that interrupt the natural flow of the conversation like "Actually, ...", "Well, ...", "Hey, wait a minute!".

In Syrett and Koev (2014) a series of experiments was conducted with appositives based on this idea. While the authors found clear preferences for direct denials as targeting the main claim in utterances containing appositives, the evidence was more mixed when it comes to whether a preference for "Hey, wait a minute!" can be used as a sufficient condition for identifying content not-at-issue. This picture fits with the corpus analysis of natural occurrences of "Hey, wait a minute!" conducted in Potts (2008), which also did not find compelling evidence that this construction is only used with not-at-issue content like appositives. Similarly, Salmon (2011) also finds that "Hey, wait a minute!" rejections can be used with other constructions than their intended purpose. For this reason, we decided against employing this test for our tests of whether the reason relation constitutes content-at-issue of indicative conditionals. Instead, we decided to investigate which content component the participants naturally interpret direct acceptances and denials as targeting in Experiment 3, following the diagnostic tests of at-issue content in Tonhauser (2012).

2.4.1 Method

Participants

Like Experiment 1, Experiment 3 was conducted over the Internet using Mechanical Turk and sampling from the USA, UK, Canada, and Australia. 339 people participated in the experiment in exchange for a small payment. The same exclusion criteria were applied as in Experiment 1. The final sample consisted of 228 participants. Mean age was 40.8 years, ranging from 19 to 98[29] years; 41.7 % of the participants were male; 73.7 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The sample differed only minimally on the demographic variables above before and after applying the exclusion criteria.

Design

Experiment 3 implemented a mixed design with the following within-subject factors: relevance (with two levels: Positive Relevance, Irrelevance), priors (with four levels: HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH), and truth table cell (TT vs. TF). There was one between-subject factor: sentences (If then vs. Therefore).

Materials and Procedure

Because we did not want to use a scenario twice for a given participant, each participant underwent only 12 of the 16 within-subject conditions. These comprise the 8 conditions elicited by crossing relevance and prior factors, and a random 4 additional

[29] It is doubtful whether the response '98' really should be taken at face value. The next highest was '72'.

conditions from the relevance x prior subdesign. Truth-table cells (TT vs. TF) were randomly assigned to the 12 conditions, for each participant anew, so that 6 conditions were allocated with TT and 6 with TF.

Experiment 3 was split into 12 blocks with the same type of structure as in Experiment 1 (unless otherwise noted). At the beginning of the experiment, the participants were presented with items containing appositives for exploratory purposes, which are not reported here. For the experimental task the participants were instructed that they should help Pierre, a foreign exchange student, to learn English by correcting his mistakes. Their task was to indicate whenever Pierre had said something that was true and to correct him whenever he had said something that was false.

Below the scenario, the participants would see two facts that they and Pierre had learned about the scenario and a statement that Pierre had made. For instance, one participant might have seen the Mark scenario text from Table 2, continued as follows.

> You and Pierre both learn that:
> Mark presses the on switch on his TV.
> Mark's TV will be turned on. [/Mark's TV will NOT be turned on]
> Pierre: IF Mark presses the on switch on his TV, THEN his TV will be turned on.
> [/Mark presses the on switch on his TV THEREFORE his TV will be turned on]

The participants were then asked whether Pierre had made a true statement and they were instructed to give their response as a forced choice between two options which varied with the condition. In this example, the participants were then given a forced choice between:

> Yes, Mark presses the on switch on his TV and his TV will be turned on vs.
> Yes, Mark presses the on switch on his TV is a reason that his TV will be turned on

The remaining conditions followed a similar structure:

1. Positive Relevance TT "Yes, TT" vs. "Yes, A is a reason for C"

Forced Choice between two Reasons to accept

2. Irrelevance TF "No, TF" vs. "No, A is not a reason for C"

Forced Choice between two Reasons to reject

3. Positive Relevance TF "No, TF" vs. "Yes, A is a reason for C"

Forced Choice between Yes/No Answers + Justifications in a Conflict Case

4. Irrelevance TT " "Yes, TT" vs. "No, A is not a reason for C"

Forced Choice between Yes/No Answers + Justifications in a Conflict Case

If the reason-relation reading is content-at-issue for the participants, then the reason-relation justification should appear most attractive.

2.4.2 Results

As Figure 7 indicates, there was a widespread tendency to treat the reason-relation reading as content-at-issue for both indicative conditionals and therefore-sentences.

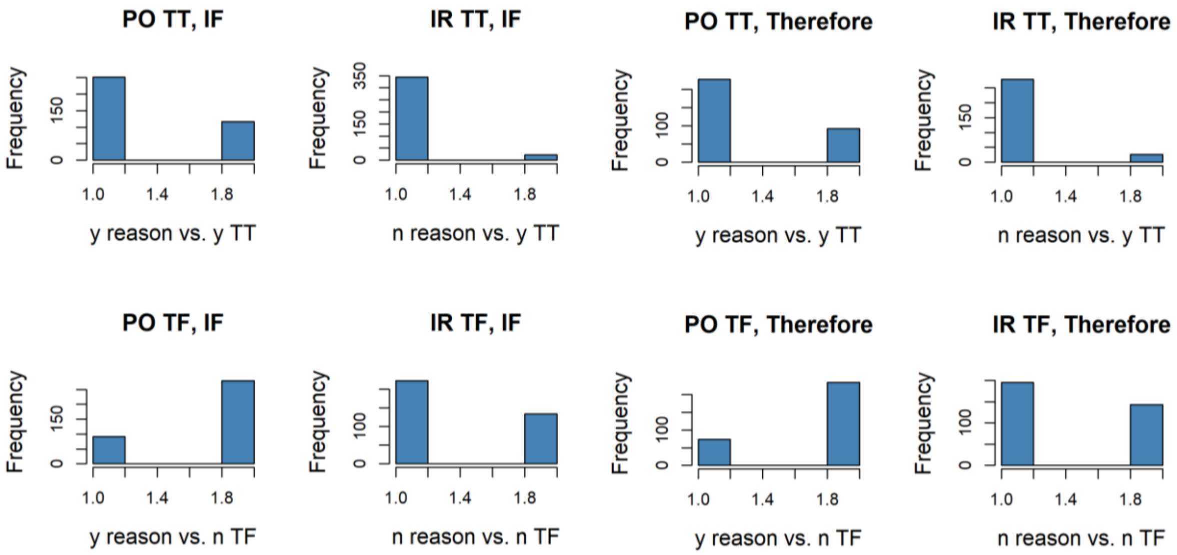


Figure 7. Histograms of Conditionals and Therefore. Note. 'y reason' = 'yes, A is a Reason for C'; 'n TF' = 'No, TF'; ‘y TT’ = ‘yes, TT’. ‘PO’ = Positive Relevance; ‘IR’ = Irrelevance.

The following constrained multi-nominal processing models were fitted to the data (Batchelder and Riefer, 1999; Skovgaard-Olsen et al., 2017):

$M_{\text{saturated}}$: model imposing no constraints. This model fits the data perfectly using one free parameter per degree of freedom provided by the data

M_{sentence}: model assuming that response probabilities are the same across sentences, while allowing for differences across the different levels of the Relevance factor

$M_{\text{relevance}}$: model assuming no differences across the different levels of the Relevance factor while allowing for differences across the levels of the Sentence factor

M_{full}: model assuming no differences across both the relevance and sentence levels

$M_{\text{ad hoc}}$: A post hoc modification of M_{sentence} to allow the sentences to differ in the Irrelevance TF cell.

A hypothesis (instantiated by a constrained model) is said to be rejected when it performs worse than the unconstrained model and/or any of the competing alternative hypotheses (even after taking differences in flexibility into account via the Fisher Information Approximation, FIA). In addition, the ratio G^2/df is included since for large samples any minor deviation from model predictions can lead to a statistically significant misfit. A ratio G^2/df between 0 and 2 is considered to indicate a good fit (Skovgaard-Olsen et al., 2017).

Table 7. Model-Comparison Results: Therefore/If then

Model	G^2	df	*p*	ΔFIA	G^2/df
$M_{\text{saturated}}$	0	0	1	9.20	--
M_{sentence}	3.94	4	.41	0	0.94
$M_{\text{relevance}}$	318.83	4	.00	157.44	79.71
M_{full}	319.17	6	.00	151.33	53.20
$M_{\text{ad\_hoc}}$	2.37	3	.50	2.02	.79

Note. df = degrees of freedom; G^2 = goodness of fit; p = p-value; ΔFIA = difference between the model's FIA and the FIA from the best-performing model.

Based on these criteria, a model (M_{sentence}) that collapses the difference between conditionals and therefore sentences is the winning model, as Table 7 shows.

The main findings are: 1) for the TT cell there is a preference for treating the reason-relation reading as at-issue content across the levels of the Relevance factor and the Sentence factor 2) in the Positive Relevance TF cell, there is a preference for treating the truth-table cell as at-issue across the levels of the Sentence factor, and 3) in the Irrelevance TF cell, there is a tendency for treating the reason-relation reading as at-issue across sentence levels.

2.4.3 Discussion

Taken together, the reason-relation reading turned out to be content-at-issue in the context of the present task, not only for the conditionals, but also for the therefore-sentences, which were used as a baseline representing a paradigmatic instance of Gricean conventional implicatures. Yet, interestingly, the participants' preferences for the TF conditions indicate that while a TF cell is sufficient to elicit a 'no' response, and thereby trump the reason relation in the Positive Relevance TF condition, the reason relation is weighted slightly higher when there are two justifications for answering 'no' in the Irrelevance TF condition.

In Skovgaard-Olsen, et al. (2017), it was found that there was little influence of relevance on truth-value assignments to both conditionals and 'therefore' sentences in a truth table task where the participants were instructed to calibrate the output of a computer program in the development phase. One of the main differences between the computer calibration task used there and the present Pierre task is the following: (1) in the Pierre task the sentences were produced based on known truth-table cells, whereas the sentences were produced before the truth-table cells became known in the computer calibration task, and (2) in the Pierre task a forced choice between Yes/No responses in the presence of justifications is required, whereas the computer calibration task involved a ternary assignment {True, False, Neither Nor} without justifications.

It has become popular to think about conventional implicatures as content-not-at-issue based on the influential work of Bach (1999) and Potts (2005). In the present experiments, we could not find evidence that the reason-relation reading of conditionals is content not-at-issue. However, as argued in Salmon (2011), the Bach-Potts notion of conventional implicatures, which centers around the example of non-restrictive relative clauses and appositives (e.g., “Mozart, who is a famous composer, started to play the piano at an early age”), is subtly different from the notion of conventional implicatures that figures in the work of Grice (1989), where the main examples came from sentences containing words like 'but', 'therefore', and 'even'.

These two notions of conventional implicatures are supposed to share the properties of a) not affecting the truth values of the sentences in which they occur, and b) being conventional aspects of the meaning of the sentences in which they occur, which cannot be cancelled without contradiction (as opposed to the pragmatic content of conversational implicatures). It was an assumption of our experiments that they would also share the property of being content not-at-issue. Based on the results of our present study, it appears that neither Gricean conventional implicatures (here represented prototypically by the therefore-sentences), nor conditionals have this property. Conversely, the success of the model $M_{sentence}$ implies that conditionals were responded to just like therefore-sentences in our test of at-issue content.

2.5 Experiment 4: Control Study

It is possible that, in Experiment 3, participants took the truth cells provided to be evidence for or against a reason relation. In Positive Relevance items, the TF cell might be taken as evidence against Positive Relevance. For instance, although we would normally assume that Mark pressing the on switch on his TV is a reason for believing that his TV will be turned on, it might be thought that this reason relation is undermined by learning that while Mark has pressed the on switch on his TV in fact his TV is not turned on. A similar problem arises in Irrelevance items, where the TT and TF conditions might be taken as evidence of reason relations. If participants did, indeed, reason this way, this reasoning would undermine the relevance relation manipulation through the scenarios based on the participants' background knowledge. Experiment 4 tests this alternative hypothesis and is a control study for Experiment 3. Its purpose is to ensure that the TF condition does not undermine the Positive Relevance manipulation, and that the TT and TF conditions do not undermine the Irrelevance manipulation.

2.5.1 Method

Participants

Like Experiment 3, Experiment 4 was conducted over the Internet using Mechanical Turk and sampling from USA, UK, Canada, and Australia. 250 people participated in the experiment in exchange for a small payment. The same exclusion criteria were applied as in Experiment 3. The final sample consisted of 155 participants. Mean age was 36.9 years, ranging from 21 to 73 years; 52.9 % of the participants were male; 64.5 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The sample differed only minimally on the demographic variables above before and after applying the exclusion criteria.

Design

Experiment 4 had the same experimental design as Experiment 3.

Materials and Procedure

Experiment 4 followed the same procedure as Experiment 3. The only difference was that instead of making judgments of direct rejections and affirmations, the participants were now asked to rate the extent to which the antecedent of Pierre's statement was a reason for/against the consequent on a five-point scale {strong reason against; reason against; neither for nor against; reason for; strong reason for}. These two randomly ordered pages differed on whether the participants were presented with the TT or TF cell as the facts that they and Pierre learned about.

2.5.2 Results

Participants assessed on a five-point scale the extent to which the first sentence in our Positive Relevance and Irrelevance conditions was a reason for/against the second. In Figure 8 the histograms are displayed. In Positive Relevance TT, the participants' reason-relation rating was ‘strong reason for’ (*Mdn* = 5) on the group level; in Positive Relevance TF it was ‘reason for’ (*Mdn* = 4). The reason relation in the Irrelevance condition was not affected by the presence of truth table cells and was in both cases assessed as ‘neither for nor against’ (Irrelevance TT *Mdn* = 3; Irrelevance TF *Mdn* = 3).

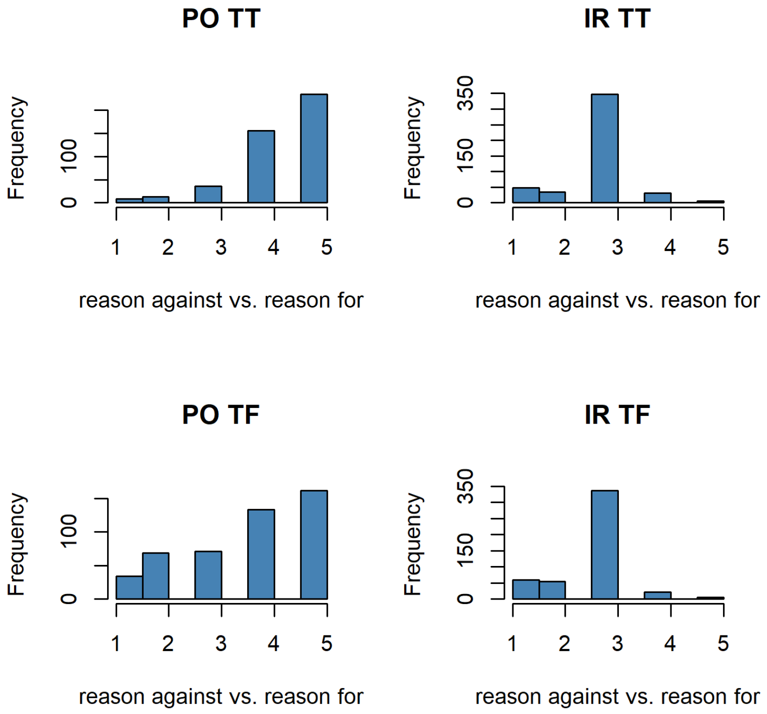


Figure 8. Histogram of Reason-Relation Assessments. Note. 1 = strong reason againist; 2 = reason against; 3 = neither for nor against; 4 = reason for; 5 = strong reason for. ‘PO’ = Positive Relevance; ‘IR’ = Irrelevance.

Linear Mixed Models

To analyze the effect of the truth table cell on the reason-relation assessments, linear mixed-models were fitted to the participants' responses. Like in Experiment 1 where a Likert scale was also used, the responses were modelled as generated by thresholds set on a latent continuous scale with a cumulative likelihood function and a logit link function (Bürkner & Vuorre, 2018). The models included the following fixed effects: (M10) modelled the rating as a function of the Item factor (TT vs. TF), the Relevance factor (Positive Relevance vs. Irrelevance), and their interaction, (M11) like M10 but without the interaction, and (M12) like M11 but without the main effect for the Relevance factor.

Table 8 reports the performance of these models as quantified by the leave-one-out cross validation criterion and the WAIC information criterion.

Table 8. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M10	4024.12	0	--	4002.1	0.9939
M11	4042.36	18.24	6.05	4019.0	0.0002
M12	4035.24	11.12	4.66	4012.4	0.0059

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC.

The information criteria clearly favour M10. This preference for M10 reflects the fact that there was a strong interaction between the Item and Relevance factors ($b_{\text{ItemTT:PositiveRelevance}}$ = 1.69, 95%-CI [1.08, 2.31], BF_{H0H1} = $4.60*10^{-11} \approx 0$). In addition, a strong main effect for the Relevance factor was found ($b_{\text{PositiveRelevance}}$ = 2.84, 95%-CI [2.14, 3.55], $\text{BF}_{\text{H0H1}} = 6.61*10^{-16} \approx 0$), but no main effect for the Item factor was found (b_{ItemTT} = 0.24, 95%-CI [-0.04, 0.52], BF_{H0H1} = 8.57). Presumably, the lack of a main effect of the truth table cells on the reason-relation ratings is due to the fact that the TT vs. TF difference only affects the Positive Relevance condition. This combination of effects indicates that it remains the case that Positive Relevance contents are evaluated higher than Irrelevance contents, even in spite of the TT/TF manipulation. However, to assess impact of the TT/TF manipulation more closely, Figures 9a and 9b were made. Figure 9a plots the central tendencies of M10, whereas Figure 9b plots the posterior predictions of M10 for the individual response categories:

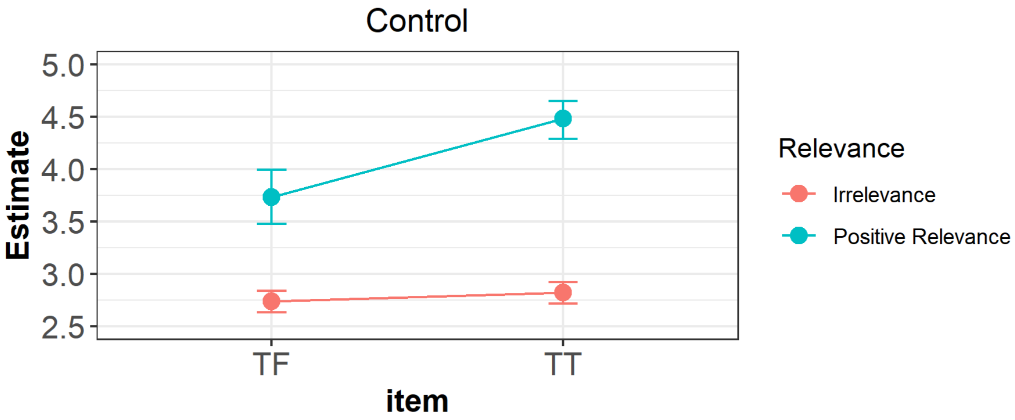


Figure 9a. Central Tendencies of M10. Note. Estimated marginal means for the reason-relation assessments in the different truth table cell x relevance conditions.

As Figure 9a shows, the central tendency remains that the participants agree that the antecedent constitutes a reason for the consequent in the Positive Relevance condition, even when the TF manipulation is added, and that the participants responded that the antecedent is neutral with respect to the consequent on the perceived reason relation scale for the Irrelevance condition.

However, as Figure 9b indicates, for the Positive Relevance condition there is a decrease from ca. 90 % agreeing that the antecedent constitutes a reason or a strong reason for the consequent in the TT cell to ca. 60% of the participants agreeing that it constitutes a reason for or a strong reason for the consequent in the TF cell. In contrast, for the Irrelevance condition there is almost no change with ca. 67 % agreeing that the antecedent is Neutral with respect to the consequent on the perceived reason relation scale for the TT items, and ca. 70% making the same judgment for the TF items.

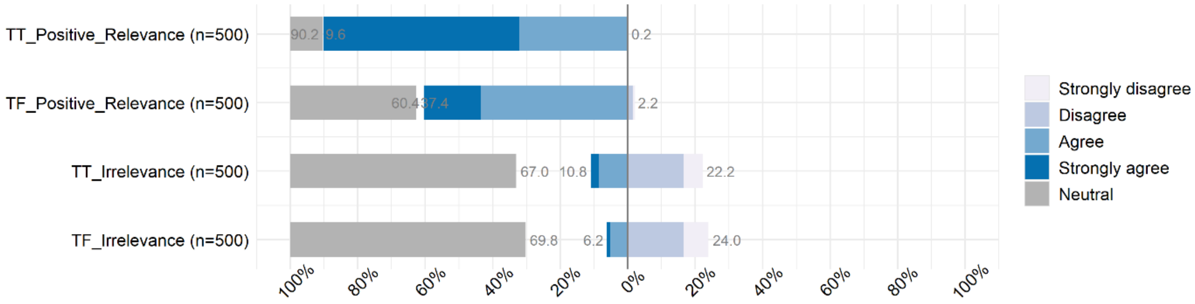


Figure 9b. Posterior Predictions of Individual Categories, M10. Note. The perceived strength and direction of the reason relation across truth table cells (TT vs. TF) was measured on a scale from ‘strong reason against’ (1), ‘reason against’ (2), ‘neutral’ (3), ‘reason for’ (4), and ‘strong reason for’.

Yet, as a central tendency it remains the case that the Positive Relevance items are rated one point higher on the reason-relation scale than the Irrelevance items on average, even in the presence of the TF cell.

2.5.3 Discussion

From the analysis we can conclude that while the truth-table cell diminishes the effect of the reason-relation manipulation, it is far from suppressing it. The median response is still to treat A as a reason for C in the Positive Relevance TF condition (in spite of an increase in ‘Neutral’ response), and for the Irrelevance conditions the truth-table cells did not have any effect.

Hence, alternative accounts of our findings based on the objection that the presence of the truth-table cells undermines the reason-relation manipulation are not supported by the data.

2.6 General Discussion

In this paper, we have been concerned with diagnosing whether the reason-relation reading is due to a pragmatic or semantic component of indicative conditionals. In addressing this question, we have empirically contrasted it with well-known linguistic phenomena - namely, conversational implicatures, entailments, presuppositions, and conventional implicatures. In the course of these investigations, we discovered that 1) attempting to cancel a commitment to the reason-relation reading of indicative conditionals is viewed as just as contradictory as cancelling a commitment to an entailment (whereas attempting to cancel a commitment to the reason-relation reading of conjunction is viewed as less contradictory than cancelling a commitment to a scalar implicature), 2) to negate a conditional in wide scope is not in general viewed as equivalent to negating the consequent of a conditional, and 3) the reason-relation reading can become at-issue content not only for conditionals but also for therefore-sentences, which constitute a paradigmatic example of Gricean conventional implicatures.

Can Mental Model Theory explain our Data?

As we pointed out in the introduction, the Mental Models Theory does not treat the relation between the antecedent and consequent of a conditional as a part of its core meaning. However, it postulates "a mechanism of *modulation* that can transfer [the core] meaning into an indefinite number of different sorts of interpretation" (Quelhas, Johnson-Laird, and Juhos, 2010, p. 1717). More specifically, which mental models are and which are not constructed in the process of the interpretation of a conditional can be affected by the semantics, that is, the meanings of its clauses, or by pragmatics, that is, the knowledge about the context of utterance and general world knowledge related to what the conditional is about. For instance, modulation can result in blocking the construction of a model corresponding to the possibility of ¬A and C which yields the bi-conditional interpretation. To give an example, in a context in which a TV is off, an interpretation of the conditional "If Mark presses the on switch, his TV will be on" might consist of the following models:

switch on	TV on
switch off	TV off

The possibility that Mark doesn't press the on switch but the TV is on is here excluded by the background knowledge (see Quelhas et al. 2010, p. 1720, for a list of possible interpretations of the conditional). Semantic modulation can also prevent us from accepting conditionals such as:

(4) If God exists, then atheism is correct.

Here, the meaning of "atheism is correct" entails that God does not exist, the mental model consisting of the antecedent and consequent of this conditional is not possible, and hence the conditional is false (Johnson-Laird et al. 2015, p. 206, Quelhas, Rasga, & Johnson-Laird, 2017, p. 24). Note, however, that here the mechanism of modulation takes the presence of the analytic relationship between the clauses of a conditional as its input, rather than an output. In other words, the presence of an analytic connection between the words used in the antecedent and consequent is the reason why the construction of certain mental models is blocked, thus blocking the construction of mental models (comparable to deleting rows in the graphical representation of the explicit model) does not explain why a broader range of conditionals expresses the existence of a connection.

More importantly, the type of a connection conveyed by the conditionals we investigated goes beyond such analytic relationships. Granted, the advocates of the Mental Models Theory propose that the mechanism of modulation can also add information to the model of a sentence. In particular, it can add relations between the clauses of a conditional (see, e.g., Johnson-Laird & Byrne, 2002, p. 651, Quelhas et al., 2010, pp. 1728-9, Khemlani et al., 2018, pp. 12-13). For instance, Quelhas et al. (2010, p. 1728) observe that:

> appropriate contents should introduce a temporal relation between antecedent and consequent events—for example:
>
> (5) If Lisa received some money, then she paid Frederico.
>
> Individuals know that payment can be made only if a payer has money, and so modulation should yield an interpretation of the conditional in which if Lisa received some money then she did so before she paid Frederico.

Given the above description, the mechanism of modulation involved in this type of case can be construed as some kind of a pragmatic inference: the form of a conditional plus general background knowledge relevant for the interpretation of its antecedent and consequent allow people to *infer* the temporal order (or, in other cases, spatial relations, causal dependencies, and other possible relationships) of the events the antecedent and consequent are about. In other words, unless we deal with the analytical relationships between words that occur in the antecedents and consequents, what is responsible for the variety of different interpretations of conditionals, on the Mental Models Theory, are pragmatic processes. Consequently, while the mechanism of modulation tells us where the perceived connections between antecedents and consequents might be coming from, the Mental Models Theory has no resources to explain why, among other things, participants judge such connections not to be cancellable without giving rise to a contradiction, nor why participants tend to find conditionals whose antecedents and consequents are not connected to be somehow defective.

Finally, we would like to note that the mechanism underlying modulation described in Johnson-Laird & Byrne (2002) and Khemlani et al. (2018) relies on deleting or adding rows of the truth table, if there are salient pragmatic factors, or lexical content in the clauses of the conditional, which exclude these truth-table cells (like in the examples discussed above). Yet part of the data, which the conventional implicature hypothesis is introduced to account for, is the strong dissociation in terms of the influence of relevance on probability/acceptability evaluations and the lack of influence of relevance on truth evaluations (Skovgaard-Olsen et al., 2017). Since this finding shows that the impact of relevance on conditionals is mostly found for types of cognitive assessments other than truth evaluations, an explanation which posits that truth-table cells are deleted or added in missing-link conditionals is unlikely to account for the complex data pattern. Consequently, semantic and pragmatic modulation of the type described above cannot explain the participants' reaction across experiments that conditionals whose antecedents and consequents are not connected are somehow defective.

Repercussions for Argumentation with Conditionals

In and of themselves, our findings have interesting repercussions for argumentation with conditionals. What they suggest is that uttering regular[30] indicative conditionals commits speakers to there being a reason relation between its antecedent and consequent, which speakers cannot escape from without retracting the original utterance, if they are to avoid leaving the impression that they are contradicting themselves. Moreover, when the interlocutor negates a speaker's conditional assertion "If A, then C", then the interlocutor need not be taken as committing to "If A, then ¬C". Finally, the results indicate that there are situations where the reason relations expressed by both conditionals and 'therefore' sentences may become content-at-issue, and indeed that there is a stronger tendency to take these reason relations as content-at-issue than the truth values of their constituents. This in turn suggests that the interlocutor need not interrupt the natural flow of the conversation by expressions like "Hey, wait a minute!" to challenge a reason relation expressed by the speaker through indicative conditionals or 'therefore' sentences. Rather, the interlocutor can challenge the speaker's reason-relation commitments directly and treat these commitments as the main point of the assertion.

To illustrate how all these phenomena may shape argumentative discourse, consider the example of a European conference on global warming in the winter of 2010-2011, which

[30] The qualification is meant to set aside the problematic case of biscuit conditionals (e.g. 'if you are hungry, there are biscuits on the sideboard') for present purposes. For further discussion see Biezma and Goebel (*in review*).

was notorious for featuring an unusually cold November in Central Europe, in some cases setting records for low temperatures. Or alternatively: think of the winter of 2017-2018 in the USA on the East coast. Suppose a politician representing climate change skepticism utters 'if global warming is real, then winter will be warmer than we are used to'. His interlocutors in turn might want to negate this utterance, without thereby committing themselves to the claim that 'if global warming is real, then winter will be colder than we are used to', which they likewise reject. According to the Negation Principle such a discourse move would, however, have been incoherent, whereas the present considerations and our results suggest that it is not.

Suppose further that a consensus forms at the conference that claims about global warming are concerned with global, long-term climate trends, which are unaffected by temporary, regional weather events. In that case, the politician from above might attempt to back-pedal and eschew a commitment to a reason relation between global warming and local weather events by performing a cancellation speech act. However, the results from Experiment 1 indicate that he or she would likely be viewed as contradicting himself/herself. Instead the politician would have to admit an error by retracting his/her earlier statement.

Finally, the results from Experiment 2 suggest that reason-relation commitments like the one above are not the sort of thing that tacitly get introduced into the common ground through indicative conditionals and therefore-sentences without themselves being the subject of direct rejections and affirmations. When it comes to controversial topics like climate change, which reason relations to accept is itself disputed territory and can become the main point of assertion. It would be a natural continuation of the discourse to make disputed reason relations the content-at-issue by targeting them with direct denial and affirmations. The interlocutors need not, in other words, interrupt the natural flow of conversation to challenge reason-relation commitments and we need not conceptualize reason-relations as uncontroversial assumptions that are automatically accommodated into the common ground.

Conventional Implicatures and At-Issue Content

As explained in Koev (2018), there are various notions of at-issue content discussed in the literature and it is, presumably, unclear how they are to be unified. One important property highlighted in Potts (2005, 2007) is that at-issue content constitutes direct proposals to update the common ground of mutually shared assumptions by the interlocutors. On this view, at-issue content is negotiable and open to direct agreement or disagreement by the addressee, and a key diagnostic is whether discourses are acceptable where the interlocutor provides a direct response ("Yes, .../No, ...") to the target content. In contrast, content not-at-issue is thought of as grammatically encoded content that is directly imposed on the common ground without negotiation.

A common assumption is that discourses are structured around questions under discussion. A second important property of at-issue content is that it provides potential answers to the question under discussion (Koev, 2018).

Following Bach (1999), Potts (2005, 2007) used the first notion of at-issue content in his treatment of appositives, which he treats as a paradigmatic instance of conventional implicatures. On this view, conventional implicatures are "secondary entailments that cooperative speakers rarely use to express controversial propositions or carry the main themes of a discourse" (Potts, 2007: 476). For instance, in 'Mozart, the famous composer, used to live here', the assumption that Mozart was a famous composer is presented as a shared assumption not really up for discussion. It is grammatically marked as not the central focus of the assertion. Yet, its content is conventionally part of the meaning of the sentence, and not produced by, say, a conversational implicature.

In continuation of this line of work, Experiment 3 set out to probe whether the reason-relation reading of conditionals is content-at-issue in an attempt to determine whether the reason-relation reading is a conventional implicature. The results showed that the participants clearly treated the reason-relation reading as content-at-issue for conditionals. Interestingly, we found the same pattern of results with respect to therefore-sentences, which Grice (1989) treated as a paradigmatic instance of conventional implicatures.

Our results thus stand in tension with the newer literature on conventional implicatures which treats not-at-issue content as a diagnostic feature. However, as shown in Salmon (2011), it turns out that Pott's (2005, 2007) notion of a conventional implicature differs subtly from Grice's (1989) both in terms of its properties and central instances. Gricean conventional implicatures are non-truth conditional, non-cancellable, not calculable based on the Gricean maxims of conversation, and detachable (e.g., substituting 'and' with 'but' in the following sentence in the same context of utterance will lose the implication of a contrast between being poor and honest: "She was poor but honest"). In contrast, Pott's conventional implicatures are also speaker-oriented in that the speaker incurs a commitment to them even when making indirect speech reports (e.g., in "John said that Ames, the former spy, is now behind bars", the speaker is also committed to Ames being a former spy). Gricean conventional implicatures lack this property (Salmon, 2011). This fits with the idea of reason relation as Gricean conventional implicatures, because in neither of the two following examples is it the case that the speaker incurs a commitment to a reason relation by making an indirect speech report:

“The politician said that if global warming is real, then winter will be warmer than we are used to”, “The politician said that the winter is surprisingly cold therefore global warming is bogus”.

A further difference between the two notions of conventional implicatures may be that Potts's notion (which centers around appositives and non-restrictive relative clauses) differs from Grice's (which centers around utterance modifiers like 'therefore', 'but', and 'even')[31] exactly with respect to the at-issue status of the content.

A first indicator that something is amiss is that in a textbook example like "she is poor but honest", it is far from the case that something uncontroversial, which is not up for negotiation, is expressed by the implied contrast between poverty and honesty. Furthermore, the skeptic’s assertion of 'if global warming is real, then winter will be warmer than we are used to' is used to express a highly controversial claim. There is no reason why his interlocutors should treat it as part of the common ground which is not up for negotiation. The same would apply if he had formulated his statement as 'The winter of 2010-2011 is unusually cold; *therefore* climate change is bogus'. Moreover, it holds for both these statements that the speaker clearly intends them as partial answers to the question under discussion of the conference. Indeed, we would expect these statements to elicit a discussion about whether the politician can really use local weather phenomena like the winter of 2010-2011 as decisive evidence against climate change without interrupting the natural flow of the discourse at the conference.

The results from Experiment 3 indicate that the participants do treat the reason-relation reading of conditionals and 'therefore' sentences as content-at-issue. Perhaps one reason is the worry that Pierre, the foreign language learner, might advance controversial, nonsensical reason relations unless he is directly confronted. Moreover, the presence of justifications in the answer options that explicitly target the presence or absence of reason relations in Pierre's utterance may also have contributed to making the reason relations content-at-issue.

What the considerations above suggest is that lack of at-issue content may not be a good characterization of (Gricean) conventional implicatures to begin with. For argumentative discourse we need some way of coordinating which reason relations to accept. When opposing views clash a central part of the dispute is which arguments to accept. If argumentative discourses always came down to which factual premises to accept, then they

[31] If the argument in Salmon (2011) goes through, then the indirectness of the evidence possessed signalled by the epistemic modal 'must' can be added to the list.

could be resolved by identifying the most reliable source of evidence on the topic and simply accepting its verdicts.

The picture that emerges out of these considerations is that the reason-relation reading of indicative conditionals is a conventional implicature, which is tapped into through probability and acceptability evaluations. Yet at times reason relations may become content at-issue in the sense of addressing the question under discussion and constituting direct proposals to update the common ground of mutually shared assumptions by the interlocutors. Since, however, (non-deductive) reason relations constitute probabilistic constraints (i.e., $P(C|A) - P(C|\neg A) > 0$), what this requires is that we go beyond Stalnaker (2016) in thinking of the common ground in terms of a set of propositions by enriching it with probabilistic structure. But this is something that we anyway have ample reason for doing when modeling epistemic and doxastic content (Yalcin, 2012).

2.7 Conclusion

In relation to the diagnostic problem with which we started; our investigations permit us to draw the following conclusions.

The Relevance Effect reported in Skovgaard-Olsen et al. (2016a) is probably not due to the influence of a conversational implicature. From Experiment 1 no support could be derived for the hypothesis that the reason-relation reading of indicative conditionals is generated by the presence of a conversational implicature. However, support could be obtained that the reason-relation reading of conjunctions could be the result of a conversational implicature. Experiment 1 thus also contributes to drawing a dividing line between the content of logical operators. In discussions, the argument is often put forward that the Relevance Effect on conditionals cannot be taken to reveal something about the semantic content of conditionals, because conjunctions also have a reason-relation reading and presumably we would not want to make it part of the semantic content of conjunctions. The results of Experiment 1 directly undercut any such argument by showing the different status that the reason-relation reading of indicative conditionals and conjunctions have with respect to conversational implicatures.

In the discussion of these results, we considered, however, alternative interpretations of our results based on varying strengths of conversational implicatures, and the possibility of other ways of phrasing the cancellation task. In response, it was pointed out that these alternative hypotheses are confronted with the burden of explaining why our way of posing the task worked so well with our comparison cases such as the conjunctions, if the results with respect to conditionals are not taken at face value.

A further conclusion of this paper is that the Relevance Effect on conditionals is probably not due to the presence of a presupposition failure of the irrelevance items. One of the most characteristic properties of presuppositions is their ability to project when embedded under logical operators, like negation. However, Experiment 2 could not find support for the hypothesis that the reason-relation reading of indicative conditionals projects when embedded under negations. And perhaps even more damaging to the presupposition failure hypothesis is the finding that while extreme individual differences could be found in the probability assignment to control items with presupposition failures, these individual differences were not reflected in the participants' probability assignment to missing-link conditionals. In our discussion of these results, we again considered, but rejected, an alternative interpretation of our results based on local accommodation.

Experiment 2 moreover yielded a further finding of interest in its own right. According to a well-known negation principle employed in various systems of conditional logic, wide scope negation equals narrow scope negation (Arlo-Costa, 2007). This principle has centrally figured in Suppositional Theory of conditionals' account of compound conditionals involving negations (Edgington 1997, 2000, 2006; Woods, 1997, ch. 6; Kölbel, 2000; and Bennett, 2003, ch. 7). Indeed, in Handley et al. (2006) this principle is even treated as a litmus test for the suppositional conditional. However, the data clearly show that while this negation principle can be maintained for the Positive Relevance condition, it is systematically violated for the Irrelevance condition. Hence, Experiment 2 provides an occasion to reevaluate how indicative conditionals interact with the negation operator.

Turning to Experiment 3, the reason-relation reading of indicative conditionals was found to be treated as more at-issue than the truth functionality of the clauses. However, this latter finding was also found with 'therefore' sentences. And, indeed, if we take the present results together with those in Skovgaard-Olsen et al. (2017), then we see that indicative conditionals and therefore-sentences behave in a similar way in a range of cognitive assessments which have a bearing on whether the Relevance Effect is a conventional implicature. To be sure, we did not find for either indicative conditionals or for therefore-sentences that the reason-relation reading was not-at-issue content. But it was found in Skovgaard-Olsen et al. (2017) for both types of sentences that the reason-relation reading shows strong fingerprints in tasks where probability or acceptability is asked; however, when the participants are asked to fill out truth-tables they almost entirely ignore the reason-relation reading. And so, the argument for the (Gricean) conventional implicature interpretation of the Relevance Effect now stands supported by a) the negative results for the relevance effect being a conversational implicature, b) the negative results for the relevance effect being a presupposition failure, c) the finding of the dissociation with relevance strongly affecting probability and acceptability but hardly affecting truth value assignments in the truth table task, d) the fact that indicative conditionals behave remarkably like 'therefore' sentences when probed for their truth, acceptability, probability, or when compared on their at-issue content.

In our discussion, we considered further alternative hypotheses with respect to a) and b) and expressed our doubts. But we welcome future empirical studies that may challenge these conclusions through variations of experimental tasks and attempts to provide unifying explanations of the complex data pattern that is emerging.

In the meantime, we conclude that our data suggest that the Relevance Effect is not a conversational implicature and is not due to presupposition failure. The best candidate, instead, is most likely a conventional implicature. These findings suggest a new direction for the debate on whether relevance is part of the semantics or pragmatics of the conditional. A final judgment will rest on the definition of semantics and pragmatics, and on how conventional implicatures are categorized according to that definition. It may be, however, that conventional implicatures - and, hence, the reason relation reading of conditionals - remain an intermediate, irreducible layer of meaning between semantics and pragmatics.

References

Arlo-Costa, Horacio (2007). The Logic of Conditionals. *The Stanford Encyclopedia of Philosophy* (spring 2016 Edition), Edward N. Zalta (ed.). URL = <http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/>.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 340-412.

Bach, K. (1999). The Myth of Conventional Implicature. *Linguistics and Philosophy, 22*(4), 327-366.

Baratgin, J., Douven, I., Evans, J. St. B. T., Oaksford, M., Over, D.E., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, 19, 547–548.

Baratgin, J., Politzer, G, Over, D. E., and Takahashi, T. (2018), The Psychology of Uncertainty and Three-Valued Truth Tables, *Frontiers in Psychology, 9:1479.*

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86.

Beaver, D. I. and Geurts, B. (2014). Presupposition. In E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*, URL = <http://plato.stanford.edu/archives/win2014/entries/presupposition/>.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Biezma, M. (2014). The grammar of discourse: The case of *then*. In T. Snider et al. (eds.), *Proceedings of SALT 24* (pp. 373-394). Cornell U: LSA and CLC Publications.

Biezma, M. and Goebel, A. (*in review*). Being pragmatic about biscuits. Retrieved from https://mariabiezma.com/biezma-some-recent-work/

Birner, B. J. (2014). *Introduction to Pragmatics*. Malden, MA: Wiley-Blackwell.

Blome-Tillmann, M. (2013). Conversational Implicatures (and How to Spot Them). *Philosophy Compass*, 8/2, 170-85.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1-28.

Bürkner, P., & Vuorre, M. (2018, June 23). Ordinal Regression Models in Psychology: A Tutorial. https://doi.org/10.31234/osf.io/x8swp

Carston, R. (1993). Conjunction, explanation and relevance. *Lingua*, 90, 27-48.

Carston, R. (2002). *Thoughts and Utterances*. Oxford: Blackwell Publishers.

Chierchia, G. and McConnell-Ginet, S. (1990). *Meaning and Grammar*. Cambridge, MA: the MIT Press.

Dancygier, B. and Sweetser, E. (2005). *Mental Spaces in Grammar: Conditional Constructions*. Cambridge University Press, Cambridge.

Davis, W. (2014). "Implicature", *The Stanford Encyclopedia of Philosophy* (Fall 2014). In: Zalta, E. N. (Ed.). Retrieved from https://plato.stanford.edu/archives/fall2014/entries/implicature/

Declerck, R. and Reed, S. (2001). *Conditionals: A Comprehensive Empirical Analysis*. Mouton de Gruyter: Berlin/New York.

Douven, I. (2015). *The Epistemology of Indicative Conditionals. Formal and Empirical Approaches*. Cambridge: Cambridge University Press.

--- (2017). How to account for the oddness of missing-link conditionals. *Synthese*, 194 (5), 1541-1554.

Douven, I. and Verbrugge, S. (2010). The Adams family. *Cognition*, *117*, 302–318.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235-329.

--- (1997). Commentary. In: Woods, M. *Conditionals*. Oxford: Oxford University Press, 95-138.

--- (2000). General Conditional Statements: A Response to Kölbel. *Mind,* 109 (433), 109-116.

--- (2006). Conditionals. In: Zalta, E. N. (ed.),*The Stanford Encyclopedia of Philosophy*. (Winter 2008 Edition). URL = <http://plato.stanford.edu/archives/win2008/entries/conditionals/>.

Egré, P., & Politzer, G. (2013). On the negation of indicative conditionals. In M. Aloni, M. Franke, and F. Roelofson (Eds.), *Proceedings of the Amsterdam Colloquium* (pp. 10-18).

Espino, O., and Byrne, R.M.J. (2012). It's not the case that if you understand a conditional you know how to negate it. *Journal of Cognitive Psychology, 24*(3), 329-334.

Evans, J. S. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 321–335.

Fugard, J. B., Pfeifer, N., Mayerhofer, B., & Kleiter, G. (2011). How people interpret conditionals: Shifts towards the conditional event. *Journal of Experimental Psychology:Learning, Memory,and Cognition, 37*, 635–648.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. New York: Academic Press.

Grice, H.P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Handley, S.J., Evans, J. St. B.T., Thompson, V.A. (2006). The negated conditional: a litmus test for the suppositional conditional? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 559-569.

Heim, I. (1983). On the projection problem for presuppositions. In Barlow, M. and Flickinger, D. and Westcoat, M. (eds.), *Second Annual West Coast Conference on Formal Linguistics*. Stanford University, 114–126.

Hinterecker, T., Knauff, M., Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology*: *Learning, Memory, & Cognition*, *42,* 1606-1620.

Jeffrey, R. C. (1991). Matter of fact conditionals. *Aristotelian Society Supplementary Volume, 65*, 161–183.

Johnson-Laird, P. N. and Byrne, R. M. J. (1991). *Deduction*. Hilsdale, NJ: Lawrence Erlbaum Associates Inc.

Johnson-Laird, P. N. and Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678.

Johnson-Laird, P. N. and Khemlani, S. S. (2014). Toward a Unified Theory of Reasoning. *Psychology of Learning and Motivation*, *59*, 1-42.

Johnson-Laird, P. N., Khemlani, S. S., and Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Science*, *19*(4), 201-214.

Kadmon, N. (2001). *Formal Pragmatics*. Malden, MA: Blackwell Publishers.

Karttunen, L. (1973). Presuppositions of Compound Sentences. *Linguistic Inquiry*, 4, 167–193.

Khemlani, S., Byrne, R. M. J., and Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science*, *42*(6),1-18.

Khemlani, S., Orenes, I., & Johnson-Laird, P.N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*, 541-559.

Kleiter, G., Fugard, A, & Pfeifer, N. (2018). A process model of the understanding of conditionals. Thinking & Reasoning, 24, 386-422.

Koev, T. (2018). Notions of At-issueness. *Language and Linguistics Compass.* https://doi.org/10.1111/lnc3.12306

Kölbel, M (2000). Edgington on Compounds of Conditionals. *Mind*, 109(433), 97-108.

Kratzer, A. (1986). Conditionals. *Chicago Linguistics Society*, *22*(2), 1–15.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Khemlani, S. S., Byrne, R. M. and Johnson-Laird, P. N. (2018), Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. Cognitive Science, 42: 1887-1924. doi:10.1111/cogs.12634

Krzyżanowska, K. (2015). *Between "If" and "Then": Towards an empirically informed philosophy of conditionals*. PhD dissertation, Groningen University.

Retrieved from http://karolinakrzyzanowska.com/pdfs/krzyzanowska-phd-final.pdf

Krzyżanowska, K, Collins, P. J. and Hahn, U. (2017a). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199-205.

--- (2017b). The Puzzle of Conditionals with True Clauses: Against the Gricean Account. In M. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Ed.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2476-2481). London, UK: Cognitive Science Society.

--- (2018). True clauses and false connections. Unpublished manuscript.

Krzyżanowska, K., Wenmackers, S., and Douven, I. (2014). Rethinking Gibbard's Riverboat Argument. *Studia Logica 102(4), 771-792*.

Lassiter, D. (2012). Presuppositions, provisos, and probability. *Semantics & Pragmatics, 5*, 1-37.

Lewis (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, D. 1976. Probabilities of conditionals and conditional probabilities. *Philosophical Review, 85*, 297–315.

McCawley, J. D. (1993). *Everything that Linguists have Always Wanted to Know about Logic, but were ashamed to ask*. Chicago: The University of Chicago Press. (Second edition)

McElreath, R. (2016). *Statistical Rethinking.* Boca Raton, FL: CRC Press.

Moss, S. (2015). On the Semantics and Pragmatics of Epistemic Vocabulary. *Semantics and Pragmatics, 8*(5), 1-81.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M. and Over, D. and Cruz, N. (2018). Paradigms, possibilities and probabilities: Comment on Hinterecker et al. (2016). *Journal of Experimental Psychology: Learning, Memory, & Cognition*. (In Press)

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 680–693.

Olsen, N. S. (2014). *Making Ranking Theory Useful for Psychology of Reasoning*. PhD dissertation, University of Konstanz. URL = http://kops.uni-konstanz.de/handle/123456789/29353.

Ospina, R. and Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, *56*, 1609-1623.

Over, D. E. (2007). New paradigm psychology of reasoning. *Thinking & Reasoning*, *4*, 431-438.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54(1), 62–97. http://dx.doi.org/10.1016/j.cogpsych.2006.05.002.

Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In Linden J. Ball and Valerie A. Thompson (Eds.), International handbook of thinking and reasoning (pp. 434-450). Hove, UK: Psychology Press.

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking and Reasoning*, *16*, 172–197.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

--- (2007). Conventional implicatures, a distinguished class of meanings. In Gillian Ramchand and Charles Reiss (Eds.). *The Oxford Handbook of Linguistic Interfaces* (pp. 475-501). Oxford: Oxford University Press.

--- (2008). Wait a minute! What kind of discourse strategy is this? (Annotated data set) Retrieved from http://christopherpotts.net/ling/data/waitaminute/.

--- (2015). Presupposition and implicature. In Shalom Lappin and Chris Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (pp. 168-202), 2nd edn,. Oxford: Wiley-Blackwell.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, 63, 1716–1739.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2017). A priori true and false conditionals. *Cognitive Science*, 41, 1003–1030.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsey, F. P. (1929/1990). General propositions and causality. In Mellor, D. H. (Eds.), *Philosophical Papers* (pp. 145-163). Cambridge University Press, Cambridge.

Recanati, F. (2011). *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Russell, B. (1905). On Denoting. *Mind*, 14, 479–493.

Salmon, W. (2011). Conventional implicature, presupposition, and the meaning of *must*. *Journal of Pragmatics*, *43*(14), 3416-3430.

Skovgaard-Olsen, N. (2016). Motivating the Relevance Approach to Conditionals, *Mind & Language*, 31(5), 555-579.

Skovgaard-Olsen, N. (*in review*). The (Lack of) Relevance of 'Then' and the Dialogical Entailment Task.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., and Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'. *Thinking & Reasoning*, 23 (4), 449-482.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016a). The relevance effect and conditionals. *Cognition*, 150, 26-36. doi:10.1016/j.cognition.2015.12.017

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016b). Relevance and Reason Relations. *Cognitive Science*, 41(S5), 1202-1215.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science, 37*, 1074–1106.

Stalnaker, R. (1968) “A Theory of Conditionals,” *Studies in Logical Theory, American Philosophical Quarterly*, Monograph: 2, 98-112.

Stalnaker, R. (2016). *Context*. Oxford: Oxford University Press.

Strawson, P. F. (1950). On referring. *Mind*, 59, 320–44. Verbrugge, S., Dieussaert, K., Schaeken, W., Smessaert, H., and Belle, W. V. (2007). Pronounced inferences: A study on inferential conditionals. *Thinking & Reasoning, 13*(2):105–133.

Tonhauser, J. (2012). Diagnosing (not-)at-issue content. In *Proceedings of Semantics of Under-represented Languages of the Americas (SULA) 6* (pp. 239-254). UMass, Amherst: GLSA.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics, 33*(1), 107-135.

von Fintel, K. (2004). Would you believe it? The king of France is back! Presuppositions and truth-value intuitions. In: Reimer, M. and Bezuidenhout, A. (eds.), *Descriptions and Beyond*. Oxford: Oxford University Press, 269–296.

von Fintel, K. (2008). What is presupposition accommodation, again? *Philosophical Perspectives*, *22*(1), 137-170.

Woods, M. (1997). *Conditionals*. Oxford: Oxford University Press.

Xue, J. and Onea, E. (2011). Correlation between presupposition projection and at-issueness: An empirical study. *Proceedings of the ESSLLI 2011 Workshop on Projective Meaning*, Ljubljana, Slovenia.

Yalcin, S. (2012). Context Probabilism. In Aloni, M., Kimmelman, V., Roelofsen, F., Sassoon, G., Schulz, K., and Westera, M. (Eds.), *Proceedings of the 18th Amsterdam Colloquium* (pp. 12-21).

Chapter 3: The Dialogical Entailment Task[32]

Niels Skovgaard-Olsen

In this paper, a critical discussion is made of the role of entailments in the so-called New Paradigm of psychology of reasoning based on Bayesian models of rationality (Elqayam & Over, 2013). It is argued that assessments of probabilistic coherence cannot stand on their own, but that they need to be integrated with empirical studies of intuitive entailment judgments. This need is motivated not just by the requirements of probability theory itself, but also by a need to enhance the interdisciplinary integration of the psychology of reasoning with formal semantics in linguistics. The constructive goal of the paper is to introduce a new experimental paradigm, called the Dialogical Entailment task, to supplement current trends in the psychology of reasoning towards investigating knowledge-rich, social reasoning under uncertainty (Oaksford & Chater, 2019). As a case study, this experimental paradigm is applied to reasoning with conditionals and negation operators (e.g., CEM and wide and narrow-scope negation). As part of the investigation, participants' entailment judgments are evaluated against their probability evaluations to assess participants' cross-task consistency over two experimental sessions.

3.1 Introduction

The empirical measurement of accepted entailments has been the subject of some recent controversy in the psychology of reasoning. In an influential paper, Evans (2002)

[32] This chapter has been published as:

Skovgaard-Olsen, N. (2019). The Dialogical Entailment Task. *Cognition*, *193*.

Authors' Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com). The supplemental materials including all data and analysis scripts are available at: https://osf.io/npc69/

Acknowledgement: I would like to thank Mike Oaksford, Eric Raidl, Nicole Cruz, David Over, Stefan Kaufmann, Keith Stenning, David Kellen, Vincenzo Crupi, Seth Yalcin, and Andrew Bacon for comments/conversations, as well as the audiences at London Reasoning Workshop (2018), What If, Konstanz (2018), and Dagstuhl Seminar (2019). Thanks also goes to Alison Scheel for her help in setting up some of the experiments.

criticizes five decades of reasoning research for following a deductive paradigm that has investigated participants' reasoning competence with a particular type of task that many participants find unnatural, based on a normative model of correct reasoning derived from classical logic. More specifically, participants were usually asked to reason with abstract stimulus materials (e.g., letters and numbers) in tasks, where they were asked to assess logical arguments, or produce logically valid conclusions, with little or no instructions on how to understand central semantic notions like *validity, soundness*, and *logical necessity*. Moreover, even when more naturalistic stimulus materials were employed, the tasks still required participants without logical instruction to set aside their background knowledge and evaluate conclusions in light of premises that they were supposed to just assume to be true. Yet, this is a type of processing that participants find unnatural as shown by well-documented context effects and belief bias effects (Klauer, Musch, & Naumer, 2000). Furthermore, in this paradigm, participants were assessed based on interpretations of natural language words like *some, if*, and *not* from first-order logic, which is something that subsequent research has shown to be particularly problematic for natural language conditionals (Evans & Over, 2004).

One type of response in the so-called New Paradigm in the psychology of reasoning has been to adopt a probabilistic task format, where participants are required to indicate their responses in terms of degrees of belief and are permitted to use their background knowledge (Elqayam & Over, 2013). This shift has been instrumental in investigating knowledge-rich inferences closer to commonsense in individual reasoning and in opening up new lines of investigation into argumentation and social reasoning (Oaksford & Chater, 2019).

The replacement of response format, however, also raises questions about how well participants' performance under the New Paradigm compares with the decades of data collected under the old Deduction Paradigm (Singmann & Klauer, 2011). Moreover, an often overlooked feature of the probabilistic representations of degrees of belief within psychology is that they also require basic logical properties like freedom from inconsistency and logical closure, which remain requirements of rational belief even within the New Paradigm (Skovgaard-Olsen, 2017a).[33] Probability theory can either be formulated in terms of set theory or in the language of propositional logic. Either way, there are certain logical properties that degrees of beliefs represented by probabilities must satisfy, like the ones listed below (Peterson, 2017, Ch. 6). Consequently, participants tested in the New Paradigm should still exhibit deductive competence to count as rational, Bayesian agents. For instance, they should

[33] For further discussion of the requirements of rational beliefs see Spohn (2012) and Raidl and Skovgaard-Olsen (2017).

still be able to assign probability 1 to logical consequences when reasoning with premises that have probability 1 (Oaksford & Chater, 2009). And, more generally, degrees of belief of rational Bayesian agents are constrained by the properties of logical truth, logical consequence, consistency, and logical equivalence as follows (Adams, 1998, p. 21-24):

> If φ is *logically true*, then their degree of belief in φ should be: $P(\varphi) = 1$,
> If φ *logically implies* ψ, then their degrees of belief in φ and ψ should conform to the inequality: $P(\varphi) \leq P(\psi)$
> If φ and ψ are *logically inconsistent*, then their degrees of belief in φ and ψ should conform to: $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$
> If φ and ψ are *logically equivalent*, then their degrees of belief in φ and ψ should conform to: $P(\varphi) = P(\psi)$

By introducing the requirement that (arbitrary complex) logical relations should be recognized in the assignment of degrees of beliefs, even when reasoning with uncertain premises, these principles illustrate how probability theory adds further requirements of rationality; not less.

P-validity

As part of the New Paradigm, a need to study inferences from uncertain premises has been identified (Stevenson & Over, 1995). One common solution has been to incorporate the work of Adams (1975, 1998) on probabilistic validity as generalizing the notion of classic validity (Cruz, Baratgin, Oaksford, & Over, 2015; Cruz, Over, Oaksford, & Baratgin, 2016; Cruz, Over, Oaksford, 2017; Evans, Thompson, & Over, 2015; Singmann, Klauer, & Over, 2014). Whereas classically valid inferences preserve truth from the premises to the conclusion, p-valid inferences cannot go from low uncertainty in the premises to high uncertainty in the conclusion. Defining the uncertainty of φ as $U(\varphi) = 1\text{-}P(\varphi)$, this idea can be explicated in terms of the uncertainty sum-rule.

> THE UNCERTAINTY SUM-RULE AND P-VALIDITY: the inference from a set of premises, Γ, to φ is probabilistically valid iff it holds for all coherent probability distributions that $U(\varphi) \leq U(\psi_1)+\ldots+ U(\psi_n)$, for $\psi_{1\ \ldots}\ \psi_n \in \Gamma$.

Or put more colloquially: the inference is p-valid if and only if the uncertainty of the conclusion is not greater than the sum of the uncertainty of the premises, for all coherent ways of assigning degrees of belief to the premises and the conclusion. In the New Paradigm, Adams' work on p-validity has been celebrated as a general solution to the problem of which inferences to accept when reasoning under uncertainty with degrees of belief that avoids the problems associated with asking participants to reason based on logical validity.

The empirical question of whether participants are then better able to reason based on p-validity is, however, not entirely clear. For instance, Evans et al. (2015) obtained mixed results when investigating the four inferences of the conditional inference task: MP (If A, C; A, therefore C), MT (If A, C; ¬C, therefore ¬A), DA (If A, C; ¬A, therefore ¬C), AC (If A, C; C, therefore A). When examining chance-corrected hit-rate levels according to p-validity, Evans et al. (2015) only found a reliable above chance performance for the valid MP and the invalid AC inference; for the valid inference MT the hit rate was below chance levels. As the authors note: "participants did not conform to p-validity on the inferences that are actually valid, MP and MT. Indeed, there was a small trend in the opposite direction" (p. 9). Similarly, Singmann et al. (2014) found that participants only conformed to p-validity for MP inferences and not for MT inferences. Moreover, when Cruz et al. (2017) stipulate the premise probability to be 100%, mean estimates for the conclusion of the valid inferences considered were around 85%-92%, in violation of the uncertainty sum-rule.

There is some discussion about whether the uncertainty sum-rule can be applied to point estimates as opposed to interval estimates representing coherence intervals of imprecise probabilities (Kleiter, 2018; see also Pfeifer & Kleiter, 2009). But here we highlight a different issue: the definition of p-validity contains a universal quantifier, which requires that the uncertainty sum-rule is conformed to *by all coherent probability distributions*. Similarly, the model-theoretic notion of classical validity contains a universal quantifier requiring that the conclusion of valid inferences is true *in all models satisfying the premises*. This universal quantifier gives classically valid inferences the modal content that they are *necessary* (i.e. that there *cannot* exist a model of a classically valid inference in which the premises are true and the conclusion is false). Similarly, the universal quantifier in the uncertainty sum-rule gives p-valid inferences the modal content that there *cannot* exist a coherent probability assignment in which the uncertainty of the conclusion is greater than the sum of the uncertainty of the premises.

In the abovementioned psychological studies advocating p-validity, it is common to investigate only a handful of premise probabilities (e.g., by stipulating that the premise probability is 60%, 80%, and 100%) and measure the probability assigned to the conclusion of valid and invalid inferences. Since, however, this type of task does not address the universal quantifier, and the modal content of p-valid inferences, it would be more accurate to say that what these studies investigate is first and foremost participants' *probabilistic* *coherence*, or whether their probability assignments are *in agreement* with the uncertainty sum-rule. In contrast, these studies do not directly investigate participants' *acceptance of entailments* in p-valid inferences—since for this, the experimental tasks would have had to be designed in a way that is suited for the modal content of p-valid inferences. To draw an analogy: from a handful of (or even many) truth-value assignments to the premises and conclusions of MP inferences, one has not shown that participants accept *the entailment* from the premises to the conclusion. For this, one would have to show that participants accept that the conclusion *cannot* fail to be true, once the premises are true.

It would appear then that there still exists a need for finding a natural way of assessing participants' acceptance of *entailments* in the New Paradigm, in spite of its many improvements to the research practice of psychologists studying human reasoning and in spite of the considerable merits of p-validity. Given the central role that entailments continue to play in the mathematical modelling of natural language through formal semantics in linguistics (see e.g., Cann, 1993; Heim & Kratzer, 1998), it would be desirable to have a substantive body of empirical data surveying the entailment judgments of ordinary people. For instance, to know which of the logical principles discussed in Arlo-Costa (2007) characterize natural language conditionals, instead of further investigations into MP, MT, AC, and DA, which are not discriminatory with respect to competing logical systems. Indeed, according to Winter (2016, Ch. 2), a central empirical adequacy criterion of semantic theories is that they respect intuitive entailment judgments. Intuitive entailment judgments thus make up one of the primary sources of data for semantic theories.

The Dialogical Entailment Task

For the reasons indicated above, the present paper seeks to present a more natural, dialogical paradigm for eliciting participants' acceptance of entailments.[34] The inspiration comes from various sources. First, from the observation that classical logic is best viewed as a competence model for adversarial reasoning when we attempt to disprove the arguments of our interlocutors (Stenning & van Lambalgen, 2008). Second, the idea is motivated by the observation that attributions of consequential commitments in argumentative contexts provide a natural setting for assessing participants' grasp of the logical consequences of their beliefs (Skovgaard-Olsen, 2017a). Finally, it is informed by linguistic work on empirical evidence for semantic theories (Tonhauser & Matthewson, 2015).

[34] This task was first put to use in Skovgaard-Olsen, Kellen, Hahn, and Klauer (2019b), when investigating and-to-if inferences.

The Dialogical Entailment Task has the following format: Samuel asserts the premise of a supposed entailment and denies its conclusion. His interlocutor, Louis, points out that Samuel has said two things that cannot both be true. The task of the participants is to assess the extent to which they agree/disagree with Louis' accusation on a Likert-scale.

In asking participants to judge whether Samuel has said two things that cannot both be true, the task builds on previous work reporting that participants find it easier to make such judgments than direct judgments concerning consistency (Johnson-Laird, Girotto & Legrenzi, 2004). Since the objection of inconsistency moreover concerns another speaker, the dialogical setting of the task is expected to make it more natural for participants to reason on the basis of the premises of the supposed entailment while setting aside their own beliefs. While it is perceived as unnatural for participants without logical training to bracket their own background beliefs, it is not unnatural for naive participants to reason on the basis of the foreign premises of another interlocutor and point out consistency problems in their line of reasoning. Finally, due to its basis on intuitive objections of inconsistency, the task does not require participants to have a sophisticated grasp of semantic notions like soundness, validity, or logical necessity (Tonhauser & Matthewson, 2015).

Earlier studies have examined which inferences participants draw in dialogical settings (Stevenson & Over, 1995; Thompson & Byrne, 2002) and investigated their degree of belief in the conclusion of informal reasoning fallacies as well as their acceptance of such arguments (Oaksford & Hahn, 2004; Hahn & Oaksford, 2007). These studies were, however, not designed to elicit participants' entailment judgments (as opposed to their acceptance of other types of inferences like, say, inductive inferences or implicatures). In fact, much of the research on argumentation within the New Paradigm has been conducted with the explicit goal of showing how everyday informal arguments that have been set aside by classical logic can nevertheless be captured by rational Bayesian reconstructions (Hahn, Harris, & Oaksford, 2012). In contrast, in Eva & Hartmann (2018) it is argued that even on a Bayesian approach to argumentation, an interest should be taken in valid arguments. The reason they give is that valid arguments have the property of ensuring that increases to the probability of one of the premises will *guarantee* that the probability of the conclusion increases. This points in the same direction as Adams' (1975) work on p-validity reviewed above but is shown to hold in a much more general framework based on minimizing the Kullback-Leibler distance between the prior and posterior probability distributions.[35] This goes to show that even within the New

[35] However, it should be noted that Eva and Hartmann’s (2018) argument is based on conjectures generalizing from examining inferences like MP, MT, AC, and DA without

Paradigm there is a need to investigate participants' acceptance of entailments in argumentative contexts.

Entailment judgments

The following principles are much discussed in conditional logics:

The Negation Principle	$\neg(\text{if } A, C) \Leftrightarrow \text{if } A, \neg C$
Conditional Excluded Middle	$(\text{if } A, C) \vee (\text{if } A, \neg C)$

As Adams (1998) says:

> The negation of a conditional, e.g., “It is not the case that if it rains it will pour,” is superficially simple to analyze, because it seems intuitively to be equivalent to the conditional denial, “If it rains it won’t pour.” In general, on this view $\sim(\varphi \Rightarrow \psi)$ seems to be equivalent to $\varphi \Rightarrow \sim\psi$. (p. 270)

Correspondingly, the Negation Principle is central to the Suppositional Theory of conditionals (Handley et al., 2006) and accepted by Stalnaker (2011, p. 233) and the three-valued logic of conditionals in Cantwell (2008a).

This principle moreover follows on general grounds connecting conditionals, subjective probability, and betting that have been influential in the New Paradigm based on work by de Finetti and Ramsey (Baratgin, Over, & Politzer 2013; Baratgin, Politzer, Over, & Takahaschi, 2018). On such accounts, the indicative conditional is explicated by the de Finetti truth table, which assigns conditionals the value ‘True’ in the $\top\top$ cell, ‘False’ in the $\top\bot$ cell, and ‘void’ in the false antecedent cells.[36] This assignment is in turn motivated by a betting analysis, according to which a conditional bet on “if A, C” is won if “A & C” turns out to be the case, lost if “A & ¬C” turns out to be the case and rendered void if “¬A” is the case. Since bets on [¬(if A, C)] and [if A, ¬C] have the same pattern of wins and losses, the Negation Principle follows for probabilistic accounts of conditionals that are based on these principles.

presenting a proof for the general case. It is also unclear how far their conclusions generalize to other frameworks. For instance, Kleiter (2018) finds that while MT is p-valid it is not *n-increasing*, in the sense that if the probability of any of the *n* premises increases, the probability of the conclusion also increases.

[36] The Jeffrey table is a variant of this, which assigns the value ‘P(C|A)’ in the false antecedent cells.

Concerning the Principle of Conditional Excluded Middle, there is a famous dispute between Lewis (1973) and Stalnaker (1980) about whether to accept it for subjunctive conditionals (e.g., 'If Oswald hadn't killed Kennedy, someone else would have'). Yet, Bacon (2015, 2019) argues that the status of the Principle of Conditional Excluded Middle is much less controversial for indicative conditionals (e.g., 'If Oswald didn't kill Kennedy, someone else did') than for subjunctive conditionals. Both the Negation Principle and the Principle of Conditional Excluded Middle require the following inference to be valid (where '⊨' indicates semantic consequence):

Target Inference: $\neg$(if A, C) ⊨ if A, $\neg$C

However, if, in contrast, participants think that [$\neg$(if A, C)] can be true because neither [if A, C] nor [if A, $\neg$C] are true, when there is no dependency between A and C, then the Target Inference should not be accepted. Accordingly, inferentialist accounts of conditionals that make inferential relations between A and C part of the truth conditions of conditionals, like Douven (2015), should reject the validity of the Target Inference.

In the experiments that follow, we will therefore investigate whether participants accept the validity of the Target Inference. To do this, the following two baselines are employed as well:

Agree Baseline: if A, $\neg$C ⊨ $\neg$(if A, C)
Disagree Baseline: if A, C ⊨ if A, $\neg$C

The idea behind the use of these baselines is to have two inferences which most theories will treat as valid,[37] and invalid respectively, as a manipulation check for the Dialogical Entailment Task (described in further details below). The test then consists in assessing whether participants' performance concerning Target Inference is more like their performance with respect to the Agree or the Disagree Baseline.

[37] On Stalnaker's logic, only the following restricted version of the Negation Principle holds: possibly(A) ⊨ $\neg$(if A, C) ⇔ if A, $\neg$C. In contrast, Stalnaker and Lewis' possible worlds semantics cannot treat the Accept Baseline as valid due to their stipulation that all so-called counterpossibles (i.e., conditionals with an impossible antecedent) are true irrespectively of the consequent. That is to say, whenever there is no accessible A-world, both [if A, $\neg$C] and [if A, C] are treated as true, and thus the Agree Baseline fails to be valid. However, this aspect of their treatment of counterpossibles is often criticized (see e.g., Mares, 2007).

In investigating these inferences, relevance manipulations are applied, which are motivated below.

The Relevance Effect

In a famous footnote, Ramsey (1929/1990) suggested that two interlocutors could settle their argument over a conditional ‘if A, then C’ by hypothetically adding the antecedent, A, to their stock of beliefs and arguing over the consequent, C, on that basis. As explained in Arlo-Costa (2007), and Skovgaard-Olsen (2017b), this little footnote outlining the so-called “Ramsey test” has inspired at least three opposing research programs in logic. We will here focus on the two which have been most influential for linguistics and psychology.

On the one hand, there is the Lewis (1973) and Stalnaker (1968) possible-worlds semantics of conditionals, which is popular in linguistics (Kratzer, 1986, 2012), that supplies an account of the truth conditions of subjunctive conditionals, according to which a subjunctive (i.e. ‘if A had been the case, then C would have occurred’) is true iff the consequent is true in all the closest possible world(-s) in which the antecedent is true. That is to say, in order for the conditional to be true, the consequent must be true in possible worlds where the antecedent is true that are otherwise minimally different from the actual world. In Stalnaker (1968), this is made precise by introducing a selection function, f(A, *w*), which selects the closest world (or, alternatively: the set of closest worlds) to *w* in which A is true. The conditional, [A > C], is then true iff the selected A-world(s) is a subset of the set of worlds in which C is true, [C] (Égré & Cozic, 2016). While Lewis (1973) only applies this analysis to subjunctive conditionals, Stalnaker (1968) takes it to hold for indicative conditionals as well.

On the other hand, the Ramsey test has inspired the probabilistic semantics of indicative conditionals of Adams (1975), which in its original form denies that indicative conditionals have truth conditions, and subscribes to either P(if A, C) = P(C|A) or acc(if A, C) = acc(C|A), for ‘if A, C’ referring to simple conditionals (which exclude nestings of conditionals). Here ‘acc(if A, C)’ stands for the acceptability of the conditional. Often this version of Adams’ thesis is preferred, because it is unclear whether P(if A, C) can still be interpreted as a probability in light of the so-called triviality results, which supply a reduction of the most obvious way of implementing this thesis (Bradley, 2007; Douven, 2015). Through the influence of the writings of Edgington (1995) and Bennett (2003), the psychological hypothesis that the probability of indicative conditionals is evaluated as the conditional probability, P(C|A), found its way into the psychological literature (Evans & Over, 2004), where it goes by the name “the Equation”.

Results by Skovgaard-Olsen, Singmann, and Klauer (2016a) recently raised an explanatory challenge for proponents of the Equation, and theories of conditionals that postulate that indicative conditionals have a core meaning which exclude relevance relations between the antecedent and the consequent. In particular, Skovgaard-Olsen et al. (2016a) found that relevance strongly moderated the evaluations of indicative conditionals, when investigating their probability and acceptability. For cases of Positive Relevance ($P(C|A)$ - $P(C|\bar{A}) > 0 \Leftrightarrow \Delta P > 0$), like “If Pete is setting his alarm clock, then Pete will get up in time for the meeting”, the conditional probability remained a good predictor of both the acceptance and probability of conditionals. For cases of Negative Relevance ($P(C|A) - P(C|\bar{A}) < 0 \Leftrightarrow \Delta P < 0$), such as “If Pete is setting his alarm clock, then Pete will be late for the meeting”, and Irrelevance ($P(C|A) - P(C|\bar{A}) = 0 \Leftrightarrow \Delta P = 0$), like “If Pete is wearing green socks, then Pete will be late for the meeting”, this relationship was disrupted. What this indicates is that participants tend to view the indicative conditional as defective under conditions, where the antecedent cannot be interpreted as providing a reason *for* the consequent, because the antecedent fails to raise its probability.

It is sometimes suggested that the Relevance Effect should be interpreted in terms of causal readings of conditionals (e.g., van Rooij & Schulz, 2018; Oaksford & Chater, 2019), given that ΔP makes up the numerator in causal power (Cheng, 1997). But it is also possible to consider causal relations as a specific instance of a more generic reason relation (Spohn, 2012), which then turns the Relevance Effect into a finding concerning the relationship between conditionals, reasons, and arguments. Possible explanations for the Relevance Effect are diverse and have been explored in several recent publications (Cruz, Over, Oaksford & Baratgin, 2016; Krzyżanowska, Collins, & Hahn, 2017; Skovgaard-Olsen, Collins, Krzyżanowska, Hahn, & Klauer, 2019a). In this paper, the goal is to investigate whether relevance effects extend to participants’ reasoning with conditionals containing negation operators, in their probability assignments and entailment judgments. Experiment 1 starts out by applying the Dialogical Entailment Task to the three types of inferences introduced above.

3.2 Experiment 1: Entailment Judgments

3.2.1 Methods

Participants

The experiment was conducted using the internet platform Mechanical Turk. Participants received a small amount of money in exchange for their participation. 116 took part in the experiment. The following exclusion criteria were used: not having English as the native language, failing to answer two SAT comprehension questions correctly in a warm-up phase, completing the task in less than 240 s or in more than 3600 s, and answering ‘not seriously at all’ to the question of how seriously they would take their participation. The final sample consisted of a total of 48 people. Mean age was 38.7 years, ranging from 21 to 68 years, 58% of the participants were female, and 68.8% of the participants had an undergraduate degree or higher. The demographics of the participants were similar before and after exclusion.

Design

The Experiment implemented a within-subjects design. Three factors were individually varied: Relevance (Positive Relevance vs. Irrelevance), Priors (HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH) and Inference Type. The Inference type factor had three levels: Agree Baseline, Disagree Baseline, and Target Inference (repeated below). Each participant thus completed 24 within-subject conditions in total.

Materials and Procedure

To reduce the dropout rate once the proper experiment had begun, participants were first shown our academic affiliations. The participants were then presented with two SAT comprehension questions in a warm-up phase and a seriousness check to ensure that the participants carefully completed their responses (Reips, 2002).

The participants were given the following task instructions:

> In the following you are going to see a short conversation, where Louis accuses Samuel of saying two things that cannot both be true. Whether you agree with Samuel's assertions is beside the point. What we are interested in is just the extent to which you agree with Louis that Samuel is saying two things that cannot both be true. When you read the sentences please pay attention to small differences in their content, so that we don't unfairly accuse Samuel of making a mistake.

Each participant completed judgments for the eight experimental conditions relating relevance and priors (Positive Relevance: HH, HL, LH, LL; Irrelevance HH, HL, LH, LL) in blocks featuring the three inference types. The order of the blocks was randomized anew for all participants. Each of these eight blocks was randomly assigned to one of 12 possible scenarios using random assignment without replacement such that each participant saw a different scenario for each condition. All items within a block were presented with the same scenario and were presented in random order.

The 12 scenarios used in this study were taken from Skovgaard-Olsen et al. (2016b). These scenarios were found to reliably induce assumptions about relevance and prior probabilities of the antecedent and the consequent in previous studies that implement our experimental conditions. Table 1 displays sample items for the Mark scenario for Positive Relevance ($\Delta p > 0$), and Irrelevance ($\Delta p = 0$), for $\Delta p = P(C \mid A) - P(C \mid \overline{A})$.

Table 1. Stimulus Materials, Mark Scenario

Scenario	Mark has just arrived home from work and there will shortly be a great movie on television, which he has been looking forward to. Mark is quite excited because he recently bought a new TV with a large screen. He has a longing for popcorn, but his wife has probably eaten the last they had while he was gone.	
	Positive Relevance	**Irrelevance**
HH	If Mark presses the on switch on his TV, then his TV will be turned on.	If Mark is wearing socks, then his TV will work.
HL	If Mark looks for popcorn, then he will be having popcorn.	If Mark is wearing socks, then his TV will malfunction.
LH	If the sales clerk in the local supermarket presses the on switch on Mark's TV, then his TV will be turned on.	If Mark is wearing a dress, then his TV will work.
LL	If Mark pulls the plug on his TV, then his TV will be turned off.	If Mark is wearing a dress, then his TV will malfunction.
	Positive relevance (PO): mean ΔP = .32 Irrelevance (IR) mean ΔP = -.01	High antecedent: mean P(A) = .70 Low antecedent: mean P(A) = .15 High consequent: mean P(C) = .77 Low consequent: mean P(C) = .27

Note. HL: P(A) = High, P(C) = low; LH: P(A) = low, P(C) = high. The bottom rows display the mean values for all 12 scenarios pretested in Skovgaard-Olsen et al. (2016b).

For the Mark scenario text in Table 1, participants assume that "Mark is pressing the on switch on his TV" raises the probability of that "his TV will be turned on", and that both of these sentences have a high prior probability (Positive Relevance, HH). Conversely, participants assume that "Mark is wearing socks" is irrelevant for whether "his TV will work", and that both have a high prior (Irrelevance, HH). The full list of scenarios can be found in the supplemental materials: https://osf.io/npc69/.

On the first page of each block, the scenario was displayed. For future reference, the scenario was repeated on the top of each page that followed in grey colour. The next three pages presented the three inference types in random order.

The participants saw two control items and a practice item before the actual experiment started, where it was emphasized that attention was needed to notice subtle differences between the wordings (e.g., use of 'not', 'false', 'wrong', 'correct', and 'if') of the various sentences presented in the experiment. For the control items, Samuel would either assert “Some of the employees are invited to the party” and deny that “not all of the employees were invited” (i.e., consistently deny a scalar implicature), or assert that “John is a bachelor” and deny that “John is unmarried” (i.e. inconsistently denying an analytical consequence of his first assertion).

In each case, Louis made the following objection to Samuel:

> Louis: Wait, you've now said two things that can't both be true.

The task of the participants was to indicate the extent to which they agreed/disagreed with Louis' statement above on a five-point Likert scale {strongly disagree, disagree, neutral, agree, strongly agree}. Agreeing with Louis' objection counts as accepting the entailment for a given inference. All other responses merely indicate lack of acceptance of the entailment.

The experimental task had the same format. This time Samuel would assert the premise and deny the conclusion of the three following inferences:

Agree Baseline:	if A, $\neg$C $\vDash$ $\neg$(if A, C)
Disagree Baseline:	if A, C $\vDash$ if A, $\neg$C
Target Inference:	$\neg$(if A, C) $\vDash$ if A, $\neg$C

In Table 2, Samuel's assertions with respect to these inferences are illustrated using the stimulus materials from Table 1 (however, without 'then' and 'will' in the consequents):[38]

Table 2. The Dialogical Entailment Task

Scenario Mark has just arrived home from work and there will shortly be a great movie on television, which he has been looking forward to. Mark is quite excited because he recently bought a new TV with a large screen. He has a longing for popcorn, but his wife has probably eaten the last they had while he was gone.		
Agree Baseline	**Reject Baseline**	**Target Inference**
	Positive Relevance	
Samuel: IF Mark presses the on switch on his TV, his TV does NOT turn on.	**Samuel:** IF Mark presses the on switch on his TV, his TV turns on.	**Samuel:** It is FALSE that IF Mark presses the on switch on his TV, his TV turns on.

[38] For all the experiments in this paper, ‘then’ in the consequents was removed from the contents. This is to see whether reason relation readings of conditionals are induced by ‘then’ in the consequents (as suggested by Iatridou, 1994; von Fintel, 1994; Biezma, 2014). Additionally, ‘will’ was removed. The future tense was replaced with present tense. See Experiment 2 for further details on these modifications.

...but it would be CORRECT to think that IF Mark presses the on switch on his TV, his TV turns on. **Louis:** Wait, you've now said two things that can't both be true.	...but it would be WRONG to think that IF Mark presses the on switch on his TV, his TV does NOT turn on. **Louis:** Wait, you've now said two things that can't both be true.	...but it would be WRONG to think that IF Mark presses the on switch on his TV, his TV does NOT turn on. **Louis**: Wait, you've now said two things that can't both be true.
Irrelevance		
Samuel: IF Mark is wearing socks, his TV does NOT work. ...but it would be CORRECT to think that IF Mark is wearing socks, his TV works. **Louis**: Wait, you've now said two things that can't both be true.	**Samuel:** IF Mark is wearing socks, his TV works. ...but it would be WRONG to think that IF Mark is wearing socks, his TV does NOT work **Louis**: Wait, you've now said two things that can't both be true.	**Samuel:** It is FALSE that IF Mark is wearing socks, his TV works. ...but it would be WRONG to think that IF Mark is wearing socks, his TV does NOT work **Louis**: Wait, you've now said two things that can't both be true.

Note. Samuel denies the conclusion of the inferences by saying 'it would be correct/wrong to think that...'. For the Agree Baseline, Samuel is denying a wide scope negated conditional [¬ (if A, C)]. To avoid using double negations, which are notoriously difficult to process, a formulation was chosen where Samuel denies the conclusion by saying that '...but it would be CORRECT to think that IF...' as opposed to '...but it would be WRONG to think that it is NOT the case that IF...'.

Finally, Experiment 1 contained an open-ended question where participants were asked to explain why they had agreed/disagreed with Louis' objection for each of the Target Inferences so that the foreign language learner Eva would be able to comprehend the task they just completed. These open-ended responses were, however, used in an exploratory fashion and are not reported for the statistical analysis below. But they can be accessed through the data set in the Online Supplementary Materials.

3.2.2 Results

Control Items

The degree to which participants agreed with accusing Samuel of an inconsistency was found to be significantly higher in the entailment control item (*Mdn* = 4.00) than in the scalar implicature control item (*Mdn* = 2.00), V = 86, *p* < .01, *r* = -.29, for the Wilcoxon signed-rank test. The experimental task was thereby found to pass a first manipulation check.

Entailment Judgments

To examine ratings of entailment for the three types of inferences, we relied on a set of mixed generalized linear models, which represent the acceptance of an entailment (a binary variable formed by answering "Agree" or "Strongly agree" to Louis' objection to Samuel) by a binominal likelihood function together with a logit link function. The models had crossed random effects for intercepts and slopes by participants and by items (Baayen, Davidson, &

Bates, 2008) to control for the effect of replicates for each participant and item in the experimental design. The models were fitted in a Bayesian framework using the R-package `brms` (Bürkner, 2017) with weakly informative priors and featured the following predictors:

> Model **M1** modelled acceptance of entailment as a function of the Inference factor (Agree vs. Disagree vs. Target), the Relevance factor (Positive Relevance vs. Irrelevance), and their interaction.
>
> Model **M2** built upon M1 but did not include the two-way interaction.
>
> Model **M3** built on M2 but did not include the Relevance factor.

Table 3 reports the performance of these models as quantified by Watanabe-Akaike information criterion (WAIC) and the leave-one-out cross validation information criterion (LOOIC).

Table 3. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1	1273.14	4.87	2.62	1266.0	0.058
M2	1269.99	1.73	0.90	1263.5	0.281
M3	1268.27	0	--	1262.0	0.661

Note. Weight = Akaike weight of LOOIC. Lower numbers of LOOIC and WAIC indicate better predictive performance in light of the trade-off between model fit and parsimony.

The information criteria displayed in Table 3 indicate that M3 was the winning model. Hypotheses concerning the presence/absence of effects are tested here and below by setting coefficients of the full model (M1) equal to zero. In this way, evidence in favour of, e.g., the H_0 that there is no main effect of Relevance can be quantified in terms of Bayes factors.

The fact that M3 was the winning model suggests that the participants' entailment judgments neither displayed a main effect of Relevance (b = 0.36, 95%-CI [-0.26, 1.01], BF_{H0H1} = 5.13) nor an interaction between Relevance and the Inference factor ($b_{Disagree:Irrelevance}$ = -0.51, 95%-CI [-1.34, 0.32], BF_{H0H1} = 3.55; $b_{Target:Irrelevance}$ = -0.26, 95%-CI [-1.17, 0.63], BF_{H0H1} = 5.79). In contrast, strong evidence was obtained for the hypothesis that the posterior probabilities of accepting the entailment in both the Disagree Baseline (b = -2.22, 95%-CI [-3.19, -1.31], BF_{H0H1} = 1.88 * 10^{-9}), and for the Target Inference (b = -1.31, 95%-CI [-2.12, -0.52], BF_{H0H1} = 0.045), were substantially below the posterior probability of accepting the entailment in the Agree Baseline. Figure 1 displays the posterior probabilities of acceptance of entailment for each type of inference.

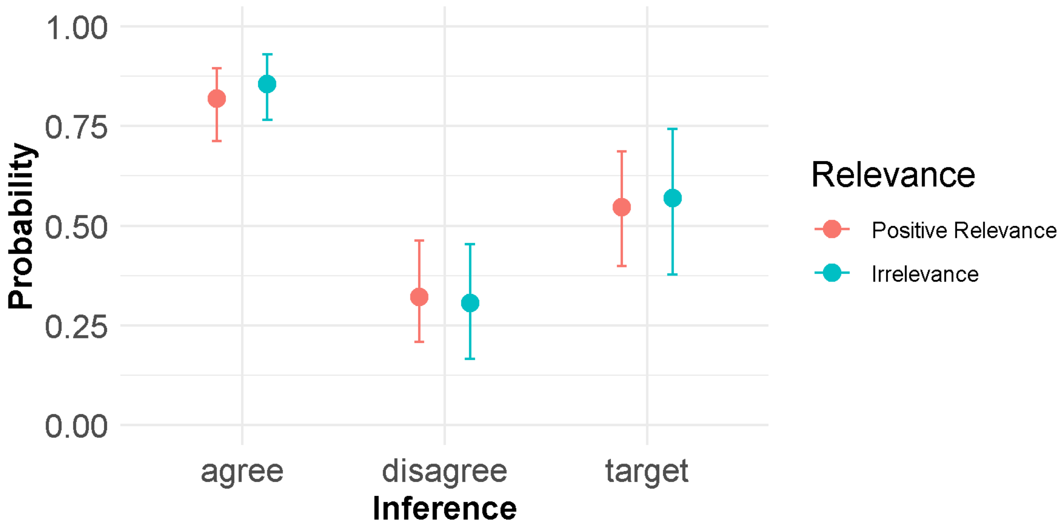


Figure 1. Weighted posterior predictive probability of acceptance of entailment. ‘agree’ = baseline for agreement; ‘disagree’ = baseline for disagreement; ‘target’ = inference to be compared with the baselines. 'Probability' on the y-axis indicates posterior probability of accepting the entailment for a given inference. The posterior predictions of M1, M2, M3 have been weighted by their Akaike weight from Table 3 to produce this plot.

3.2.3 Discussion

As a manipulation check of the Dialogical Entailment Task, participants’ performance with respect to two control items and two baselines were investigated. As expected, it was found that the participants accepted the entailment for the Agree Baseline and the Entailment Control Item and did not accept the entailment for the Disagree Baseline and the Scalar Implicature control item. Having established this, we turned to the comparisons between the Target Inference and the two baselines.

The results of Experiment 1 show strong evidence that participants have a lower posterior probability of accepting the entailment for the Disagree Baseline and the Target Inference than for the Agree Baseline. At the same time, the results indicate that participants lack a strong preference with respect to the Target Inferences in either direction, with posterior probabilities of acceptance of just above 50% at the group level. Since a main effect of relevance and an interaction with the Relevance factor were not found, this lack of preference concerning the Target Inference has to be accounted for on other grounds.

To further investigate participants’ performance with the Target Inference, Experiment 2 investigates the extent to which participants’ performance in the Dialogical Entailment task is consistent with their probability assignments to conditionals with negation operators, across relevance levels.

3.3 Experiment 2: Cross-Task Consistency

Experiment 2 was split into two sessions separated by one week, which are reported consecutively in this paper. The first session suffices to test the hypothesis that recent work in linguistics on the contribution of ‘then’ in conditionals can adequately account for the Relevance Effect (more on this below). The second session was introduced to compare participants’ responses across sessions with the following cross-task consistency constraint that ensures that probabilistic reasoning is consistent with deductive logic (Joyce, 2004; Oaksford, 2014):

$$A \vDash B \qquad \text{only if} \qquad P(B) \geq P(A)$$

Accordingly, the second session featured a replication of Experiment 1 ca. 1 week later after the participants had assigned probabilities to conditionals with and without negation operators, across relevance conditions.

3.3.1 Session 1: Negations, Then, and Probabilities

On the Meaning Contribution of ‘Then’

In Iatridou (1994), the dependency of the consequent on the antecedent is attributed to the contribution of ‘then’. More specifically, Iatridou suggests that utterances of ‘if A, then C’ are equivalent to utterances of ‘if A, C’ with the presupposition added that not all not-A worlds are C worlds. On this view, the conditional “If it's sunny, then Michael takes the dog to Pastorius Park” carries the assertion that "In every case in which it is sunny, Michael takes the dog to the Pasterius Park". In contrast the semantic contribution of *then* is to add the presupposition that "Not in every case in which it isn't sunny does Michael take the dog to Pastorius Park". According to Iatridou (1994), the presence of this presupposition in turn accounts for why the following special conditional constructions do not allow for the presence of ‘then’:

If John is dead or alive, (#then) Bill will find him.
Even if John is drunk, (#then) Bill will vote for him.
If I were the richest linguist on earth, (#then) I (still) wouldn’t be able to afford this house.

Similarly, it has been suggested in von Fintel (1994) that ‘then’ carries a separate meaning as a conventional implicature, and the syntactic motivation for these proposals is thoroughly discussed in Bhatt and Pancheva (2006).

In line with this, Biezma (2014) puts forward a general theory on the non-truth functional meaning of ‘then’. The central claim is that ‘then’ operates at the level of discourse structures by establishing an anaphoric relation between two discourse moves. As part of its felicity conditions, it is claimed that non-temporal uses of ‘then’ require that two propositions enter into a causal explanatory relationship, whereby the antecedent proposition provides a reason for the consequent proposition. In paraphrase, when ‘then C’ occurs alone, the meaning conveyed is ‘C because of A’, where A may remain an implicit part of the antecedent discourse.

One of the central advantages of the theories reviewed above is that apparently the Relevance Effect of conditionals reported in Skovgaard-Olsen et al. (2016a) can be explained by pointing to the occurrence of ‘then’ in the investigated stimulus materials (‘if A, *then* C’).[39] This in turn would allow us to adopt the Lewis (1973), Stalnaker (1968), and Kratzer (1986) framework to provide a semantics for ‘If A, C’ while predicting the influence of reason relations on the evaluation of the felicity conditions of ‘if A, *then* C’, which in turn should affect probability and acceptability evaluations. On this view, ‘if A, C’ merely provides a description of the worlds in the context set (to wit, that in the most similar A-worlds to the actual world, C is also true), whereas ‘if A, *then* C’ establishes a causal, explanatory claim whereby the antecedent provides causal information about the consequent.

Usually in psychology and philosophy, indicative conditionals are treated as a unit consisting of an antecedent and a consequent joined by ‘if…, then…’ (Johnson-Laird & Byrne, 2002; Johnson-Laird, Khemlani, & Goodwin, 2015; Stalnaker, 1980). However, if Iatridou (1994) von Fintel (1994), and Biezma (2014) are right, this tradition is mistaken in holding that ‘if…then’ is a primitive unit of meaning. In this they are in agreement with Grice (1989, pp. 63), who insisted that his preferred semantics of the natural language conditional applies to ‘if A, C’, and that it is obvious that it would fail for ‘if A, *then* C’.

One central purpose of Session 1 of Experiment 2 is to test this conjecture.

The Negation Task

As a test of whether Iatridou (1994), von Fintel (1994), and Biezma’s (2014) theories are able to account for the Relevance Effect, the Negation Task from Skovgaard-Olsen et al.

[39] I thank María Biezma, Maribel Romero, and Eva Csipak for discussion.

(2019a) was selected. In this task, participants are asked to assign probabilities to the following conditionals across manipulations of the antecedent's relevance for the consequent (see below):

AFFIRMATIVE CONDITIONAL: if A, C
WIDE-SCOPE NEGATION: ¬(if A, C)
NARROW-SCOPE NEGATION: if A, ¬C

where the negation operator takes a *wide scope* over the whole conditional in the first case, and a *narrow scope* over only the consequent of the conditional in the second case.

However, while a previous version of the task featured conditionals with 'then' and 'will' in the consequents, a central goal of the present study was to investigate whether we can replicate previous findings with conditionals without 'then' and 'will'.

One of the central findings produced by the Negation Task is that the following probabilistic version of the Negation Principle can only be maintained for Positive Relevance, when the antecedent raises the probability of the consequent ($\Delta P > 0$), because for Irrelevance, where the antecedent leaves the probability of the consequent unaffected ($\Delta P = 0$), the Negation Principle is systematically violated (Skovgaard-Olsen et al., 2019a):

THE NEGATION PRINCIPLE: ¬(if A, C) ⇔ if A, ¬C
Probabilistic version: P(¬(if A, C)) = P(if A, ¬C)

Yet, in Handley et al. (2006), the probabilistic version of the Negation Principle has been taken to be a litmus test for the Suppositional Theory of conditionals, which explicates the meaning of indicative conditionals in terms of the Ramsey test and the Equation, (P(if A, then C) = P(C|A)), as outlined above.

3.3.1.1 Methods

Participants

The experiment was conducted using the internet platform Mechanical Turk. Participants received a small amount of money in exchange for their participation. 141 took part in Session 1 of the experiment. The same exclusion criteria were used as in Experiment 1. The final sample for Session 1 consisted of a total of 78 people. Mean age was 38.4 years, ranging from 20 to 72 years, 61.5% of the participants were female, and 70.1% of the participants had an undergraduate degree or higher. The demographics of the participants differed minimally before and after exclusion.

Design

Session 1 implemented a within-subjects design. Three factors were individually varied: Relevance (Positive relevance vs. Irrelevance), Priors (HH, HL, LH, LL) and Sentence Type. The Sentence Type variable had five levels: two of these measured conditional probability judgments ($P(C|A)$, $P(\overline{C}|A)$), the remaining measured probability assignments to affirmative conditionals [P(if A, C)], their wide scope negation [P(¬(if A, C)], and their narrow scope negation [P(if A, ¬C)]. Each participant thus completed 40 within-subject conditions in total.

Materials and Procedure

First, participants were given a brief general introduction:

> In the course of the experiment we ask you to provide probabilities for various sentences. To fill in your responses please use the slider, which you can click on. Entering a number in the box will not work.

They were then presented with four practice items in random order. As practice items, participants were asked to assign a probability on a scale from 0 to 100% to a categorical sentence with an existential presupposition failure (e.g., “The queen of the USA is in her mid-thirties”, which falsely presupposes that there is a queen of the USA) and its wide and narrow scope negations. After this, participants were instructed to pay attention to subtle differences in the wording of the sentences used for the rest of the experiment, such as whether they contain words like 'not', 'false', and 'if'.

Each participant completed probability assignments for the eight experimental conditions relating Relevance and Priors (Positive Relevance: HH, HL, LH, LL; Irrelevance HH, HL, LH, LL) with the same counterbalancing and randomization procedure as in Experiment 1. On the first page of each block, the scenario was displayed. For each of the following five pages presenting the five sentence types in random order, the scenario was repeated on the top of the page for reference in grey colour.

The items have been modified for the purpose of this study, however. Most importantly, ‘then’ in the consequent was removed from all contents. This is to see whether the traces of the reason relation reading are induced by ‘then’, as Iatridou (1994), von Fintel (1994), and Biezma (2014) conjecture. Additionally, ‘will’ has been removed. The future tense was replaced with present tense. The wording of the wide scope negation has been modified as well, compared to the Negation Task in Skovgaard-Olsen et al. (2019a). ‘It is not the case that’ was replaced with ‘it is false that’.

3.3.1.2 Results

Probability Judgments

Like in Experiment 1, a set of mixed generalized regression models were fit to the data. The models had crossed random effects for intercepts and slopes by participants and by scenarios (Baayen, Davidson, and Bates, 2008) to control for the effect of replicates for each participant and item in the experimental design. The models featured the following predictors:

> Model M4 modelled the ratings as a function of the DV factor, encoding the three different types of conditionals (Affirm [if A, C], Wide [¬(if A, C)], Narrow [if A, ¬C)], and the Relevance factor, encoding the two relevance levels. The model also included the interaction of these two factors.
> Model M5 built upon M4 but did not include the two-way interaction.
> Model M6 built on M5 but did not include the Relevance factor.

In line with Experiment 1, these models were implemented in a Bayesian framework with weakly informative priors, using R package brms (Bürkner, 2017). Since the dependent variable consisted of continuous proportions containing zeros and ones, the values were first transformed to be within the [0,1] interval and a beta-likelihood function was used. Table 4 reports the performance of these models as quantified by WAIC and LOOIC.

Table 4. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M4	-4565.82	0	--	-4531.9	0.989
M5	-4555.03	10.79	7.10	-4519.9	0.005
M6	-4555.70	10.12	7.69	-4519.4	0.006

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC. Note that information criteria can take both positive and negative values and that the lowest value on the real line still indicates best fit.

The information criteria in Table 4 display a clear preference for M4. Consistent with this, very strong evidence for a main effect of Relevance (b_{IR} = -1.17, 95%-CI [-1.41, -0.94], BF_{H0H1} = -6.05 * 10^{-59}), the DV factor (b_{Wide} = -1.16, 95%-CI [-1.39, -0.92], BF_{H0H1} = -8.7 10^{-154}; b_{Narrow} = -1.10, 95%-CI [-1.34, -0.87], BF_{H0H1} = 2.97 * 10^{-22}), and the two-way interaction ($b_{IR:Wide}$ = 1.77, 95%-CI [1.41, 2.14], BF_{H0H1} = 2.51 * 10^{-16}; $b_{IR:Narrow}$ = 0.96, 95%-CI [0.65, 1.27], BF_{H0H1} = 4.04 * 10^{-15}) were found. The interaction is illustrated in Figure 2 with the characteristic cross-over of the lines representing Positive Relevance and Irrelevance, which makes the wide-scope negated conditionals the highest rated for the Irrelevance condition.

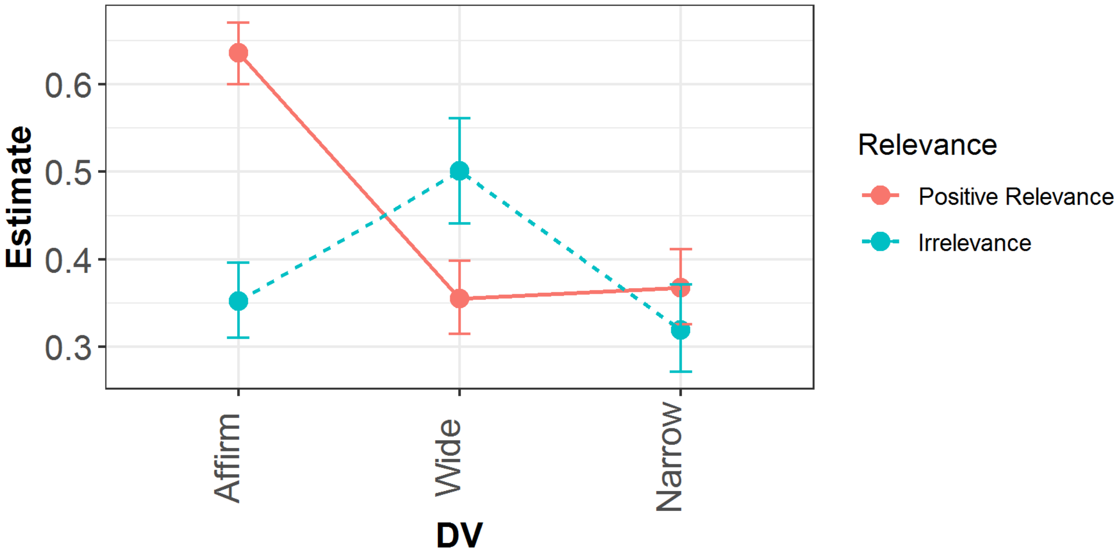


Figure 2. Posterior mean estimates of M4.

In Appendix 1A, further analyses are reported with a comparative data set from Skovgaard-Olsen et al. (2019a, Experiment 2), which differed from the present only in involving conditionals featuring ‘then’ and ‘will’ in the consequent. As the results show, very strong evidence could be obtained for the H_0 stating that there is no difference between the two datasets for all main effects and interactions in which the ‘Experiment’ factor figured (representing the identity of the two datasets). One central advantage of the present Bayesian framework is that evidence in favour of H_0 can be quantified in terms of Bayes factors, whereas classical statistics only permits inferences about whether H_0 could or could not be rejected at the $\alpha = 0.05$ level (Wagenmakers et al. 2018). In the present context, where replications of the effects in Skovgaard-Olsen et al. (2019a, Experiment 2) without ‘then’ and ‘will’ are tested, this makes Bayesian statistics ideally suited.

3.3.1.3 Discussion

Replicating Skovgaard-Olsen et al. (2019a, Experiment 2), it was found that there is a strong interaction between negation operators and relevance conditions making wide scope negated conditionals the highest rated conditionals in the Irrelevance condition (see Figure 2). The analysis reported in Appendix 1A provide further support for the H_0 that there were no differences between the present data set and the dataset reported in Skovgaard-Olsen et al.

(2019a, Experiment 2). Participants thus appear to treat the difference between ‘if A, C’ and ‘if A, *then* C’ to be little more than a stylistic difference when assigning probabilities to [if A, C], [if A, ¬C], and [¬(if, A, C)] across relevance levels. This in turn agrees with the notion in Geis and Lycan (1993) that genuine conditionals can take the proform ‘then’ in their consequents without change in meaning, in contrast to superficially similar constructions that are not conditional in meaning, like so-called biscuit conditionals:[40]

> If you want any, there are biscuits on the sideboard
> #If you want any, *then* there are biscuits on the sideboard.

The replication of Skovgaard-Olsen et al. (2019a, Experiment 2) strongly suggests that it is not the presence of ‘then’ that is driving the Relevance Effect. For instance, in both experiments, the marked drop of the marginal means of [if A, C] from ca. 65% in the Positive Relevance condition to ca. 35% in the Irrelevance condition was found, which was originally reported in Skovgaard-Olsen et al. (2016a).

This tells against attempts to use accounts of the meaning contribution of ‘then’ along the lines of Iatridou (1994), von Fintel (1994), and Biezma (2014) as an explanation for the Relevance Effect. We can thus conclude that it is something about indicative conditionals, and not about the presence of ‘then’ in the consequents, which gives rise to the expectation that the antecedent is a reason for the consequent. In Skovgaard-Olsen et al. (2019a), several linguistic categories at the interface between pragmatics and semantics were investigated and accumulating evidence was presented that the Relevance Effect is produced by a conventional implicature. Based on the present results, we can conclude that this conventional implicature does not arise due to the presence of ‘then’ or ‘will’ in the examined stimulus materials.

In both Skovgaard-Olsen et al. (2019a, Experiment 2) and the present experiment, it is found that the probabilistic version of the Negation Principle can only be maintained for the Positive Relevance condition. In contrast, this principle is systematically violated for the Irrelevance condition in both experiments. While the affirmative conditional [if A, then C] was rated the highest, and [¬(if A, then C)] was rated the lowest, in the Positive Relevance condition, this relationship switched in the Irrelevance condition with the affirmative conditional being rated the lowest and [¬(if A, then C)] being rated the highest. This is in spite of the fact that the probabilistic version of the Negation Principle has been taken as a litmus test for the Suppositional Theory of conditionals in Handley et al. (2006).

[40] See Iatridou (1994), Biezma (2014), Bhatt and Pancheva (2006), and Zakkou (2017) for further discussion.

A further way of interpreting our results is that the relevance manipulation invites two different resolutions of the ambiguity of the scope of the negation operator. To illustrate, Bhatt and Pancheva (2006) point out that the following sentence is ambiguous between two readings: "Mary doesn't yell at Bill if she is hungry". The two readings become salient with the following continuations:

...but if she is sleepy.
...since hunger keeps her quit.

In the first continuation, "if she is hungry, Mary yells at Bill" is rejected and the conditional "if she is sleepy, Mary yells at Bill" is accepted. In the second continuation, the conditional "if she is hungry, Mary yells at Bill " is rejected and the conditional "If she is hungry, Mary does not yell at Bill" is accepted.

One way of interpreting the interaction between the Relevance factor and the negation operator for probability assignments, which was raised by one of the reviewers, is that Positive Relevance and Irrelevance brings out this ambiguity in the scope of the negation operator and that Irrelevance forces the wide-scope interpretation (in which both 'if A, then C' and 'if A, then ¬C' are rejected) whereas Positive Relevance typically goes along with the narrow-scope interpretation (according to which 'if A, then not-C' and '¬(if A, then C)' are equivalent). Further research will have to determine the merits of this interpretation. So far, possible scope ambiguities like this are an underexplored topic in the psychology of reasoning. However, their importance has recently been stressed by Over, Douven, and Verbrugge (2013).

3.3.2 Session 2: Negations and Entailments

The Dialogical Entailment Task

A week later, the same participants from Session 1 were invited to participate in the Dialogical Entailment task from Experiment 1.

Investigating participants' entailment judgments with respect to the inferences from Experiment 1 allows us to apply the following cross-task consistency constraint that ensures that their probability judgments in session 1 are consistent with their entailment judgments in session 2:

$$A \vDash B \qquad \text{only if} \qquad P(B) \geq P(A)$$

Hence, it holds for the inferences under investigation that they are licensed by conformity to the inequality constraints outlined in Table 5:

Table 5. Applying the Cross-Task Consistency Constraint

	Inference		License
Agree Baseline	if A, ¬C ⊨ ¬(if A, C)	only if	P(¬(if A, C)) ≥ P(if A, ¬C)
Disagree Baseline	if A, C ⊨ if A, ¬C	only if	P(if A, ¬C) ≥ P(if A, C)
Target	¬(if A, C) ⊨ if A, ¬C	only if	P(if A, ¬C) ≥ P(¬(if A, C))

Based on the results from Session 1, it is very clear that the participants have acquired a license to accept the Agree Baseline inferences and that the participants do not have a license to accept the Disagree Baseline inferences. Matters are, however, less clear when it comes to the Target inference. The reason is the interaction with Relevance and the negation operator that was found, which lead to violations of the probabilistic version of the Negation Principle for the Irrelevance condition. More specifically, in Session 1 it was found for the Positive Relevance condition that P(¬(if A, C)) ≈ P(if A, ¬C). Yet, for the Irrelevance condition it was found that P(¬(if A, C)) > P(if A, ¬C), at the group level. According to Table 5, the participants are in other words only licensed to accept the Target Inference for the Positive Relevance condition. If, however, participants accept the Target Inference for the Irrelevance condition, then it would lead to violations of the above cross-task consistency constraint that ensures that probabilistic reasoning is consistent with deductive logic (Joyce, 2004; Oaksford, 2014). A central purpose of Session 2 is to investigate whether participants violate this cross-task consistency constraint for the Target Inferences.

3.3.2.1 Method

Participants

Unless otherwise noted, session 2 of Experiment 2 resembled Experiment 1. Only participants who had taken part in Session 1, and had not been excluded by the exclusion criteria in Session 1, were invited to take part in Session 2 one week later. 57 participants took part in Session 2. The participants were paid a small amount of money for their participation and a bonus of 1$ for having taken part in both sessions.

Two sets of responses of Session 2 had to be excluded due to double participation. The final sample consisted of 55 participants. Mean age was 38 years, ranging from 22 to 72, 61.8% of the participants were female; 69% indicated that the highest level of education that they had completed was an undergraduate degree or above.

Design

Session 2 had the same experimental design as Experiment 1. In total, the participants were thus presented with 24 within-subject conditions.

Materials and Procedure

Like in Session 1, each participant worked on one randomly selected scenario for each of the 8 prior × relevance within-subject conditions. The task in Session 2 followed the procedure of Experiment 1 and used the same materials.

3.3.2.2 Results

Entailment Judgments

To examine the ratings of entailment for the three types of inferences, we relied on the same set of mixed generalized linear models as in Experiment 1:

> Model M7 modelled participants' acceptance of an entailment (1 vs. 0) as a function of the Inference factor (Agree Baseline vs. Disagree Baseline vs. Target Inference), the Relevance factor (Positive Relevance vs. Irrelevance), and their interaction.
> Model M8 built upon M7 but did not include the interaction.
> Model M9 built upon M8 but did not include the Relevance factor.

Table 6 reports the performance of these models as quantified by WAIC and LOOIC.

Table 6. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M7	1460.19	0	--	1455.9	0.847
M8	1464.66	4.47	5.45	1460.3	0.091
M9	1465.41	5.22	6.13	1461.0	0.062

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC.

The information criteria displayed in Table 6 favour M7 indicating that there was an interaction making the Target Inferences slightly higher rated in the Positive Relevance condition than in the Irrelevance condition (b = 0.96, 95%-CI [0.26, 1.66], BF_{H0H1} = 0.25). But no main effect of Relevance could be found (b = -0.25, 95%-CI [-0.78, 0.28], BF_{H0H1} = 7.25). In contrast, very strong evidence in favour of a main effect of the Inference factor could be obtained. Both the posterior probabilities of accepting the entailment in the Disagree Baseline (b = -2.80, 95%-CI [-3.63, -2.02], $BF_{H0H1} = 5.73 * 10^{-43}$) and for the Target

Inference (b = -2.20, 95%-CI [-2.86, -1.59], BF_{H0H1} = 5.96 * 10^{-19}) were substantially below the posterior probability of accepting the entailment in the Agree Baseline, as displayed in Figure 3.

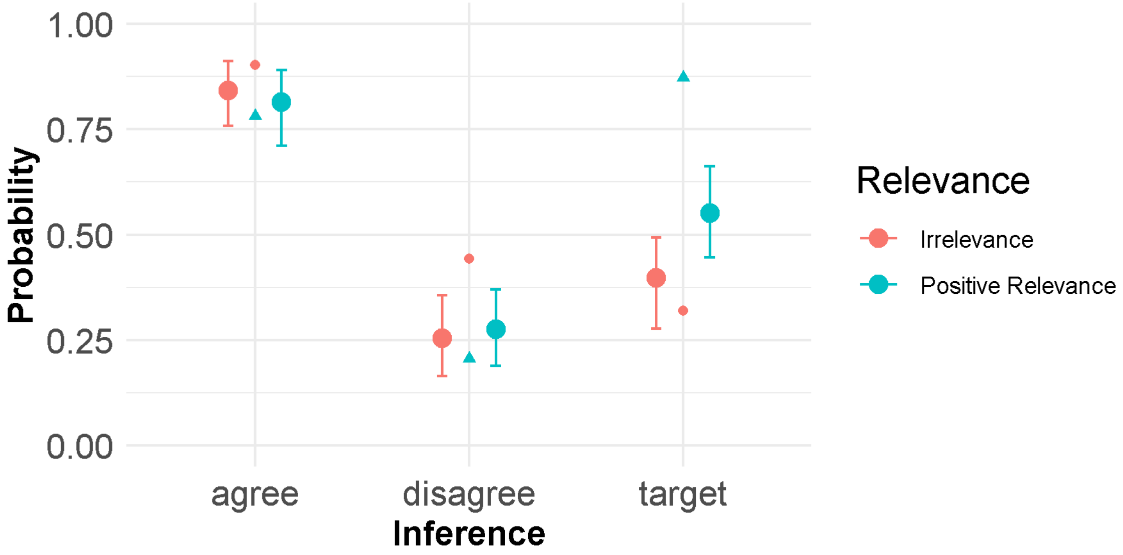


Figure 3. Weighted posterior predictive probability of acceptance of entailment. ‘agree’ = baseline for agreement; ‘disagree’ = baseline for disagreement; ‘target’ = inference to be compared with the baselines. The large dots indicate the posterior probability of accepting the entailment for a given inference. The little dots and triangles indicate the predicted acceptance based on the majority assignment of latent classes in session 1 (see Appendix B). The posterior probabilities of M7, M8, and M9, were weighted by the Akaike Weights from Table 6 to produce this plot.

As outlined in Appendix 1B, a Bayesian mixture model was applied to identify latent classes for whether the participants possessed a license to accept the entailments in session 2 based on the cross-task consistency constraint in Table 5 and their performance in Session 1. Figure 3 displays the predicted acceptance of entailment based on the assignments of latent classes of inference licenses in session 1 as little dots and triangles. The prediction assumes that P(*acceptance of entailment*) = 1 – P(*missing license*). It was found that the central tendency in the posterior probability of acceptance of entailment in session 2 was highly correlated with the predicted acceptance based on the majority assignment of latent classes in session 1, r = 0.84, t(4) = 3.13, p = 0.035.[41] The main exception was the unused license for accepting the Target Inference in the Positive Relevance condition. Here the majority response (n = 33) would predict an 87% posterior probability of acceptance of the entailment in session 2 (see Figure 3). In contrast, the participants’ actual responses were more in line with the minority response (n = 22) of having a posterior probability of 46% of acceptance of the entailment in this condition.

[41] Using a weighted average of both latent classes yields: r = 0.79, t(4) = 2.64, p = 0.058.

3.3.2.3 Discussion

It was found that while the participants had a higher posterior probability of accepting the entailment with the Target Inference than in the Disagree Baseline, the participants had a lower posterior probability of accepting the entailment with the Target Inference than in the Agree Baseline. Like in Experiment 1, participants' performance at the group level appears to exhibit a lack of preference concerning the Target Inference with a posterior probability of ca. 50% of accepting the entailment. In contrast to Experiment 1, an interaction between the Relevance and Inference factor was found, rendering M7 the preferred model. This interaction may have been the result of being exposed to the stimulus materials in Session 1 one week earlier, and it indicates a slight decrease in posterior probability of the entailment in response to the Target Inferences with irrelevance items.

Applying the cross-task consistency constraint from Table 5, we can observe that while it is consistent for participants to accept the entailment in the Agree Baseline in Session 2 following their Session 1 responses, it would have violated the cross-task consistency constraint, if the participants had accepted the Target Inference. Since the participants did not show a strong preference for accepting the Target Inference, they did not exhibit gross violations of the cross-task consistency constraint, even in the Irrelevance condition.

As shown in Figure 3, the participants' conformity to the Negation Principle for positive relevance items in session 1 of Experiment 2 gave them a license to accept the Target Inference in the Positive Relevance condition in session 2. Yet, the participants displayed a similar lack of preference with respect to the Target Inference in the Positive Relevance condition as in the Irrelevance condition. On closer inspection, however, it would have appeared problematic, if the participants had selectively exploited this license by disagreeing with Louis' objection for the Target Inference in the Irrelevance condition while agreeing with Louis' objection for the Positive Relevance condition. Doing so would have required that the participants agreed that Samuel's statements *cannot* both be true when seeing one type of item while accepting that they *can* both be true, when seeing a different type of item. In the first case, participants would have had to accept that there are no models satisfying the premise and the negation of the conclusion while agreeing, in the second case, that there are such models.

3.4 Experiment 3: Preclusion of Joint Acceptance

Some of the open-ended explanations of why the participants agreed/disagreed with Louis in Experiment 1 indicated that there may be differences in how the participants parse wide-scope negations. Table 7 outlines some of these readings:

Figure 7. Examples of Different Readings of Negation Operator in Experiment 1

Samuel:
It is false that if A, C.
….but it would be wrong to think that if A, not-C

Negation of antecedent	**Narrow-scope**	**Mixture**	**Heuristic to reduce complexity**
If *not* A, C. If *not* A, not-C.	If A, *not* C. If A, *not* (not-C).	If *not* A, C. If A, *not* (not-C).	~~It is false that if A~~, C. ….~~but it would be wrong to think that if A~~, not-C.
Lucas Scenario	*Maria Scenario*	*Julia Scenario*	*Martin Scenario*
"…the first one says if Lucas professor is not employed by the university he is attending that he meets the deadline. The second sentence implies if Lucas professor is not employed by the university he is attending that he [d]oes not meet the deadline…"	"First he says if Maria visits Adrian it's false that Craig would be jealous. Then he says, it would be wrong to think that her visit does not make Craig jealous."	"it's true that if she's not having surgery. she loses weight. It's also right that if she's having surgery, she loses weight. Either way she can lose weight. same thing."	"Both statements start with False or Wrong, so you take the reverse of the statement, and they both say Martin is raising his hand discreetly, so you can disregard that portion of the statements. The second half of each statement, therefore, so the opposite of each other - the first one says he gets the attention of the waitress, the second one says he does not…"

Note. Examples of open-ended responses from Experiment 1, used here for exploratory purposes.

Faced with such a variety of different ways of parsing the sentences, Experiment 3 sought to fix the parsing of the sentences through Louis' objections. This time, Louis' objection interprets the wide-scope negations in Samuel's statements as categorical rejections of conditional statements. Accordingly, Louis' objection to the Target Inference now takes the form of that Samuel cannot both *reject* "if A, C" and *reject* "if A, ¬C" at the same time.

Another side-effect of this reformulation is that whereas the original formulation of the task concerns the more traditional semantic question of whether the premise and the negation of the conclusion of an inference can be true at the same time, the reformulated version concerns rational acceptability/assertability and whether warrant to assert the premise precludes a warrant for denying the conclusion. Preservation of rational acceptability from the premises to the conclusion has traditionally been associated with the pragmatics of making assertions. However, there have also been attempts to replace classical notions of logical consequence with more use-oriented notions of inference based on rational assertability/acceptability (Tennant, 2002; Khlentzos, 2004). E.g., in Yalcin (2012), a consequence relation is defined based on that no information state that accepts the premises can fail to accept the conclusion, to model epistemic content.

3.4.1 Method

Participants

A total of 124 people from USA, UK, Canada, and Australia took part in the Online study, which was run on Mechanical Turk. The same exclusion criteria were used as in Experiment 1. Since some of these criteria were overlapping, the final sample consisted of 87 participants. Mean age was 41 years, ranging from 23 to 71, 56% of the participants were female; 79% indicated that the highest level of education that they had completed was an undergraduate degree or above. Applying the exclusion criteria had only slight effects on the demographic variables.

Design

Experiment 3 had the same experimental design as Experiment 1. In total, the participants were thus presented with the same 24 within-subject conditions.

Materials and Procedure

Experiment 3 followed the procedure of Experiment 1 and used the same materials. The only differences were that 1) the participants were instructed that Louis accuses Samuel of making two claims that he cannot *assert* at the same time, 2) Louis' objections were replaced by the objections in Table 8, 3) the participants were cautioned not to conflate agreeing/disagreeing with Samuel's statements and with Louis' objections, and 4) that the participants read Samuel's assertions on a separate page before processing Louis' objections. When presenting Louis' objections, Samuel's statements and the scenario texts were displayed as reminders in grey at the top of the page.

Table 8. Louis' Acceptability Objections

Target Inference	Disagree Baseline	Agree Baseline
Samuel: It is FALSE that IF A, C …But it would be WRONG to think that if A, not-C	**Samuel:** IF A, C …But it would be WRONG to think that IF A, not-C	**Samuel:** IF A, not-C …But it would be CORRECT to think that IF A, C
Louis: Wait, you cannot both **reject** that: "IF A, C" and **reject**: "IF A, not-C" at the same time!	**Louis:** Wait, you cannot both **accept** that: "IF A, C" and **reject**: "if A, not-C" at the same time!	**Louis:** Wait, you cannot both **accept** that: "IF A, not-C" and **accept**: "if A, C" at the same time!

Note. In the experiment, the words "accept" and "reject", which are marked in bold here, were made salient through a blue color to the participants. Here the structure of the objections is illustrated; in the actual experiment the propositional letters A and C were filled out with the same naturalistic scenarios as in Experiment 1.

3.4.2 Results

The same type of analysis was applied as in Experiment 1 with the following models:

> Model M10 modelled acceptance of entailment as a function of the Inference factor (Agree vs. Disagree vs. Target), the Relevance factor (Positive Relevance vs. Irrelevance), and their interaction.
>
> Model M10 built upon M11 but did not include the two-way interaction.
>
> Model M12 built on M11 but did not include the Relevance Factor.

Table 9 reports the performance of these models as quantified by WAIC and LOOIC.

Table 9. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M10	2130.94	0	--	2122.3	0.452
M11	2132.63	1.70	3.72	2124.2	0.194
M12	2131.42	0.49	3.76	2123.1	0.354

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC.

As the information criteria suggest, the full model, M10, was the winning model, but the edge given to this model was very slight as witnessed by the intermediary Akaike weights given to all models. In line with this, no main effect of relevance could be found (b = 0.09, 95%-CI [-0.43, 0.61], BF_{H0H1} = 10.5), and the Relevance factor was also not involved in an interaction ($b_{Disagree:Irrelevance}$ = 0.13, 95%-CI [-0.50, 0.74], BF_{H0H1} = 8.42; $b_{Target:Irrelevance}$ = -0.46, 95%-CI [-1.09, 0.16], BF_{H0H1} = 3.2). Like in the previous studies, strong evidence could be obtained that the posterior probability of accepting the entailment in the Disagree Baseline was below the Agree Baseline (b = -2.54, 95%-CI [-3.28, -1.84], BF_{H0H1} = -2.36 * 10^{-23}). In contrast, there was now only anecdotal evidence for a difference between the Target Inference and the Agree Baseline (b = -0.57, 95%-CI [-1.06, -0.07], BF_{H0H1} = 0.92). These findings are illustrated in Figure 5, which displays the weighted predictive posterior probabilities of all three models, when collapsing across the Relevance factor.

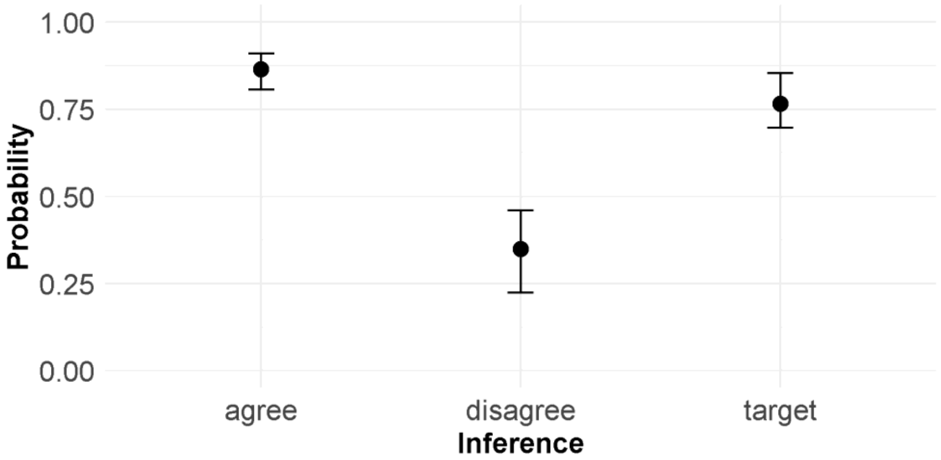


Figure 5. Weighted posterior predictive probability of acceptance of entailment. ‘agree’ = baseline for agreement; ‘disagree’ = baseline for disagreement; ‘target’ = inference to be compared with the baselines. The posterior probabilities of M10, M11, and M12, were weighted by the Akaike Weights from Table 9 and collapsed across the Relevance factor to produce this plot.

3.4.3 Discussion

It is striking that only anecdotal evidence could be obtained for a difference between the Target Inference and the Agree Baseline in Experiment 3. This indicates that participants accept the entailment of the Target Inference when Louis’ objection is presented as in Experiment 3. Experiment 3 thereby documents a *facilitation effect* compared to Experiments 1 and 2, where only a lack of preference with respect to the Target Inference could be found (with posterior probability of accepting the entailment around 50%). Apparently, fixing the parsing of the negation operator (as a wide-scope rejection of the whole statement), and changing the task to judging preservation of rational acceptability, has the effect of rendering the Target Inference acceptable to the participants.

3.5 General Discussion

In this paper, evidence was found against an unrestricted adoption of the Negation Principle both in its probabilistic version – with and without ‘then’ and ‘will’ in the examined conditionals – as well as against its truth-conditional version in an entailment task.

THE NEGATION PRINCIPLE: $\neg$(if A, then C) $\Leftrightarrow$ if A, then $\neg$C
Probabilistic version: P($\neg$(if A, then C)) = P(if A, then $\neg$C)

This principle has, however, played a prominent role in the psychological literature, where it has been cited by proponents of the Suppositional Theory of conditionals as a litmus test of their theory (Handley et al, 2006). In addition, the principle has played a role in the possible-worlds account of conditionals that is popular in linguistics (Stalnaker, 2011). The Negation Principle is moreover accepted by Adams (1998, p. 270) and certain three-valued logics of conditionals in philosophy (e.g., Cantwell, 2008a). Moreover, it follows from accounts emphasizing the connections between conditionals, subjective probability, and conditional bets based on de Finetti truth tables (Baratgin, Over, & Politzer 2013; Baratgin, Politzer, Over, & Takahaschi, 2018).

The probabilistic version of the Negation Principle is violated due to an interaction between the reason relation of indicative conditionals and the negation operator, which strongly affect their probabilities. The evidence suggests that participants only conform to this principle for Positive Relevance conditions; for Irrelevance it is systematically violated.

The significance of the violation of the Negation Principle for its probabilistic version both with and without ‘then’ and ‘will’ is that it rules out an explanation of the Relevance Effect in Skovgaard-Olsen et al. (2016a) as based on the non-truth conditional contribution of the discourse marker ‘then’, along the lines of Iatridou (1994), von Fintel (1994), and Biezma’s (2014). Had the Relevance Effect been due to the influence of ‘then’ in the investigated materials, we would expect the effects on probability evaluations of the contrast Positive Relevance ($\Delta P > 0$) vs. Irrelevance ($\Delta P = 0$) to go away once conditionals without ‘then’ in the consequent were investigated. But this turned out not to be the case; in fact, it was found that the results on the Negation task in Skovgaard-Olsen et al. (2019a) could be exactly replicated without the occurrence of ‘then’ (and ‘will’) in the consequent (see Appendix 1A). This suggests that as far as probabilistic relevance effects are concerned, there is no difference between ‘if A, C’ and ‘if A, then C will be the case’.

In Experiments 1 and 2, it was found that the participants do not have strong preferences concerning the Target Inference [¬(if A, C) ⊨ if A, ¬C], which is also required by the Principle of Conditional Excluded Middle (CEM).

$$\text{Conditional Excluded Middle} \qquad (\text{if } A, C) \lor (\text{if } A, \neg C)$$

In a famous dispute between Lewis (1973) and Stalnaker (1980), Stalnaker defended this principle while making the concession that in practice issues of vagueness introduce ties in which possible worlds are most similar to the actual world. As a result, situations may arise where neither [if A, C] nor [if A, ¬C] can be treated as true for practical purposes, although the inference principle of Conditional Excluded Middle continues to remain valid on the idealized theory. Bacon (2015, 2019) argues that while there is a dispute among Stalnaker and Lewis about the status of the principle of Conditional Excluded Middle for subjunctive conditionals, the principle is self-evident for indicative conditionals. Indeed, Bacon (2019, p. 20) proposes to treat the validity of the principle of Conditional Excluded Middle as: “a piece of data that any account of indicatives ought to be able to accommodate, not a controversial principle like its subjunctive cousin”.[42]

In contrast, Khemlani, Orenes, and Johnson-Laird (2014) hold that [if A, then C] and [if A, then ¬C] make contrary but not contradictory assertions, because it is possible for both of them to be false. Interestingly, the data in Experiments 1 and 2 suggest that the participants do not treat the Target Inference [¬(if A, C) ⊨ if A, ¬C] as a valid entailment in relation to indicative conditionals. Yet, both the Negation Principle and the principle of Conditional Excluded Middle require the Target Inference to be valid.

At the same time, a facilitation effect was found in Experiment 3 indicating that the participants *do* accept the Target Inference when the parsing of the negation operator is fixed (as a wide-scope rejection of the whole statement) and the task is changed to judging preservation of rational acceptability, instead of preservation of truth. The implication appears to be that while our results are not supportive of the entailment of the Target Inference when validity is judged by classical logic, the Target Inference would fare better on consequence relations based on preservation of acceptance, like the one expounded in Yalcin (2012).

Grice (1989, p. 80-83) discusses the possibility of using a denial of conditional as a refusal to assert the conditional in question, but not because it does not represent the facts. To illustrate: “to say “It is not the case that if X is given penicillin, he will get better” might be a way of suggesting that the drug might have no effect on X at all” (p. 81). Similarly, Adams (1998, p. 270) points out that “to assert “It is not the case that if *φ*, then *ψ*” *can* mean that “If *φ*, then *ψ*” isn’t probable enough to be asserted”. Accordingly, the fact that P(¬(if A, C)) received the highest value in the Irrelevance condition in Session 1 of Experiment 2 could be taken as an indicator that the participants treat both [if A, C] and [if A, ¬C] as unassertable. From this perspective, it is, however, strange that the participants would not permit Samuel to deny both [if A, C] and [if A, ¬C] in Experiment 3, where the facilitation effect was found. One possibility is that the participants were reacting to the oddity of why Samuel would connect unrelated sentences such as “Mark is wearing socks” and “Mark’s TV is working”

[42] Part of Bacon’s (2019) theoretical argument for the Principle of Conditional Excluded Middle for indicative conditionals relies on Adams' thesis (P(if A, C) = P(C|A) for simple conditionals). However, Adams' thesis has already been shown to break down for missing-link conditionals in Skovgaard-Olsen et al. (2016a), which is a result that the data from Session 1 (Experiment 2) replicated for bare indicative conditionals without 'then' and 'will' in the consequent.

out of the blue in sentences, if *he did not presuppose* that they were supposed to be connected. In retrospect, it might have been better to let a neutral interlocutor assert the missing-link conditionals, and have Samuel react to these assertions by denials, instead of making Samuel the originator of the missing-link items. Future research will have to determine whether the facilitation effect is robust with respect to such variations.

Finally, the participants' cross-task consistency was examined in Experiment 2 by investigating whether the participants accepted entailments for which they had no license based on their probability assignments one week earlier. It was found that this was not the case, but that the participants did have an unused license to endorse the Target Inference for the Positive Relevance condition. On closer inspection, it was found, however, that by using this license, participants would have had to adopt the doubtful cognitive state of, on the one hand, accepting that there are no models satisfying the premise and the negation of the conclusion (when responding to the positive relevance items) while agreeing, in the second case, that there are such models (when responding to irrelevance items).

A further contribution of the present paper consists in the introduction of a novel experimental task for investigating participants' acceptance of entailments, which avoids the pitfalls of previous research into deductive reasoning identified in Evans (2002). In line with work by Stenning and van Lambalgen (2008) on deductive logic being most suited for adversial contexts, and with work on the argumentative nature of logical norms for rational beliefs in Skovgaard-Olsen (2017a), the Dialogical Entailment Task proposes to investigate participants' acceptance of entailments in argumentative contexts.

In this paper, the Dialogical Entailment Task was put to use to investigate the participants' acceptance of a Target Inference [¬(if A, C) ⊨ if A, ¬C] across relevance levels. While relevance did play a role on some of the open-ended responses in Experiment 1 of why the participants had agreed/disagreed with Louis (which were used here only for exploratory purposes), in general strong effects of relevance were not found in the entailment task (as opposed to the probabilistic Negation Task). Similarly, no relevance effects on the examined and-to-if entailment judgments were found in Skovgaard-Olsen et al. (2019b), echoing the lack of relevance effects for truth-value judgments in Skovgaard-Olsen et al. (2017).

There is room for improvements of the Dialogical Entailment Task in future studies. One obvious way of improving it would be to elicit the counterexamples produced by participants who do not accept a given inference principle. Furthermore, alternative entailment relations to the classical notion of logical validity could be tested. In Experiment 3 one such variant was investigated (i.e., preservation of rational acceptability), but many further kinds exist. For instance, versions of the Dialogical Entailment Task implementing p-validity could be investigated (e.g., by having Samuel assign high probabilities to the premises of an inference and a low probability to its conclusion). Furthermore, Cantwell (2008b) recommends using preservation of non-falsity as a notion of validity for three-valued logic. Finally, Chemla, Egré, and Spector (2017) and Chemla and Egré (2018) have investigated an even more general family of entailment relations for many-valued logics by, inter alia, exploiting the possibility of exhaustively investigating all possible truth tables through computer-aided search.

These developments indicate the importance of extending the Dialogical Entailment Task to further types of entailment relations, in particularly when three-valued truth tables of indicative conditionals are investigated, such as in Baratgin et al. (2018).

3.6 Conclusion

Given that intuitive entailment judgments arguably make up one of the primary sources of data for semantic theories, it would be desirable to have a substantive body of empirical data surveying the entailment judgments of ordinary people. In this paper, a novel Dialogical Entailment Task was developed to obtain data of participants' intuitive entailment judgments in the aftermath of the methodological criticism in Evans (2002) of a previous deductive paradigm in the psychology of reasoning.

Combining this task with participants probability assignments across relevance conditions, evidence was reported against the Negation Principle [¬(if A, then C) ⇔ if A, then ¬C] both in its probabilistic version – with and without 'then' and 'will' – as well against its truth-conditional version. In its probabilistic version, it was found that the Negation Principle was only conformed to for positive relevance items; for irrelevance items it was systematically violated. As an inference principle concerning truth-preservation from the premises to the conclusion, it was found that the participants did not have strong preferences in either direction (Experiments 1, 2). Yet, when the entailment task was posed using preservation of rational acceptability, while disambiguating potential scope ambiguities, a facilitation effect was found (Experiment 3).

The Relevance Effect reported in Skovgaard-Olsen et al. (2016a) was found using indicative conditionals containing neither 'then' and 'will' in the consequent as stimulus materials. Consequently, it is possible that these results could be completely accounted for based on the meaning contribution of 'then' advanced in Iatridou (1994), von Fintel (1994), and Biezma (2014). Against such an account, it was found that the strong interaction for probability evaluations between relevance and the negation operator reported in Skovgaard-Olsen et al. (2019a) could be completely replicated using indicative conditionals without ‘then’ (and ‘will’) in the consequents. We can therefore conclude that it is not the presence of ‘then’ in the investigated stimulus materials that is driving the Relevance Effect.

References

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Adams, E. (1998). *A Primer of Probability Logic*. Stanford, CA: CLSI publications.

Arlo-Costa, Horacio (2007). The Logic of Conditionals. In E. N. Zalta (eds.), *The Stanford Encyclopedia of Philosophy* (spring 2016 Edition). Retrieved from: <http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/>.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 340-412.

Bacon, A. J. (2015). Stalnaker’s thesis in context. *The Review of Symbolic Logic*, *8*(1), 131-163.

Bacon, A. J. (2019). On the Semantics of Indicatives. Ms, University of Southern California. Retrieved from: http://yalcin.work/workshop

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking & Reasoning*, *19*, 308-328.

Baratgin, J., Politzer, G., Over, D. E., and Takahashi, T. (2018). The Psychology of Uncertainty and Three-Valued Truth Tables. *Frontiers in Psychology*, *9* (1479), 1-17.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 159-219.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bhatt, R. and Pancheva, P. (2006). Conditionals. In Everaert, M. and van Riemsdijk, H. (Eds.), *The Blackwell companion to syntax* 1 (pp. 638–687). Oxford: Blackwell.

Biezma, M. (2014). The grammar of discourse: The case of *then*. In T. Snider et al. (eds.), *Proceedings of SALT 24* (pp. 373-394). Cornell U: LSA and CLC Publications.

Bradley, R. (2007). A Defence of the Ramsey Test. *Mind*, *116*(461), 1-21.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P., & Vuorre, M. (2018, February 28). Ordinal Regression Models in Psychological Research: A Tutorial. Retrieved from http://doi.org/10.17605/OSF.IO/X8SWP

Cann, R. (1993). *Formal Semantics*. New York: Cambridge University Press.

Cantwell, J. (2008a). The logic of conditional negation. *Notre Dame Journal of Formal Logic, 49(3)*, 245-260.

Cantwell, J. (2008b). Indicative conditionals: Factual or Epistemic? *Studia Logica*, *88*(1), 157-194.

Chemla, E. and Egré, P. (2018). From Many-Valued Consequence to Many-Valued Connectives. Retrieved from https://arxiv.org/abs/1809.01066

Chemla, E., Egré, P., and Spector, B. (2017). Characterizing logical consequence in many-valued logic. *Journal of Logic and Computation*, *27*(7), 2193-2226.

Cheng, P.W. (*1997). From covariation to causation*: A *causal* power theory. *Psychological Review*, *104*, 367-405.

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology, 6,* 192.

Cruz, N., Over, D., Oaksford, M., & Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1104–1109). Austin, TX: Cognitive Science Society.

Cruz, N., Over, D. E., & Oaksford, M. (2017). The elusive oddness of *or*-introduction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 663-668). Austin, TX: Cognitive Science Society.

Douven, I. (2015). *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*. Cambridge: Cambridge University Press.

Edgington, D. (1995). On Conditionals. *Mind, 104,* 235-327.

Égré, P. and Cozic, M. (2016). Conditionals. In: Aloni, M. and Dekker, P. (eds.), *The Cambridge Handbook of Formal Semantics*. Cambridge: Cambridge University Press, 490-524.

Elqayam, S. & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction. In S. Elqayam, J. Bonnefon, and D. E. Over (eds.), *Thinking & Reasoning,* 19:3-4, 249-265.

Eva, B., & Hartmann, S. (2018). Bayesian argumentation and the value of logical validity. *Psychological Review, 125*(5), 806-821.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978-996.

Evans, J. St. B. T. & Over, D. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Thompson, V., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology, 6*, 398.

Geis, M. L. and Lycan, W. G. (1993). Nonconditional Conditionals. *Philosophical Topics, 21*(2), 35-56.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA.: Harvard University Press.

Hahn, U. and Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review*, *114*(3), 704-732.

Hahn, U., Harris, A. J. L., and Oaksford, M. (2012). Rational argument, rational inference. *Argument and Computation*, *4*(1), 21-35.

Handley, S.J., Evans, J. St. B.T., Thompson, V.A. (2006). The negated conditional: a litmus test for the suppositional conditional? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 559-569.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Blackwell Publishing.

Iatridou, S. (1994). On the contribution of then. *Natural Language Semantics*, 2(3), 171-199.

Johnson-Laird, P. N. and Byrne, R. M. J. (2002). Conditionals: A Theory of Meaning, Pragmatics, and Inference. *Psychological Review*, *109*(4), 646-678.

Johnson-Laird, P. N., Girotto, V., and Legrenzi, P. (2004). Reasoning From Inconsistency to Consistency. *Psychological Review*, *111*(3), 640-661.

Johnson-Laird, P. N., Khemlani, S. S., and Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Science*, *19*(4), 201-214.

Joyce, J. M. (2004). Bayesianism. In Miele, A. R. and Rawling, P. (Ed.), *The Oxford Handbook of Rationality* (pp. 132-155). Oxford: Oxford University Press.

Khemlani, S., Byrne, R. M. J., and Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science*, *42*(6),1-18.

Khemlani, Orenes, and Johnson-Laird (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, *151*, 1-7.

Khlentzos, D. (2004). *Naturalistic Realism and the Antirealist Challenge.* Cambridge, MA: the MIT Press.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852–884.

Kleiter, G. D. (2018). Adams’ p-validity in the research on human reasoning. *Journal of Applied Logics*, *5*(4), 775-825.

Kratzer, A. (1986). Conditionals. *Chicago Linguistics Society, 22*(2), 1–15.

Kratzer, A. (2012). *Modals and Conditionals*. Oxford: Oxford University Press.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Krzyżanowska, K, Collins, P. J. and Hahn, U. (2017). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199-205.

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lewis, D. (1975). Adverbs of Quantification. In E. Keenan (eds.), *Formal Semantics of Natural Language*, 3-15. Cambridge: Cambridge University Press.

Lewis (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Mares, E.D. (2007). *Relevant Logic: A Philosophical Interpretation*. Cambridge: Cambridge University Press.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian Models. *Frontiers in Psychology*, *5* (332), 1-5.

Oaksford, M., & Chater, N. (2009). The uncertain reasoner: Bayes, logic and rationality. *Behavioral and Brain Sciences*, *32*, 105–120.

Oaksford, M. and Chater, N. (2019). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, ISSN 0066-4308. (In Press)

Oaksford, M. and Hahn, U. (2004). A Bayesian Approach to the Argument from Ignorance. *Candadian Journal of Experimental Psychology*, *58*(2), 75-85.

Oaksford, M. and Over, D. and Cruz, N. (2018). Paradigms, possibilities and probabilities: Comment on Hinterecker et al. (2016). *Journal of Experimental Psychology: Learning, Memory, & Cognition*. (In Press)

Peterson, M. (2017). *An Introduction to Decision Theory (Second Edition)*. Cambridge: Cambridge University Press.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206-217.

Ramsey, F.P. (1929). General Propositions and Causality. In: H. A. Mellor (eds.), *F. Ramsey: Philosophical Papers*. Cambridge: Cambridge University Press, 1990.

Raidl, E., & Skovgaard-Olsen, N. (2017). Bridging Ranking Theory and the Stability Theory of Belief. *Journal of Philosophical Logic, 46*(6), 577–609. https://doi.org/10.1007/s10992-016-9411-0

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49* (4), *243*-256.

Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning, 17*(3), 247-281.

Singmann, H., Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology, 5,* article 316.

Skovgaard-Olsen, N. (2017a). The problem of logical omniscience, the preface paradox, and doxastic commitments. *Synthese*, *194*(3), 917-939.

Skovgaard-Olsen, N. (2017b). *Putting Inferentialism and the Suppositional Theory of Conditionals to the Test.* (Psychology Dissertation, University of Freiburg)

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., and Klauer, K. C. (2019a). Cancellation, Negation, and Rejection. *Cognitive Psychology*, *108*, 42-71.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., and Klauer, K. C. (2019b). Norm Conflicts and Conditionals. *Psychological Review*. http://dx.doi.org/10.1037/rev0000150

Skovgaard-Olsen, N., Kellen, D., Krahl, H., and Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'. *Thinking & Reasoning*, 23 (4), 449-482.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016a). The relevance effect and conditionals. *Cognition*, 150, 26-36.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016b). Relevance and Reason Relations. *Cognitive Science*, 41(S5), 1202-1215.

Spohn, W. (2012). *The Laws of Beliefs*. Oxford: Oxford University press.

Stalnaker, R. (1968). A theory of conditionals. *Studies in Logical Theory*, *2*, 98–112.

Stalnaker, R. (1980). A defense of conditional excluded middle. InW. L. Harper, G. Pearce, & R. Stalnaker (eds.), *Ifs* (pp. 97–104). Dordrecht: Reidel.

Stalnaker, R. (2011). Conditional propositions and conditional assertions. In A. Egan & B. Weatherson (eds.), *Epistemic modality* (pp. 227–248). Oxford: Oxford University Press.

Stalnaker, R. (2016). *Context*. Oxford: Oxford University Press.

Stenning, K., and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT University Press.

Tennant, N. (2002). *The Taming of the True*. Oxford: Oxford University Press.

Thompson, V. A. and Byrne, R. M. J. (2002). Reasoning Counterfactually: Making Inferences About Things That Didn't Happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1154-1170.

Tonhauser, J. and Matthewson, L. (2015). Empirical evidence in research on meaning. Retrieved from http://ling.auf.net/lingbuzz/002595.

van Rooij, R. and Schulz, K. (2018). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, 1-17.

von Fintel, K. (1994). *Restrictions on quantifier domains*. (University of Massachusetts Amherst dissertation)

von Fintel, K. (2011). Conditionals. In K. von Heusinger, C. Maienborn & P. Portner (eds.), *Semantics: An international handbook of meaning*, vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 33.2), 1515–1538. Berlin/Boston: de Gruyter.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part 1: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35-57.

Winter, Y. (2016). *Elements of Formal Semantics*. Edinburgh: Edinburgh University Press.

Yalcin, S. (2012). A counterexample to modus tollens. *Journal of Philosophical Logic*, *41*, 1001-1024.

Zakkou, J. (2017). Biscuit Conditionals and Prohibited ‘Then’. *Thought*, *6*(2), 84-92.

Appendix 1A: Comparison with Skovgaard-Olsen et al. (2019a)

As part of the analysis of Experiment 2, the data from its participants were compared to the data from Skovgaard-Olsen et al. (2019a, Experiment 2), which is publicly accessible at the *Open Science Framework*: https://osf.io/hz4k6/.

Like Experiment 2 of this paper, Skovgaard-Olsen et al. (2019a) conducted their experiment over the Internet using Mechanical Turk and sampling from USA, UK, Canada, and Australia. 105 people participated in the experiment in exchange for a small payment. The exclusion criteria were the same as in Experiment 2 of this paper. The final sample consisted of 67 participants. Mean age was 41.3 years, ranging from 23 to 71 years; 41.8 % of the participants were male; 68.7 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The sample differed only minimally on the demographic variables above before and after applying the exclusion criteria.

Results

Experiment 1 and 2

To investigate whether the findings from Skovgaard-Olsen et al. (2019a, Experiment 2) could be replicated with conditionals without 'then' and 'will' in the consequent, a set of mixed linear models were fitted to the data. The models had crossed random effects for intercepts and slopes by participants and by scenarios (Baayen, Davidson, & Bates, 2008) to control for the effect of replicates for each participant and item in the experimental design. To investigate whether a replication of the previous results was possible without 'then' and 'will', the models included an 'Experiment' factor that indicated whether the data originated from Skovgaard-Olsen et al. (2019a) and included 'then' and 'will' (Exp 1), or whether the data came from the present replication without 'then' and 'will' (Exp 2). The models featured the following predictors:

> Model **M1A** modelled the ratings as a function of the DV factor, encoding the three different types of conditionals (Affirm [if A, C], Wide [¬(if A, C)], Narrow [if A, ¬C)], the Relevance factor, encoding the two different relevance levels, and of the Experiment factor (Exp1 vs. Exp2). The model also included all the interactions between these three factors.
> Model **M2A** built upon M1A but did not include the three-way interaction between DV, Relevance, and Experiment.

Model **M3A** built upon M2A but did not include the two-way interaction between DV and Experiment.

Model **M4A** built upon M3A but did not include the two-way interaction between Relevance and Experiment.

Model **M5A** built upon M4A but did not include a main effect of the Experiment factor. M5A thus effectively eliminated the Experiment factor from the model of the two data sets.

In line with the previous studies, these models were implemented in a Bayesian framework with weakly informative priors, using R package brms (Bürkner, 2017). One advantage of the Bayesian framework is that it allows us to quantify the evidence in favour of the null-hypothesis in terms of Bayes factors, whereas classical statistics would only have allowed us to conclude that H_0 could not be rejected (Wagenmakers et al. 2018). Since the dependent variable consisted of continuous proportions containing zeros and ones, the values were first transformed to be within the interval [0,1] and a beta-likelihood function was used.[43] Table 1A reports the performance of these models as quantified by WAIC and LOOIC.

Table 1A. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1A	-8564.17	3.74	3.70	-8496.1	0.078
M2A	-8564.06	3.85	2.74	-8497.4	0.074
M3A	-8565.73	2.18	1.68	-8498.4	0.170
M4A	-8565.78	2.13	1.55	-8500.2	0.175
M5A	-8567.91	0	--	-8501.2	0.505

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC.

Table 1A indicates that M5A was the winning model. Consistent with this, evidence of varying degrees could be obtained in favour of the null-hypotheses which set the coefficients of these fixed effects equal to zero for all effects involving the Experiment factor, reflecting the fact that the 95% credible interval in all cases crossed zero. For the three-way interaction, strong evidence in favour of the null-hypothesis was found ($b_{IR:Narrow:Exp2}$ = -0.29, 95%-CI [-0.71, 0.13], BF_{H0H1} = 19.0; b_$_{IR:Wide:Exp2}$ = -0.22, 95%-CI [-0.76, 0.32], BF_{H0H1} = 26.61). For the two-way interaction between DV and Experiment, strong evidence in favour of the null-hypothesis was found ($b_{Narrow:Exp2}$ = 0.25, 95%-CI [-

[43] Note that in Skovgaard-Olsen et al. (2019a), a zero-or-one inflated beta likelihood function was used to report a similar qualitative pattern as in Figure 1A below. Both are compromise solutions when modelling continuous proportions containing zeros and ones.

0.05, 0.55], BF_{H0H1} = 16.23; $b\_{}_{Wide:Exp2}$ = 0.05, 95%-CI [-0.26, 0.37], BF_{H0H1} = 58.32). For the two-way interaction between Relevance and Experiment, strong evidence in favour of the null-hypothesis was found ($b_{IR:Exp2}$ = 0.10, 95%-CI [-0.23, 0.44], BF_{H0H1} = 47.46). For the main effect of Experiment, strong evidence in favour of the null-hypothesis was found ($b\_{}_{Exp2}$ = -0.10, 95%-CI [-0.31, 0.11], BF_{H0H1} = 57.38).

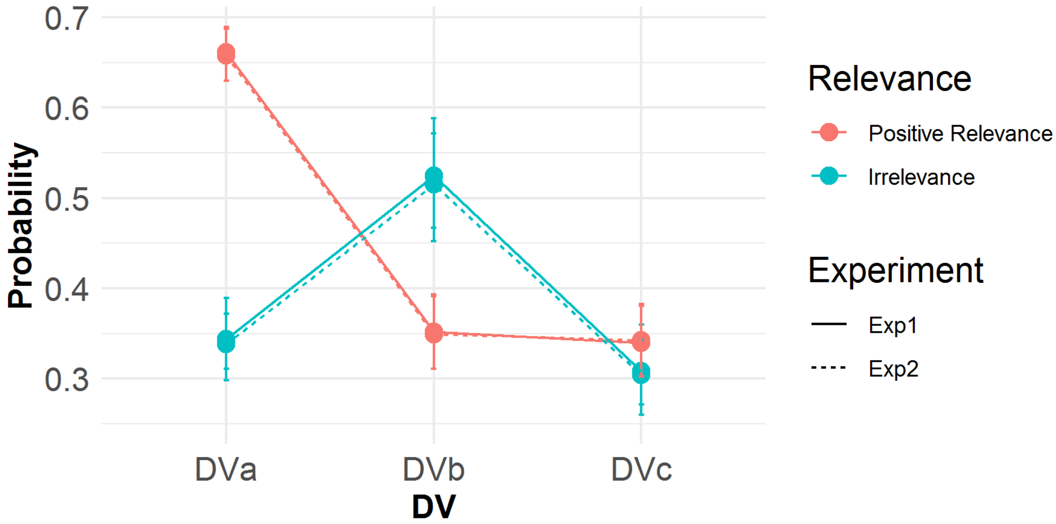


Figure 1A. Predictive posterior means from M1A, M2A, M3A, M4A, M5A weighted by the Akaike weights from Table 1A. 'Exp1' = conditionals with 'then' and 'will'; 'Exp2' = conditionals without 'then' and 'will'. ‘DVa’ = affirmative conditional; ‘DVb’ = wide scope negation; ‘DVc’ = narrow scope.

As Figure 1A indicates, the estimated marginal mean posterior probabilities across experiments were almost identical for all six measures.

Appendix 1B: Bayesian Mixture Model

It was assumed that participants’ responses came from a mixture distribution consisting of a group of participants, who had a license to accept a given entailment in session 2 of Experiment 2 based on their probability assignments in session 1 (e.g., accepting the entailment “¬(if A, C) ⊨ if A, ¬C” after conforming to the inequality “P(if A, ¬C) ≥ P(¬(if A, C))” in session 1), and a group of participants who lacked such a license (e.g., conforming to “P(if A, ¬C) < P(¬(if A, C))” in the first session 1). The upper half of Table 1B below displays the license and inference pairs:

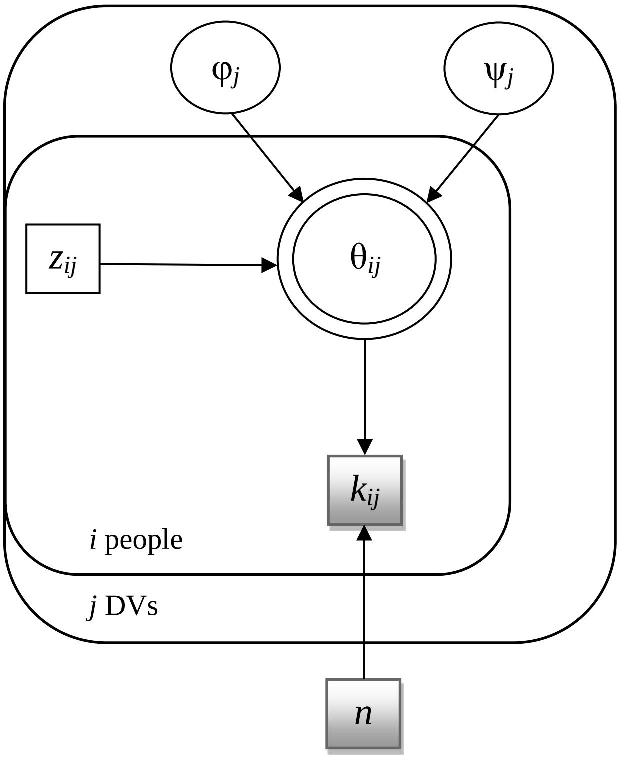
Table 1B. Applying the Cross-Task Consistency Constraint

	Inference		License	
Agree Baseline	if A, ¬C ⊨ ¬(if A, C)	only if	P(¬(if A, C)) ≥ P(if A, ¬C)	
Disagree Baseline	if A, C ⊨ if A, ¬C	only if	P(if A, ¬C) ≥ P(if A, C)	
Target Inference	¬(if A, C) ⊨ if A, ¬C	only if	P(if A, ¬C) ≥ P(¬(if A, C))	
	P(missing license)			
	Positive Relevance		*Irrelevance*	
Agree Baseline	φ = *0.54* [0.50, 0.61]	ψ = *0.22* [0.13, 0.32]	φ = *0.53* [0.50, 0.59]	ψ = *0.097* [0.04, 0.17]
Disagree Baseline	φ = *0.80* [0.71, 0.87]	ψ = *0.48* [0.43, 0.50]	φ = *0.56* [0.50, 0.66]	ψ = *0.34* [0.22, 0.46]
Target Inference	φ = *0.54* [0.50, 0.61]	ψ = *0.13* [0.05, 0.22]	φ = *0.68* [0.55, 0.80]	ψ = *0.31* [0.19, 0.43]
	P(acceptance of entailment) = 1 – P(missing license)			
Agree Baseline	*0.46*	*0.78*	*0.47*	*0.90*
Disagree Baseline	*0.20*	*0.52*	*0.44*	*0.66*
Target Inference	*0.46*	*0.87*	*0.32*	*0.69*

Note. The 95%-credible intervals for the parameter estimates are listed in square brackets. The bottom row indicates the predicted posterior probabilities of acceptance of the entailments based on the latent classes in session 1. The grey boxes in the bottom row indicate the modal session 1 classification (n = 33).

To classify participants into two latent classes, the prior recommendations and Bayesian mixture models in Lee and Wagenmakers (2014) were followed. Essentially, information or ignorance regarding the model parameters is represented by *prior distributions*. The observed data is then used to update our knowledge about the parameters, resulting in *posterior parameter distributions* (Kruschke, 2014; Lee & Wagenmakers, 2014; Skovgaard-Olsen et al. 2019b). As shown in Table 2B, participants' conformity to/violation of a given inequality (e.g., the Target Inference license) was modelled as produced by binominal rate parameters that come from two distributions (the φ_j distribution that was constrained to be above 0.5 or the ψ_j distribution, which was constrained to be below 0.5). An uninformed indicator variable (z_{ij}) classified which distribution a given participant belonged to in a given experimental condition. Based on the posterior probabilities of the indicator variables z_{ij}, each individual was classified per condition as possessing or lacking an inference license. Since Positive Relevance and Irrelevance were modelled separately, and there were three types of inference licenses (see Table 1B), six binominal rate parameters were assigned to a given participant based on four trial replications (HH, HL, LH, LL). Identifiability was ensured by applying the constraint that the two binominal rate parameters were identical across participants for the two latent classes for a given DV. The lower half of Table 1B lists the estimated parameters.

Table 2B. Bayesian Mixture Model



z_{ij} ~ Bernoulli(0.5)

φ_j ~ beta(1,1)T(0.5,1)

ψ_j ~ beta(1,1)T(0,0.5)

$$\theta_{ij} \leftarrow \begin{cases} \varphi_j \ if \ z_{ij} = 1 \\ \psi_j \ if \ z_{ij} = 0 \end{cases}$$

k_{ij} ~ Binominal(θ_{ij},n)

Note. beta(1,1)T(0.5, 1) indicates that the beta-distribution with the shape-parameters $\alpha = 1$ and $\beta = 1$ is truncated to only take values from the interval [0.5, 1]. DV ∈ {Agree license$_{\text{PO}}$, Agree license$_{\text{IR}}$, Disagree license$_{\text{PO}}$, Disagree license$_{\text{IR}}$, Target license$_{\text{PO}}$, Target license$_{\text{IR}}$}.

Chapter 4: Indicatives, Subjunctives, and the Falsity of the Antecedent[44]

Niels Skovgaard-Olsen,
Peter Collins

It is widely held that there are important differences between indicative conditionals (e.g., “If the authors are linguists, they have written a linguistics paper”) and subjunctive conditionals (e.g., “If the authors had been linguists, they would have written a linguistics paper”). A central difference is that indicatives and subjunctives convey different stances towards the truth of their antecedents. Indicatives (often) convey neutrality: for example, about whether the authors in question are linguists. Subjunctives (often) convey the falsity of the antecedent: for example, that the authors in question are not linguists. This paper tests prominent accounts of how these different stances are conveyed: whether by presupposition or conversational implicature. Experiment 1 tests the presupposition account by investigating whether the stances project – remain constant – when embedded under operators like negations, possibility modals, and interrogatives, a key characteristic of presuppositions. Experiment 2 tests the conversational-implicature account by investigating whether the stances can be cancelled without producing a contradiction, a key characteristic of implicatures. The results provide evidence that both stances – neutrality about the antecedent in indicatives and the falsity of the antecedent in subjunctives – are conveyed by conversational implicatures.

[44] This chapter has been published as:

Skovgaard-Olsen, N. and Collins, P. (2021). Indicatives, Subjunctives, and the Falsity of the Antecedent. *Cognitive Science*, *45*(11).

Authors’ Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).
Supplementary Materials: https://osf.io/w8p97/

4.1 Introduction[45]

Consider these sentences:

(1) "If the authors are linguists, they have written a linguistics paper".

(2) "If the authors had been linguists, they would have written a linguistics paper".

What, if anything, do they convey about the authors' profession? Sentence (1) seems to be silent on this issue: the authors may or may not be linguists. Sentence (2), in contrast, seems to convey that the authors are not linguists. This difference underlies a major distinction between types of conditional sentences (sentences of the form "If *A*, (then) *C*"). Sentences like (1) are typically known as *indicative* conditionals; sentences like (2), as *subjunctive* or *counterfactual* conditionals. Conditionals in general are essential to everyday language and reasoning; counterfactual conditionals, in particular, to causal and moral thinking (Byrne, 2016). The relationship between these two types is one of the mysteries about conditionals (Bennett, 2003; Quelhas, Rasga et al., 2018).

It is widely accepted that indicatives and subjunctives convey different stances towards the truth or falsity of the antecedent (the “A” clause” of "If *A*, then *C*"). But it is not clear how. Classically, researchers have distinguished between two general ways to convey meaning: semantics and pragmatics. These terms have competing definitions, but a reasonable working definition is that semantics can be understood as literal, context-independent, non-inferential, and truth-conditional meaning; and pragmatics can be understood as non-literal, context-dependent, inferential, and non-truth-conditional meaning[46] (Birner, 2014).

This paper seeks to identify how conditionals’ stances towards the antecedent are conveyed. In doing so, it addresses an important debate in linguistics, the philosophy of language, and the psychology of reasoning on the status of these stances. The paper investigates whether the stances are conveyed by a presupposition (for presupposition accounts, see, e.g., Declerck & Reed, 2001; Fillenbaum, 1974; Khemlani, Byrne, & Johnson-Laird; Levinson, 1981) or a conversational implicature (for conversational implicature accounts, see, e.g., Iatridou, 2000; Ippolito, 2003; Leahy, 2011, 2018; Mittwoch, Huddleston

[45] **Acknowledgement:** We would like to thank the reviewers, Eric Raidl, David Over, Ruth Byrne, and the audience at the Annual Conference of the German Linguistic Society (2021) for helpful discussion. We also thank Nico Vowinkel for his help with setting up the experiments.

[46] We adopt this as a working definition as a way of defining *typical* (though not necessary) characteristics. Of these typical characteristics of pragmatic meanings, perhaps the most controversial is non-truth-conditionality, since some would argue that pragmatic meanings can be truth-conditional (Birner, 2014; Carston, 2002; Recanati, 2011).

e al., 2002). We will explain these phenomena fully in the introductions to Experiments 1 and 2 respectively. Here it suffices to note that, if the stances were conveyed by a presupposition, a good case could be made for these stances being part of the conventional, semantic meaning of the conditionals. But if the stances were conveyed by a conversational implicature, the stances would clearly be a pragmatic phenomenon, and not part of the conventional meaning of the conditionals.

Important though these theoretical debates are, this is an issue with far wider relevance. For instance, whether the stance is conveyed semantically or pragmatically – and, if pragmatically, *how* - bears on how strongly the speaker is committed to that stance. Recent theories have held that, since speakers are less committed to pragmatic meanings, such meanings are plausibly deniable (e.g., Fricker, 2012; Lee & Pinker, 2010; Pinker et al., 2008). Imagine a court case in which a key issue is whether a witness had ever had a Swiss bank account. Imagine, further, that the prosecuting attorney failed to follow a clear line of questioning and, commenting later, the witness states "If I *had* had a Swiss bank account, I would have answered a direct question about it." This utterance appears to suggest that the witness did not have a Swiss bank account. But how strongly did the witness commit to that? And if he really *did* have a bank account, was his statement a lie? Experimental data suggest that participants prefer indirect to direct meanings when committing problematic acts when the hearer is likely to be antagonistic and when the potential costs are high (Lee & Pinker, 2010). Data also suggest that participants prefer to trust speakers who implied (more technically, 'implicated'), rather than explicitly said or presupposed, false information Mazzarella et al., 2018). Moreover, how the stance is conveyed may have implications for individual differences. For instance, researchers have been interested in the relationship between pragmatic reasoning and autism (Geurts, Kissine et al., 2020).

Our question bears on another key debate: whether there can be a single, unified semantic theory of indicative and subjunctive conditionals. This debate has long proved controversial, with some researchers advancing a unifying account (e.g., Edgington, 2008; Johnson-Laird & Byrne, 2002; Quelhas et al., 2018; Over et al., 2007; Pfeifer & Tulkki, 2017; Stalnaker, 1968, 1975; Spohn, 2013; Starr, 2014; Williamson, 2020) while others argue against it (e.g., Bennett, 2003; Lewis, 1973, 1976). This paper contributes to the debate by investigating salient semantic and pragmatic accounts of one key difference in meaning between indicative and subjunctive conditionals, the stances towards the antecedent, and ascertaining whether these stances belong to the conventional, context-independent, semantic meaning of the conditionals or their non-conventional, context-dependent, pragmatic meaning. In the rest of the introduction, we first outline the range of stances a conditional can be used to convey towards its antecedent, before previewing the experiments.

The Truth/Falsity of the Antecedent: Defining Indicatives and Subjunctives

Theoretical and corpus-linguistic work suggests that conditionals are, in fact, compatible with a range of stances towards their antecedent. They can convey that the speaker takes the antecedent to be true, false, or somewhere in between. To illustrate, consider the following examples from Declerck and Reed (2001). These examples illustrate categories from their extensive typology, which relates the grammatical (morpho-syntactic) form of a conditional to its stance towards the antecedent.

(3) “If I had a problem, I always went to my grandmother” (ibid., p. 50).

This conditional conveys that its antecedent is known to be true. Conditionals like this, with factual antecedents, often describe past repetitive habits (ibid.). Compare example (3) with the next example:

(4) “I hope Liverpool won their home match yesterday. If they did, they still have a chance of winning the championship” (ibid., p. 54).

This conditional conveys that its antecedent is an open – a real – possibility. Compare example (4) with the next example:

(5) “I would have been happy if we had found a solution” (ibid., p. 54).

This conditional conveys that the antecedent is false in the actual world: it is counterfactual.

What sets the counterfactual-antecedent (5) apart from the others[47] is a distinctive use of verbal morphology in the antecedent[48]. The morphology appears to be standard past

[47] See also tentative-antecedent examples, such as the following, which should be read as referring to the future: “I would be happy if we found a solution” (Declerck & Reed, 2001, p. 54). This conditional is tentative about the antecedent: it is possible, but unlikely, that the antecedent will prove true. There is “fake tense” here too, with the past-tense morphology conveying remoteness of possibility or tentativeness.

[48] The verbal morphology in the consequent appears less distinctive. For example, speakers can use the modal auxiliary (Huddleston, 2002; Mittwoch, Huddleston et al., 2002) – some would say past tense (e.g., Iatridou, 2000) – "would" in factual-antecedent conditionals. We could paraphrase example (3) as "If I had a problem, I would always go to my grandmother". "Would" can also appear without "have" in the consequent of counterfactual conditionals, as in this example: "If the colonial powers hadn't invaded, the Americas would be very different" (Starr, 2019).

perfect, “had found”. But this morphology does not simply situate the antecedent in a particular time: it is, in a sense, a "fake tense" (Iatridou, 2000). The counterfactual-antecedent refers to the past but uses the extra layer of past tense – the past-perfect “had” – to indicate that the antecedent situation did not actually obtain. This use of morphology has led von Fintel (2012) to refer to counterfactuals as "additional-past conditionals". But counterfactual-antecedent conditionals can also occur in the following form, referring to the present:

(6) "If he were rich, he would be smart" (Iatridou, 2000, p. 232).

Here the antecedent conveys counter-factuality through "were", which some class as being in the subjunctive mood (e.g., Starr, 2019) and others as being in a distinct "irrealis" mood (Huddleston, 2002; Mittwoch, Huddleston et al., 2002).

Following convention, we will focus on the distinction between indicative and subjunctive, or counterfactual, conditionals here, although the label “subjunctive” has well-known problems (see, e.g., Starr, 2019; von Fintel, 2012). We take it, moreover, that by "indicative" most researchers would mean conditionals like (4) above, which we will call "open-antecedent conditionals" to indicate that usually the speaker does not know whether the antecedent or consequent are true or false (Mittwoch, Huddleston et al., 2002). We take it, also, that by "subjunctive" or "counterfactual" most researchers would mean conditionals like example (5) with the distinctive extra-layering of "fake past" in the antecedent and a modal auxiliary "would" or "would have" in the consequent.

Previous Findings

There is experimental data to support the theoretical and intuitive distinctions between indicative (open-antecedent) and subjunctive conditionals. For instance, in Thompson and Byrne (2002), when participants indicated "What, if anything, you think [the speaker] meant to imply?" by indicative and subjunctive conditionals, different patterns emerged for indicatives and subjunctives. Some 54% of participants took the speaker of an indicative to imply nothing; of the remaining participants, 24% took the speaker of an indicative to imply the truth of the antecedent and 44% the truth of the consequent. These data suggest that, at least for many participants, indicatives are compatible with either the truth or falsity of the antecedent (and consequent). For subjunctives, in contrast, around half (48%) of participants took the speaker of a subjunctive to imply the falsity of the antecedent and around half (47%) the falsity of the consequent, a far higher rate than for indicatives (respectively, 2% and 1%).

A distinction emerges between indicatives and subjunctives in other tasks investigating conditional inferences (Byrne & Tasso, 1999; Thompson & Byrne, 2002).

Moreover, in Quelhas, Rasga, and Johnson-Laird (2018) participants selected among different paraphrases of indicative and subjunctive conditionals. Participants tended to choose a paraphrase of indicative conditionals to the effect that antecedent and consequent were both possible, and a paraphrase of subjunctive conditionals to the effect that both antecedent and consequent once were possible but no longer are. A substantial minority also selected a paraphrase for the subjunctives to the effect that antecedent and consequent were both possible. Given this range of data, and further evidence from processing studies (e.g., Santamaria, Espino et al., 2005; De Vega, Urrutia et al., 2007; Ferguson & Sanford, 2008; Stewart, Haigh et al., 2009), we can grant that indicative and subjunctive conditionals can convey different stances towards their antecedent, with subjunctives often conveying the falsity of their antecedents. But just how, and when, are these stances conveyed?

Entailment

A first, semantic possibility is that conditionals semantically entail their stances towards the antecedent: for instance, that subjunctives semantically entail the falsity of the antecedent. One sentence entails a second if the second sentence is true in every model satisfying the first sentence. The sentence “There is a polar bear in the zoo enclosure” entails “There is a mammal in the zoo enclosure”: the first cannot be true without the second also being true. Famous examples like (7) and (8) below, however, suggest that this constraint is too strong for accounting for the falsity of the antecedent of subjunctive conditionals:

> (7) "If Jones had taken arsenic, he would have shown exactly those symptoms which he does in fact show" (Anderson, 1951, p. 37).

Since a speaker of this conditional could use (7) to argue that Jones had, in fact, taken arsenic, the sentence does not entail that the opposite is true – i.e. that Jones did not take arsenic (von Fintel, 1997, 2012; Stalnaker, 1975, 2014). Such conditionals are commonly referred to as "Anderson conditionals"; they will feature in our experiments below.
A similar case is example (8):

> (8) "If the butler had done it, we would have found blood on the knife. The kitchen knife was clean; therefore the butler did not do it" (Iatridou, 2000, p. 232).

The second sentence, here, does not seem redundant: the *modus tollens* argument does not seem to beg the question. But if the first sentence had already entailed that the butler did not do it, the argument would have been superfluous (Iatridou, 2000, Stalnaker, 1975, 2014).

Similarly, if subjunctives ‘A > C’ are given the truth conditions of being true if a base conditional (‘if A, C’) is true *and* the antecedent is false, we immediately run into trouble with modus ponens (MP), modus tollens (MT), affirmation of the consequent (AC), and denial of the antecedent (DA):

$$MP\colon \frac{\begin{array}{c} A > C \\ A \end{array}}{\therefore C} = \frac{\begin{array}{c} If\ A,C \\ \neg A \\ A \end{array}}{\therefore C} \quad MT\colon \frac{\begin{array}{c} A > C \\ \neg C \end{array}}{\therefore \neg A} = \frac{\begin{array}{c} If\ A,C \\ \neg A \\ \neg C \end{array}}{\therefore \neg A} \quad AC\colon \frac{\begin{array}{c} A > C \\ C \end{array}}{\therefore A} = \frac{\begin{array}{c} If\ A,C \\ \neg A \\ C \end{array}}{\therefore A} \quad DA\colon \frac{\begin{array}{c} A > C \\ \neg A \end{array}}{\therefore \neg C} = \frac{\begin{array}{c} If\ A,C \\ \neg A \\ \neg A \end{array}}{\therefore \neg C}$$

In MP inferences we see that the conclusion is now inferred from an inconsistent premise set, in MT one of the premises presupposes what the conclusion is supposed to establish, in AC the conclusion is inconsistent with one of the premises, and in DA one of the premises is redundant. Normally, AC and DA are considered invalid forms of inferences, but not due to these problems.

To account for the stances towards the antecedent, we need other, more flexible linguistic phenomena. In this paper we consider two such phenomena: presupposition and conversational implicature. We will define these terms below.

The Experiments

Two experiments, below, use classic diagnostics for being a presupposition (Experiment 1) or a conversational implicature (Experiment 2) to address the question of how conditionals convey the stances toward the antecedent. For these experiments, novel stimulus materials were developed which manipulate participants’ belief states (i.e., neutrality, belief, or disbelief) via occluded pictures. These stimulus materials were pretested to investigate whether participants made the appropriate belief state assumptions as a function of the picture shown, and whether they rank-ordered indicative and subjunctive conditionals accordingly.[49]

4.2 Experiment 1: Presuppositions

It is a common idea that there is some difference in status between the stances of indicative and subjunctive conditionals towards the antecedent and other content of the conditional. Within mental models theory, for instance, it has been common to speak of the falsity of antecedent and consequent as part of the default meaning (e.g., Khemlani, Byrne, & Johnson-Laird, 2018) but also of the "presupposed facts" (see, e.g., Byrne, 2005, 2016, 2017; Espino & Byrne, 2018). This notion of presupposed facts connects with a long tradition in linguistics and philosophy according to which counterfactual conditionals presuppose the falsity of their

[49] The pilot study can be found on the osf repository: https://osf.io/w8p97/.

antecedents (see, e.g., Fillenbaum, 1974; Declerck & Reed, 2001; Levinson, 1981). Presupposition is a linguistic category that is often used for capturing further aspects of content that are not directly represented in a sentence's truth conditions, which, however, make up a precondition for the sentence being true, or appropriately assertable.

To presuppose information is to linguistically mark it as taken for granted (Beaver & Geurts, 2014) or to act as if it could be made an uncontroversially part of the shared common ground between speaker and interlocutor (Potts, 2007, 2015). Precise definitions of the term "presupposition" are contested. But on a common view, presuppositions are marked, linguistically, with presupposition triggers.[50] Triggers include, e.g., the following:

> (9) factive verbs, such as "know"
> "The reader knows that this paper is fantastic" presupposes that the paper in question is fantastic.
> (10) aspectual verbs, such as "continue"
> "The reader continued to enjoy the paper" presupposes that the reader was enjoying the paper.
> (11) definite descriptions, such as “The [Noun Phrase]”
> “The broken glass glittered in the sunlight” presupposes that there was broken glass.

In some lists, one would also see the antecedent of counterfactual conditionals (e.g., Levinson, 1981) but, as we will see, their inclusion is contentious. Some researchers also argue that the openness of the indicative conditionals is due to a presupposition (see, e.g., Declerck & Reed, 2001, Byrne & Johnson-Laird, 2019). If presuppositions convey the different stances of indicatives and subjunctives towards their antecedent, then the presuppositions attach to some element of the antecedent: presumably, the morphological form of the main verb in the antecedent. How well, then, does a presupposition account for intuitions and linguistic data? To answer this question, we must consider a characteristic known as ‘projection’. This characteristic is at work in examples (12) and (13):

> (12) "The East German ambassador laughed."
> (13) "The East German ambassador did not laugh."

[50] This is a simplification. Some theories take presuppositions to be more pragmatic: to be performed by the speaker, rather than triggered conventionally (Stalnaker, 1972, 1974, 2014). There is also debate about the extent to which presuppositions can be wholly conventional as attaching to particular lexical items or whether they can be reconstructed from general conversational principles (Simons, 2006; Beaver & Geurts, 2014).

Here there is a presupposition trigger, the definite description “The East German ambassador”, which presupposes the existence of the said ambassador at the relevant time. In (13), this trigger is embedded under negation, but the presupposition survives: it *projects* under negation. Such projection behaviour is a hallmark of presuppositions, and it is not one that is found with semantic entailments (Simons, 2006). Indeed, it is a classic diagnostic test for being a presupposition to see whether information projects under various operators (Beaver & Geurts, 2014). In the so-called "family of sentences test" (see, e.g., Kadmon, 2001), one considers whether a candidate for being a presupposition survives in a set of related sentences: in negation, questioning, embedding under modals, and embedding in the antecedent of a conditional. Table 1 illustrates this test for the East German ambassador examples, and how the test might apply to indicative and subjunctive conditionals.

Table 1. Family of Sentences Test

	Projects?			
Test Sentence	*There was laughter*	*There is an East German ambassador*	*Speaker is open to the possibility that the East German ambassador will laugh*	*Speaker doubts that the East German ambassador laughed*
The East German ambassador did not laugh.	No	Yes	It is not the case that if the East German ambassador laughs…	It is not the case that if the East German ambassador had laughed….
Did the East German ambassador laugh?	No	Yes	Will the guest be offended, if the East German ambassador laughs?	Would the guest have been offended, if the East German ambassador had laughed?
Possibly, the East German ambassador laughed	No	Yes	Possibly, if the East German ambassador laughs…	Possibly, if the East German ambassador had laughed…
Diagnosis	*Entailment*	*Presupposition*	*Unclear*	*Unclear*

A range of existing empirical work has used such embedding to test for projection. For instance, studies have shown projection under negation for the presuppositions of factive verbs “realize” and “know” – i.e. the truth of the complement (Chemla & Bott, 2013); for the presupposition of “stop” – i.e. that “stop X” presupposes “used to X” (Romoli & Schwarz, 2015); and for the presuppositions of “the” and “win” – i.e. “the X” presupposes X’s existence, and “win X” presupposes competing for X (Smith & Hall, 2011).

However, it turns out that presuppositions do not always survive; presuppositions that project can sometimes nevertheless be directly denied (Simons, 2006; Kadmon, 2001). For instance, example (14) directly denies the presupposition in example (13):

(14) "The East German ambassador did not laugh. There is no East German ambassador, because East Germany no longer exists."

Importantly, though, direct denial only seems to work when the presupposition trigger is embedded under an operator (Beaver & Geurts, 2014). Compare the successful denial in (14), where the presupposition trigger is embedded under negation, with the attempted but infelicitous denial that follows (15):

(15) "The East German ambassador laughed. There is no East German ambassador."

These rather specific contexts, then, do not undermine the use of projection as a diagnostic test. Can projection behavior, then, account for the stances towards the antecedent conveyed by conditionals? With indicative conditionals, there seems to be no great problem. If we ultimately want a theory that can allow all stances towards the antecedent, we might wonder whether presuppositions can do the required work: whether there are distinct triggers for the different stances. But there are promising differences in form between conditionals that convey different stances on the truth of the antecedent which might serve as triggers (see, e.g., Declerck & Reed, 2001). But with subjunctive conditionals, there seem to be considerable difficulties. As we have seen, presuppositions can be cancelled through direct denial when they are embedded under an operator. But a presupposition account predicts that the falsity of the antecedent should be conveyed when there is no embedding. Examples (7) and (8) already challenges this notion via their cancellation of the falsity of the antecedent of the respective subjunctives (though see Stalnaker, 2014 and Zakkou, 2019 for further discussion).

In Experiment 1, we test the presupposition account more systematically. Experiment 1 explores whether the stances towards the antecedent – neutrality for indicatives, and disbelief for subjunctives – exhibit the projection behaviour of presuppositions. To investigate this, we apply the family of sentences test (Kadmon, 2001) to see whether these belief-state assumptions project past negation operators ("it is not the case that…"), possibility-modals ("possible, …"), and interrogatives ("Martin, do you think that … ?"). More specifically, we test: (1) for stand-alone indicatives, whether neutrality towards the antecedent projects past these three operators; (2) for stand-alone subjunctives, whether disbelief towards the antecedent projects past the operators; and (3) for Anderson conditionals, whether belief in/neutrality towards the antecedent projects past the operators.

Translated into a statistical model, the presupposition hypothesis holds that there should be no differences across the various types of operator (referred to as the "DV Type factor" below). This model (M5) is tested against a collection of other models which allow for differences between the operators, as explained below.

4.2.1 Method

Participants, and sampling procedure shared for all experiments

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 118 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk from the USA, UK, Canada, and Australia. They were paid a small amount of money for their participation. The following *a priori* exclusion criteria were used: not having English as native language, completing the task in less than 240 seconds or in more than 3600 seconds, failing to answer at least one of two simple SAT comprehension questions correctly in a warm-up phase, and answering 'not serious at all' to the question 'how serious do you take your participation' at the beginning of the study. The final sample consisted of 78 participants. Mean age was 37.41 years, ranging from 21 to 65. 38.46% of participants identified as female; 61.54% identified as male. 79.49 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a within-participants design with the following factors varying within participant: DV Type (assert vs. negation vs. possible vs. question), Prior (high probability (H) vs. low probability (L)) and Conditional Type (indicative vs. subjunctive). To allow for four trial replications for each cell of the design, each participant in total went through 64 within-subject conditions.

Materials and Procedure for All the Experiments

For a pilot study,[51] a pool of 24 different pictures was created, and 16 pictures selected for further studies based on which pictures had the highest rate of inducing the intended belief state assumptions consistently across the four conditions. In all the experiments reported below, the various within-participants conditions were thus randomly assigned to a pool of the 16 different pictures. Random assignment was performed without replacement such that each participant saw a different picture for each condition. This ensured that the mapping of

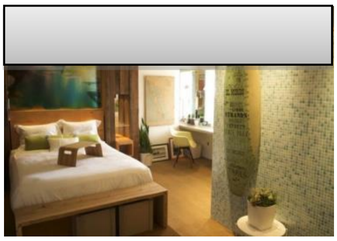
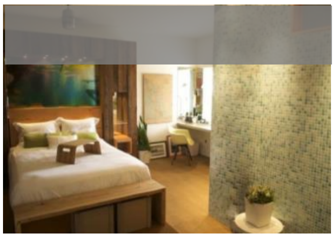
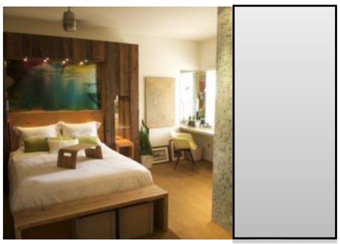
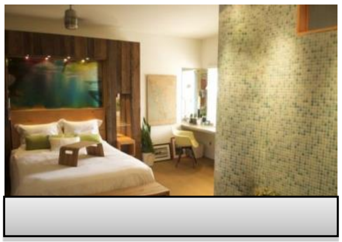
[51] See: https://osf.io/w8p97/.

condition to picture was counterbalanced across participants preventing confounds of condition and picture content.

To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002). Moreover, to ensure that the pictures were displayed properly if the participants completed the study on a smartphone, participants were asked to turn their smartphone in horizontal orientation, if they were using one.

The 16 possible pictures all implemented the four conditions indicated in Table 2. The pictures feature familiar places like bedrooms, cafés, and kitchens, where we stereotypically have expectations about likely objects (e.g., a pendant lamp in a bedroom) and unlikely objects (e. g. a surfboard in a bedroom). As Table 2 shows, the pictures additionally featured grey boxes that manipulate the assertability of indicative and subjunctive conditionals. These boxes operationalize the Occlusion variable (see also Baratgin, Over et al., 2013):

Table 2. Stimulus Materials and Experimental Conditions

Indicative, occluded	*Subjunctive, not occluded*
P(there is a pendant lamp in the bedroom) = H	
IF there is a pendant lamp in the bedroom, THEN it hangs above the bed.	IF there had been a pendant lamp in the bedroom, THEN it would have hung above the bed, where indeed something is hanging.
	
P(there is a surfboard in the bedroom) = L	
IF there is a surfboard in the bedroom, THEN it stands against the wall	IF there had been a surfboard in the bedroom, THEN it would have stood against the wall.
	

Note. 'H' = high prior probability; 'L' = low prior probability. Note that the upper right corner is an example of the so-called "Anderson conditional".

To create a situation in which indicative conditionals are assertable (left column), we used a grey box to hide the location specified by the consequent of the conditional. For instance, due to the grey box in the lower left picture, participants cannot verify for certain whether there is a surfboard standing against the wall, but they are expected to deem it unlikely. Our pilot study confirmed that participants make these judgments of high vs. low prior probability.

To create a situation in which subjunctive conditionals are assertable (right column), we either placed a transparent grey box where the object was supposed to be (upper right corner), or a non-transparent grey box in an irrelevant location that had no bearing on the presence of the object mentioned in the conditional (lower right corner). For instance, when assessing the conditional ‘If there had been a surfboard, then it would have stood against the wall’ based on the picture in the lower right corner, participants can see for certain that there is no surfboard standing against the wall, and thus maintain disbelief in the presence of a surfboard on the picture. In contrast, the transparent[52] grey box in the upper right corner was introduced to create a situation for asserting so-called Anderson conditionals (e.g., “If there had been a pendant lamp in the bedroom, then it would have hung above the bed, where indeed something is hanging”) which take the subjunctive form but are asserted without doubting the antecedent. Due to the transparent grey box, participants can verify that there is an object that appears to fit the description at the place mentioned in the consequent. Nevertheless, the lack of full transparency is intended to make the guarded form of the subjunctive mood for the conditional assertion sound more natural.

A feature of the conditionals in Table 2 is that the consequent depends for its truth on the antecedent. The conditionals were designed in this way, because it enabled us to manipulate belief states based on the pictures and the grey boxes in a way that would also permit the formulation of Anderson conditionals. Since Experiments 1 and 2 only concern belief states targeting the antecedent, this feature does not matter for their purpose.

Procedure specific to Experiment 1

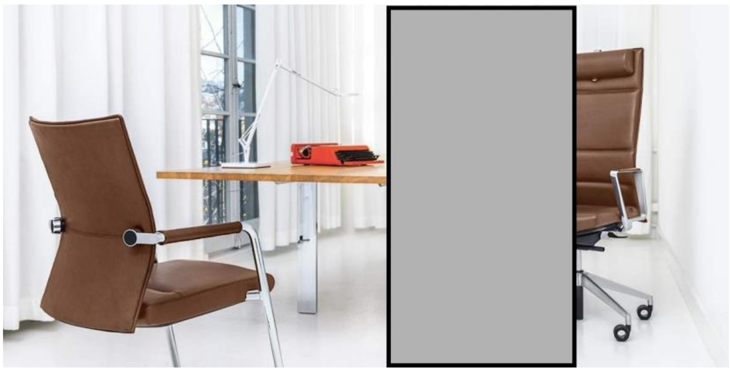
The experiment was split into 16 blocks, each implementing one of the four trial replications of the four Prior × Conditional Type within-subject conditions. For each block a picture was randomly assigned from the pool of 16 pictures used. The order of the blocks was randomized and there were no breaks between blocks. Within a given block, participants were presented with the four DV Types on separate pages in random order with the same picture.

[52] Note that in their rendering on the computer screen, the pictures were larger and so the grey box really was transparent to the participants.

Before beginning with the actual experiment, participants completed four practice trials with one of the excluded pictures, where it was emphasized that it was important to pay attention to subtle differences between the wordings on the various pages. To complete these trials, participants were given the following instruction:

> In the following, you are going to see pictures and statements made by Dennis concerning the pictures shown. Your task is to indicate which assumptions you would make concerning what Dennis believes based on what he says.

On each page, participants were then presented with a statement by Dennis in response to the selected image, corresponding to the within-subject condition displayed at the moment. For instance, a participant might first have seen the following image:



Together with the following statement:

> **Dennis**:
> Possibly, IF there is a monitor in the office, THEN it stands on the table. (*possible*)

The task of the participants was to indicate which of the following three statements best describes Dennis' state of mind when reading his statement:

> Dennis disbelieves that there is a monitor in the office.
> Dennis neither believes nor disbelieves that there is a monitor in the office.
> Dennis believes that there is a monitor in the office.

On the three pages that followed, participants were given the same task with the following three statements in random order:

> IF there is a monitor in the office, THEN it stands on the table. *(assert)*

It is NOT the case that IF there is a monitor in the office, THEN it stands on the table. *(negation)*

Martin, do you think that IF there is a monitor in the office, THEN it stands on the table? *(question)*

4.2.2 Results

Table 3 reports descriptive statistics for participants' belief state ascriptions.

Table 3. Descriptive Statistics.

	Assert	Negation	Possible	Question
Indicative, HH	50.96% Neutral	50.32% Disbelief	63.46% Neutral	73.40% Neutral
Indicative, LL	57.69% Neutral	48.08% Disbelief	63.46% Neutral	70.19% Neutral
Subjunctive, HH	37.18% Belief	55.45% Disbelief	37.50% Neutral	42.95% Neutral
Subjunctive, LL	45.83% Disbelief	55.45% Disbelief	47.12% Neutral	54.17% Neutral

Note. Due to the categorical nature of the response variable, the descriptive statistics is reported as percentages of the modal values. 'HH' = high prior probability of antecedent and consequent; 'LL' = low prior probability of antecedent and consequent.

Given the design, there were replicates for each participant and picture. Hence, it was not appropriate to assume that the data were independently and identically distributed. Accordingly, linear mixed-effects models with crossed random effects for intercepts and slopes by participants and by pictures were used (Baayen, Davidson, et al., 2008).[53] This analysis was conducted using the statistical programming language R (R Core Team, 2013) and the package brms for mixed-effects models in Bayesian statistics (Bürkner, 2017) with a multinominal likelihood and a logit link function for categorical regression. The following family of models was fit to the data, which vary in their fixed effects:

(M1) a maximal model that treats participants' selections as a function of the DV Type factor (assert vs. negation vs. possible vs. question), the Prior factor (high vs. low), the Conditional factor (indicative vs. subjunctive) and their three and two-way interaction.

(M2) a model that is obtained from the maximal model (M1) by removing the three-way interaction.

(M3) a model that is obtained from (M2) by removing the two-way DV Type:Prior interaction.

(M4) a model that is obtained from (M3) by removing the two-way Conditional:DV Type interaction.

[53] Conditional*Prior was kept fixed as random effects by participants and by pictures.

(M5) a model that is obtained from (M4) by completely removing the DV type factor. (M5) thereby implements the presupposition model.

Hypotheses concerning the presence/absence of effects are tested here and below by setting coefficients of the maximal model (M1) equal to zero. In this way, evidence in favour of, e.g., the H_0 that there is no simple effect of the DV type factor can be quantified in terms of Bayes factors, where classical significance testing would only have permitted us to conclude that H_0 could not be rejected (Wagenmakers et al., 2018). To be able to quantify the strength of evidence both against and in favour of H_0, we rely on the following qualitative interpretation of Bayes factors (Lee & Wagenmakers, 2014): (Anecdotal evidence for H_1) $\frac{1}{3} < BF_{H0H1} < 1$, (Moderate evidence for H_1) $\frac{1}{10} < BF_{H0H1} < \frac{1}{3}$, (Strong evidence for H_1) $\frac{1}{30} < BF_{H0H1} < \frac{1}{10}$, (Very Strong evidence for H_1) $\frac{1}{100} < BF_{H0H1} < \frac{1}{30}$, (Extreme evidence for H_1) $BF_{H0H1} < \frac{1}{100}$. Values above 1 indicative evidence in favour of H_0 since this scale is mirrored by applying the following ratio: $BF_{H0H1} = \frac{1}{BF_{H1H0}}$. Table 4 reports the performance of the models as quantified by the leave-one-out cross validation criterion and WAIC.

Table 4. Model Comparison

	LOOIC	Δelpd	SE	WAIC	Weight
M1	8150.2	0	--	8147.7	0.611
M2	8152.3	-1.1	3.7	8149.8	0.213
M3	8152.6	-1.2	5.3	8150.3	0.177
M4	8202.8	-26.3	10.1	8200.5	0.000
M5	8795.1	-322.5	27.6	8793.2	0.000

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy.

The information criteria showed a preference for M1-M3 and clearly rejected the model (M5) corresponding to the presupposition hypothesis of no effect of embedding indicative and subjunctive conditionals under negation, possibility, and interrogation operators. Since the differences between M1-M3 were small, Figure 1 plots the posterior predictions of all three models as weighted by their respective model weights from Table 4.

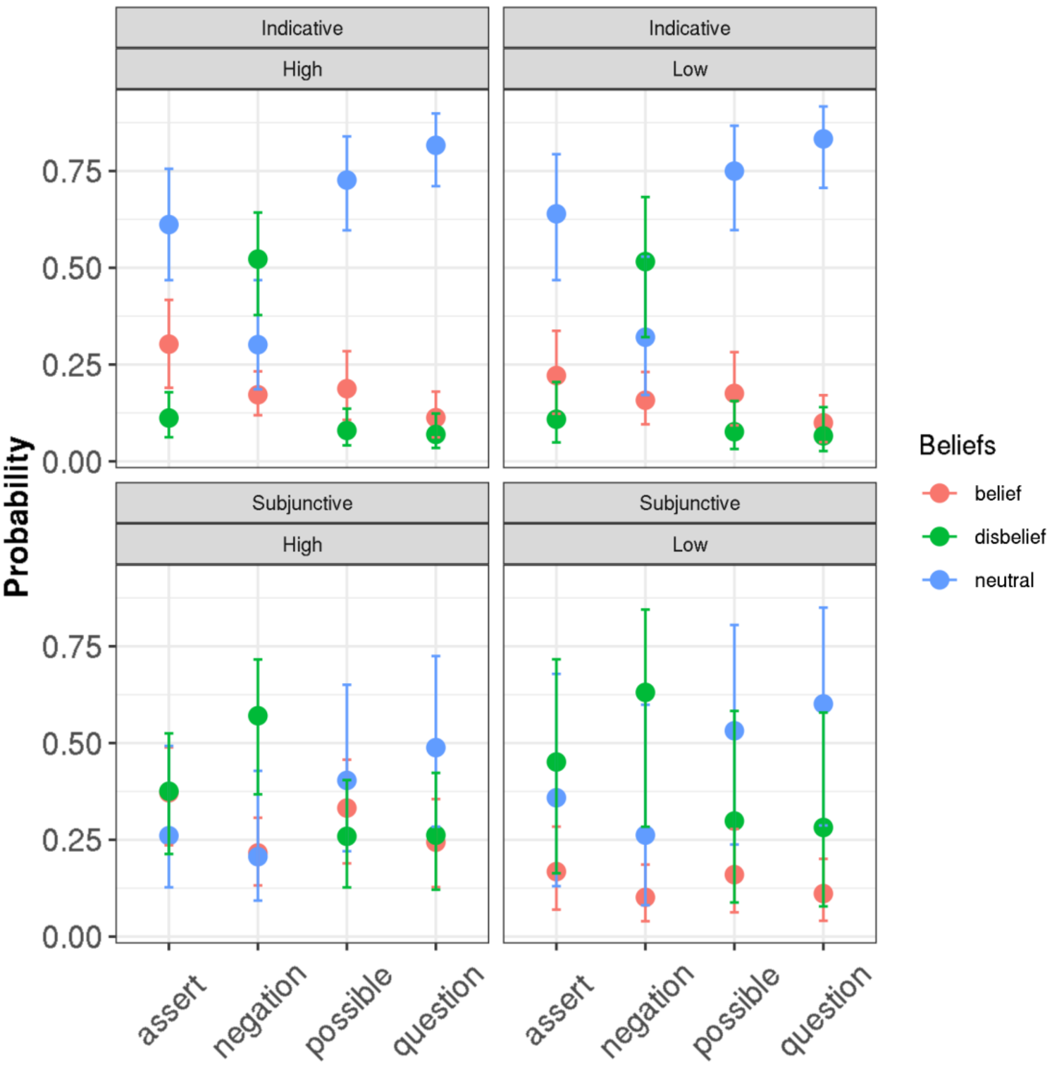


Figure 1. Posterior Predictions based on M1-M3. The posterior predictive probabilities of selecting belief/neutrality/disbelief across within-subject conditions are displayed. ‘High’ = high prior probability of antecedent and consequent. ‘Low’ = low prior probability of antecedent and consequent. Error-bars represent 95% credible intervals.

The results indicate that there was a contrast between ‘assert’ and the other DV types across conditions. In particular, strong evidence could be obtained that use of 'Negation' increased the posterior probability of Disbelief ($b_{\text{Negation\_Disbelief}}$ = 2.21, 95%-CI [1.72, 2.72], BF_{H0H1} < .001) and that embedding under 'Possible' and 'Question' both increased the posterior probability of 'Neutral' ($b_{\text{Possible\_Neutral}}$ = 0.77, 95%-CI [0.35, 1.18], BF_{H0H1} = .02; $b_{\text{Question\_Neutral}}$ = 1.58, 95%-CI [1.13, 2.03], BF_{H0H1} < .001). There was, moreover, weaker evidence of a three-way interaction in particular based on the following contrast, which indicates a higher posterior probability of selecting the ‘Neutral’ category for a specific level of the Condition and Prior factors ($b_{\text{Subjunctive:PriorLL:Question\_Neutral}}$ = 1.21, 95%-CI [0.29, 2.12], BF_{H0H1} = .37).

4.2.3 Discussion

As a manipulation check, we can gauge the belief-state attributions of stand-alone assertions for their plausibility across conditions. What we find is a general tendency to attribute doxastic neutrality towards the antecedent for indicative conditionals (across Prior levels), disbelief/neutrality in the counterfactual conditionals (subjunctive, low prior), and an elevated posterior probability of selecting ‘belief’ with the Anderson conditionals (subjunctive, high prior) compared to the counterfactual conditionals (Table 3, Figure 1). Since these belief-state attributions overall match prior theoretical expectations, the results from Experiment 1 can be used to test the presupposition hypothesis. Translated into a statistical model, the presupposition hypothesis holds that there should be no differences across the various levels of the DV Type factor. Accordingly, if the presupposition hypothesis had accounted for the data, we would expect M5 to be the winning model. In contrast, M5 turned out to be the worst fitting model. What we find instead is that the DV Type factor enters into an interaction with the Conditional factor, and that participants attribute somewhat different belief states depending on whether the conditional is embedded under an operator. Negation increases the probability of attributing disbelief; Possible and Question increase the probability of attributing neutrality. The results thus speak against the presupposition hypothesis.

That these effects were found most strongly with projection past the negation operator is not surprising, since embedding under a possibility modal and an interrogative has the same valence as the bare assertion case, when the latter expresses neutrality. But in fact it was found that both the possibility modal and the interrogative contributed to attenuating the expression of doxastic neutrality.

As Experiment 1 shows, presupposition as defined by the classic family of sentences test is not a flexible enough phenomenon to handle the different stances towards the antecedent. This finding naturally prompts us to investigate a more flexible phenomenon: conversational implicature.

4.3 Experiment 2: Conversational Implicatures

Conversational implicatures are the paradigm case of natural-language pragmatics. They arise when a speaker implicitly and intentionally communicates something other than the conventional meaning of the utterance.

Take the following example: "I ate most of the pizza" (Birner, 2013, p.45). The speaker literally states only that they ate most of the pizza but appears to convey – to conversationally implicate – that they did not eat all of it. Implicatures, it is said, arise because of how we expect conversations to go: we expect speakers to behave cooperatively. The classical account, here, is Grice (1989): we expect speakers to say enough, but not too much; to avoid saying false or un-evidenced things; to be relevant; to avoid obscurity and ambiguity, and to be brief and orderly. Implicatures can arise when these expectations are observed or flouted – ostentatiously *not* observed. Let us assume that the speaker is cooperative and, in particular, has said enough, but not too much (has respected the *Maxim of Quantity*). Our cooperative speaker did not make the stronger statement “I ate all of the pizza”, and so – we presume - does not believe that the stronger statement is true. As hearers, we therefore conclude that the speaker did not eat all of the pizza.

Different theories account for implicatures with different theoretical constructs (see, e.g., Horn, 1984; Levinson, 2000; Sperber & Wilson, 1995), but a central property is that implicatures are defeasible: they can be cancelled without producing a contradiction (Blome-Tillmann, 2003). Hence, the speaker above could legitimately say “I ate most of the pizza – in fact, all of it.” That implicatures are so cancellable makes them an attractive option for explaining the different stances conveyed by indicatives and subjunctives. For indicative conditionals, some have proposed that it is an implicature that conveys the "open possibility" sense of the antecedent (Mittwoch, Huddleston et al., 2002), a proposal that obviates the need for distinct presupposition triggers for each stance on the antecedent. More commonly, researchers have proposed that it is an implicature that conveys the "not known" sense of the antecedent (e.g., Grice, 1989; Mittwoch, Huddleston, & Collins, 2002). After all, if the speaker of "If *A*, *C*" had known that both "*A*" and "*C*" were true, they could have said simply "*A* and *C*"; that the speaker did not do so suggests that they do not know (Grice, 1989).

For subjunctives, the implicature account plays an important role. On this account, speakers can use subjunctive conditionals to conversationally implicate, in context, that the antecedent is false. With this account, we can accept, for instance, that example (6) – "If he were rich, he would be smart" – can sometimes, perhaps often, suggest that the "he" in question is not rich (or smart), but the sentence need not give rise to this implicature. Implicature-based accounts differ in detail, but have attracted numerous supporters (e.g., Iatridou, 2000; Ippolito, 2003; Leahy, 2011, 2018; Mittwoch, Huddleston et al., 2002).

The cancellability of conservational implicatures offers a diagnostic test: if information is conveyed by a conversational implicature, then it should be cancellable. Skovgaard-Olsen, Collins et al. (2019) designed a cancellation task that applied this diagnostic test. In this cancellation task, the candidate for being an implicature is uttered by a fictional character. For the current research question, a character, Samuel, might say:

> Samuel: “If there had been a pendant lamp in the bedroom, then it would have hung above the bed.”

Samuel then attempts to cancel the potential implicature: that there is not, in fact, a pendant lamp in the bedroom. A second character, Louis, accuses Samuel of contradicting himself, and participants are asked whether they agree with Louis. If this information is an actual implicature, then it should be possible for Samuel to cancel it: participants should disagree with Louis.

Alongside the candidate implicature are two baselines. The first baseline is an uncontroversial implicature: Samuel might say that it is “possible” that there is such a lamp, but deny suggesting that it is not highly likely. This baseline is an instance of a modal scalar implicature: when a speaker uses a weaker modal term, “possible”, they may implicate, or be mistaken for implicating, that a stronger modal term would be inappropriate. Hence, the speaker here would be suggesting that it is possible but not highly likely that there is such a lamp. Scalar implicatures are readily cancellable. The second baseline is an entailment: Samuel states that “this is a picture of a bedroom AND …” before going on to deny suggesting that it is a picture of bedroom. This should not be cancellable.

The cancellation task allows us to ask whether cancelling the stance towards the antecedent is more like cancelling a scalar implicature or cancelling an entailment. It therefore allows us to experimentally test whether indicatives and subjunctives convey their stances towards their antecedents with a conversational implicature.

4.3.1 Method

Participants

The same sampling procedure and exclusion criteria were used as in Experiment 1. A total of 120 people completed the experiment. Since some of the exclusion criteria were overlapping, the final sample consisted of 93 participants. Mean age was 34.46 years, ranging from 19 to 68. 50.54% of participants identified as female; 48.39% identified as male; and .11% preferred not to respond. 65.59 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.[54]

Design

[54] We are here ignoring the entry ‘2’ for the age of one of the participants.

The experiment had a within-subject design with three factors: Occlusion (with two levels: occluded vs. not-occluded), Prior (with two levels: high (H) vs. low (L)) and Cancellation Type (with three levels: scalar vs. entailment vs. belief-state). To allow for four trial replications for each cell of the design, each participant in total went through 48 within-subject conditions.

Materials and Procedure

The experiment was split into 16 blocks of three pages, one block for each level of the Occlusion × Prior factors and their four trial replications. Each block contained one page for each of the three levels of the Cancellation Type factor. 16 different pictures were randomly assigned to each of the 16 blocks. The order of the blocks was randomized anew for each participant and there were no breaks between the blocks. The three pages within each block were randomized and showed one within-subject condition from the pool of 16 selected pictures with different types of cancellations.

We cued participants to the intended interpretation of the cancellations with instructions and practice trials. For Experiment 2, the participants were given the following instructions together with four sample items:

> In the following you will see several pictures of familiar settings (e.g., bathrooms, kitchens). As you will notice, different parts of the pictures are hidden by grey boxes. Note that some of these boxes are transparent.
> The responses we will ask you to make relate to a picture shown and a corresponding dialogue between Samuel and Louis. In the dialogues, Samuel will say what he thinks is true – what he believes. Sometimes he will indicate what he thinks is false – what he disbelieves. And sometimes he will indicate that he doesn’t have a view – that he is open to either believing or disbelieving it. Louis in turn accuses Samuel of contradicting himself. It will be your task to evaluate Louis' objection. Is he right?

The task of the participants was to indicate the extent to which they agreed or disagreed with Louis' statement on a five-point Likert scale {strongly disagree, disagree, neutral, agree, strongly agree}. Before beginning the experiment proper, participants moreover saw three practice trials, where we emphasized that it was important to pay attention to both subtle differences between the wordings of the various types of cancellations used in the experiment and the varying placement of the grey boxes.

On the following three pages, participants were presented with one of the three types of cancellation in random order (perceived contradiction of cancellation of entailment, of scalar implicature, and of belief state assumptions). The task of the participants was always to assess the extent to which they agreed with Louis' claim that Samuel contradicted himself. Using the bedroom picture from Table 3, the three types of cancellation were implemented across the four conditions as shown in Table 5.

Table 5. Cancellation Types in Experiment 2

Entailment	Scalar Implicature	Belief State
Indicative, occluded H		
Samuel: This is a picture of a bedroom AND IF there is a pendant lamp in the bedroom, THEN it hangs above the bed ...but I am not suggesting that this is a picture of a bedroom.	**Samuel:** This is a picture of a bedroom AND IF there is a pendant lamp in the bedroom, THEN it is possible that it hangs above the bed ...but I am not suggesting that if so, it isn’t highly likely that it hangs above the bed.	**Samuel:** This is a picture of a bedroom AND IF there is a pendant lamp in the bedroom, THEN it hangs above the bed ...but I am not suggesting that I am open to believing or disbelieving that there is a pendant lamp in the bedroom.
Indicative, occluded L		
Samuel: This is a picture of a bedroom AND IF there is a surfboard in the bedroom, THEN it stands against the wall ...but I am not suggesting that this is a picture of a bedroom.	**Samuel:** This is a picture of a bedroom AND IF there is a surfboard in the bedroom, THEN it is possible that it stands against the wall ...but I am not suggesting that if so, it isn’t highly likely that it stands against the wall.	**Samuel:** This is a picture of a bedroom AND IF there is a surfboard in the bedroom, THEN it stands against the wall ...but I am not suggesting that I am open to believing or disbelieving that there is a surfboard in the bedroom.
Subjunctive, not occluded H		
Samuel: This is a picture of a bedroom AND IF there had been a pendant lamp in the bedroom, THEN it would have hung above the bed, where indeed something is hanging ...but I am not suggesting that this is a picture of a bedroom.	**Samuel:** This is a picture of a bedroom AND IF there had been a pendant lamp in the bedroom, THEN it is possible it would have hung above the bed, where indeed something is hanging ...but I am not suggesting that if so, it isn’t highly likely that it would have hung above the bed.	**Samuel:** This is a picture of a bedroom AND IF there had been a pendant lamp in the bedroom, THEN it would have hung above the bed, where indeed something is hanging ...but I am not suggesting that I doubt that there is a pendant lamp in the bedroom.
Subjunctive, not occluded L		
Samuel: This is a picture of a bedroom AND IF there had been a surfboard in the bedroom, THEN it would have stood against the wall ...but I am not suggesting that this is a picture of a bedroom.	**Samuel:** This is a picture of a bedroom AND IF there had been a surfboard in the bedroom, THEN it is possible it would have stood against the wall ...but I am not suggesting that if so, it isn’t highly likely that it would have stood against the wall.	**Samuel:** This is a picture of a bedroom AND IF there had been a surfboard in the bedroom, THEN it would have stood against the wall ...but I am not suggesting that I doubt that there is a surfboard in the bedroom.

Note. For the entailments, the conclusion of And Elimination was cancelled. ‘H’ = high prior probability. ‘L’ = low prior probability.

The goal of the experiment was to find out whether cancellations of assumptions concerning belief states of indicative and subjunctive conditionals are more like cancellations of entailments or cancellations of scalar implicatures.

4.3.2 Results

Some initial descriptive statistics are reported in Table 6.

Table 6. Descriptive Statistics

	Entailment	**Belief-State**	**Scalar Implicature**
Indicative H	*Mdn* = 5, *MAD* = 0	*Mdn* = 2, *MAD* = 1.48	*Mdn* = 3, *MAD* = 1.48
Indicative L	*Mdn* = 5, *MAD* = 0	*Mdn* = 2, *MAD* = 1.48	*Mdn* = 3, *MAD* = 1.48
Subjunctive H	*Mdn* = 5, *MAD* = 0	*Mdn* = 2, *MAD* = 1.48	*Mdn* = 3, *MAD* = 1.48
Subjunctive L	*Mdn* = 5, *MAD* = 0	*Mdn* = 2, *MAD* = 1.48	*Mdn* = 3, *MAD* = 1.48

Note. Due to the ordinal nature of the perceived contradiction ratings, the descriptive statistics are reported via medians (*Mdn*) and median absolute deviations (*MAD*).

In the analysis below, we have collapsed across the levels of the Priors factor to focus on the contrast between indicative conditionals (investigated in the occluded conditions) and subjunctive conditionals (investigated in the not-occluded conditions), which is the contrast of most direct importance.

Given the design, there were replicates for each participant and pictures. Hence, it was not appropriate to assume that the data were independently and identically distributed. Accordingly, the appropriate analysis was to use linear mixed-effects models, with crossed random effects for intercepts and slopes by participants and by pictures (Baayen, Davidson, et al., 2008). This analysis was conducted using R-package `brms` for mixed-effects models in Bayesian statistics (Bürkner, 2017). The following family of nested models was fit to the data:

(M6) a maximal model that treats participants' ratings of perceived contradiction as a function of the Cancellation factor (scalar implicature vs. entailment vs. belief state), Sentence Type (subjunctive vs. indicative), and their interaction.

(M7) a model that is obtained from the maximal model (M6) by removing the two-way interaction.

(M8) a model that is obtained from (M7) by removing the simple effect for the Sentence factor.

(M9) a model that is obtained from (M8) by removing the simple effect for the Cancellation factor.

Effects of the Cancellation Type factor are of theoretical importance for testing the conversational implicature hypothesis. In selecting the class of models above, we investigated whether the effects of the Cancellation Type factor varies across indicative and subjunctive conditionals. Since the responses obtained from the five-point Likert scale are ordinal responses, the responses were modelled as generated by thresholds set on a latent continuous scale via a cumulative model and a logit link function (Bürkner & Vuorre, 2019). Table 7 reports the performance of the models as quantified by the leave-one-out cross validation criterion and WAIC.

Table 7. Model Comparison

	LOOIC	Δelpd	SE	WAIC	Weight
M6	9271.5	0	--	9265.5	0.43
M7	9273.1	-0.8	1.9	9267.3	0.19
M8	9271.7	-0.1	1.9	9265.9	0.39
M9	9296.0	-12.3	2.6	9288.8	0.00

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOOIC. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy.

The modest differences between M6-M8 indicate that the difference between indicative and subjunctive conditionals did not matter much for participants' perceived degree of contradiction. In contrast, the clear rejection of M9 indicates that strong differences in the type of Cancellation were found. Since the differences between M6-M8 were small, Figure 2 plots the posterior predictions of all three models as weighted by their respective model weights from Table 7. Note that, as M8 excludes the interaction and the simple effect of Sentence type, the plot collapses across the Sentence factor. For purposes of plotting, we here aggregate "Disagree strongly"/"Disagree" and "Agree strongly"/"Agree", although these response options were fitted separately above.

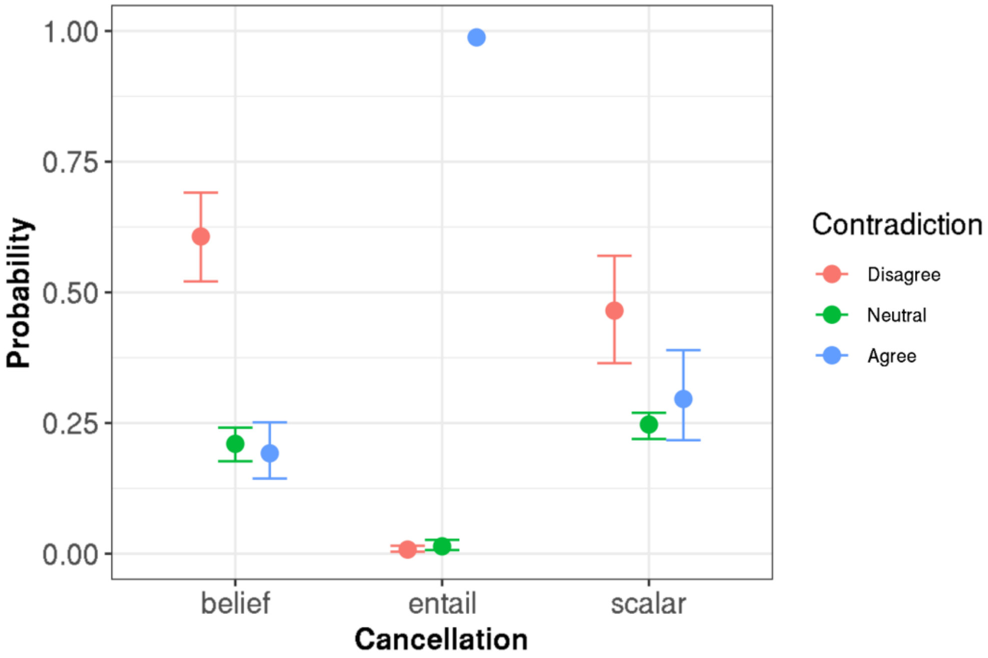


Figure 2. Posterior Predictions of M6-M8. Level of (dis)agreement that Samuel was contradicting himself, split by type of cancellation (of the belief state, entailment, and scalar implicature). This figure collapses across the levels of the Sentence factor. For each of the three types of cancellation, the "Strongly agree" and "Agree" ordinal categories were aggregated to "Agree" and "Strongly disagree" and "Disagree" were aggregated to "Disagree". Error-bars represent 95% credible intervals.

As a manipulation check, it can be observed across sentences that participants clearly distinguished between attempts to cancel a commitment to entailments and conversational implicatures, for both indicative and subjunctive conditionals. It was thus found that cancellations of entailments were viewed as more contradictory than cancellations of scalar implicatures for both indicatives ($b_{\text{Entail - Scalar}}$ = 4.59, 90% CI [3.93, 5.30], BF_{H1H0} > 100) and subjunctives ($b_{\text{Entail - Scalar}}$ = 5.01, 90% CI [4.26, 5.77], BF_{H1H0} > 100).

Next, the cancellation of belief states were compared to these two baselines. Strong evidence was found that cancellations of belief states were viewed as less contradictory than cancellations of entailments for both indicatives ($b_{\text{Belief - Entail}}$ = -5.29, 90% CI [-5.99, -4.58], BF_{H1H0} > 100) and subjunctives ($b_{\text{Belief - Entail}}$ = -5.30, 90% CI [-6.08, -4.53], BF_{H1H0} > 100).

In addition, moderate evidence was found that cancellations of belief states were viewed as less contradictory than cancellations of scalar implicatures for indicatives ($b_{\text{Belief - Scalar}}$ = -.70, 90% CI [-1.07, -.30], BF_{H1H0} = 9.07) but not for subjunctives ($b_{\text{Belief - Scalar}}$ = -.28, 90% CI [-.65, .08], BF_{H1H0} = .18), where indeed the H_0 of no difference between the cancellation of belief state assumptions and scalar implicatures was supported.

4.3.3 Discussion

The analysis validated our two baselines for the cancellation test by showing that there was very strong evidence that commitments to entailments were viewed as more cancellable than commitments to scalar implicatures. Next, our results showed that speakers can cancel, without contradicting themselves, the neutrality towards the antecedent of an indicative conditional and the disbelief towards the antecedent of a subjunctive conditional. Indeed, cancelling a commitment to the suggested belief state was viewed as *less* contradictory than cancelling a commitment to a scalar implicature for indicative conditionals. For subjunctive conditionals, strong evidence was found that the belief state assumptions concerning the antecedent was just as cancellable as scalar implicatures. The data thus supports the view that a conversational implicature is present in both indicative and subjunctive conditionals. Differences in the content of these conversational implicatures may accordingly help account for the meaning differences between indicative and subjunctive conditionals. Converging evidence for this conclusion was found in Experiment 1, where the posterior probability of selecting 'Belief' was increased from subjunctives used to convey counterfactual conditionals to subjunctives used as Anderson conditionals.

4.4 General Discussion

It is a familiar point that indicative and subjunctive conditionals differ with respect to the belief-state status of the antecedent, illustrated by Adams' (1970) Oswald-Kennedy pair, where one can consistently accept the first while rejecting the second:

(*indicative*) If Oswald did not shoot Kennedy, someone else did.
(*counterfactual*) If Oswald had not shot Kennedy, someone else would have.

The formulation of this minimal pair, with two conditionals differing in meaning, has led to a number of attempts to either provide a unifying account of indicative and subjunctive conditionals (Stalnaker, 1975; Edgington, 2008; von Fintel, 2012; Spohn, 2013; Starr, 2014; Williamson, 2020), argue why disjunct accounts are needed (Lewis, 1973, 1976; Bennett, 2003), or argue for a unifying account by questioning that this indeed constitutes a minimal pair (Quelhas et al., 2018). For proponents of the first approach, it is tempting to formulate one semantics of conditionals and look to linguistic phenomena closer to pragmatics, like conversational implicatures or presuppositions, to account for the meaning differences between the two types of sentences above. Our findings cast light on the plausibility of such an approach.

Conversational Implicatures and Presuppositions

Throughout Experiments 1 and 2, it was found that a conversational implicature best accounts for the diverging belief state assumptions concerning the antecedents of indicative and subjunctive conditionals. Central to the evidence is the cancellability of the belief states: speakers could cancel the neutrality towards the antecedent in indicatives and the disbelief towards the antecedent in subjunctives without participants perceiving a contradiction.

According to the Stalnaker-Karttunen-Heim approach to presuppositions, a sentence carrying a presupposition can only be felicitously uttered in contexts that entail the presupposition (Kadmon, 2001, Ch. 5), or which can be updated so as to entail the presupposition (Simons, 2006). On this view, cancellation of presuppositions cannot be accounted for, if presuppositions are supposed to be entailed by the context on a classical, monotonic consequence relation. In contrast, on the so-called Cancellation Approach of Gazdar (1979) and Soames (1982), presuppositions are defeasible and can be cancelled by contextual assumptions or prior conversational implicatures (Kadmon, 2001, Ch. 6).

However, as Beaver and Geurts (2014) note, it appears that the main examples of cancellation of presuppositions concern cases, where the sentence carrying the presupposition has been embedded in a compound sentence. For instance, in examples like “If it’s *the* knave that stole the tarts, then I’m a Dutchman: there is no knave here”, the presupposition of the embedded sentence that there is a knave is cancelled. In contrast, cancelling unembedded presuppositions is typically seen to be as infelicitous as cancelling a commitment to an entailment (e.g., “It’s the knave that stole the tarts, but there is no knave”). Based on this observation, Beaver and Geurts (2014) formulate the following generalization:

Table 8. Predictions

	Entailments	Presuppositions	Conversational implicatures
Project from embeddings	0	1	0
Cancellable when embedded	--	1	--
Cancellable when unembedded	0	0	1

Note. The horizontal lines indicate that Beaver and Geurts (2014) do not provide values for those cells.

This generalization fits with the further observation that, mostly, the presuppositions of unembedded affirmative statements are entailments (Simons, 2006). Accordingly, the presuppositions of unembedded affirmative statements should not be cancellable without contradiction. These observations about cancellation pose a challenge to the view that presupposition gives rise to the differing stances towards the antecedents conveyed by indicative and subjunctive conditionals, inasmuch as only further embeddings of the conditionals should permit cancellation. Yet, the results from Experiment 2 show that the stances towards the antecedent were cancellable for both indicatives and subjunctives, and even more cancellable than a commitment to scalar implicatures.

The finding of this cancellation effect thus provides support for a conversational implicature account (Iatridou, 2000; Leahy, 2011, 2018) over a presupposition (Kutschera, 1974; Stalnaker, 1975, 2014; von Fintel, 1997) or entailment account. This rejection of a presupposition account is further strengthened by our results in Experiment 1, where it was found that the belief-state assumptions concerning the antecedents of indicative and subjunctive conditionals do not project through embedding under various operators.

The Source of the Conversational Implicatures

A challenge for a conversational implicature account is that it must be shown *in principle*[55] how the conversational implicature to the falsity of antecedent of subjunctive conditionals could be reconstructed based on general maxims of communication (Grice, 1989). In Leahy (2018), this conversational implicature is accounted for by applying the notion of scalar implicatures to the presuppositions of a sentence. Leahy further holds that the presuppositions of counterfactuals (Ø) is logically weaker than the presuppositions of indicative conditionals (i.e. that the antecedent is epistemically possible). These constraints generate the expectation that the choice of the subjunctive means that the speaker was not warranted in making the stronger presuppositions of the corresponding indicative conditional. One difficulty with this view is, however, that, our data suggest that it is not, in fact, a presupposition of indicative conditionals that the antecedent is epistemically possible. In addition, participants considered the belief-state assumption of the antecedent to be more cancellable than scalar implicatures for both indicatives and subjunctives in Experiment 2.

Another possibility runs as follows: in the choice of a conditional construction ("if *A*, then *C*") over a conjunction ("*A* & *C*"), the speaker signals that they are not warranted in making the stronger assertion of committing to the truth of *A*. Rather, by making a conditional assertion, the speaker can express their view about a relationship between *C* and *A* while remaining uncommitted about *A*. By further choosing the subjunctive mood (e.g., 'if [past tense], would …'), where past tense morphology is employed which does not have a literal past tense interpretation (Iatridou, 2000; Ippolito, 2003), further distance is expressed. If

[55] Note that the circumstance that rational reconstructions in terms of abductive reasoning like this can be carried out does not mean that they play a role for the underlying psychological processes, or that they could not have become conventionalized in time (for discussion see Geurts, Kissine, & van Tiel, 2020).

interpreted doxastically, there are only three possibilities for categorical beliefs: either the speaker *believes A,* the speaker is *neutral* about *A,* or the speaker *disbelieves A.* If the speaker had been in a position to believe *A,* a conjunction could have been used. Instead, the speaker chose a conditional construction. If the speaker wished to remain neutral about *A,* a conditional in the indicative mood could have been used. Instead, the speaker chose a more convoluted formulation employing fake past tense to express further distance. Given that the speaker does not believe *A*, and is not content with remaining neutral about *A*, their interlocutors are warranted in inferring that the speaker disbelieves, or doubts, *A*.

Anderson Conditionals, Modus Tollens, and Presuppositions

In Anderson conditionals, the speaker complicates the interpretational task of his/her interlocutors even further. The speaker does this by combining a conditional construction with past tense morphology that is not to be taken literally ("If Jones had taken arsenic, he would have shown exactly those symptoms…") with a factive relative clause ("…which he does in fact show"), which cancels the doxastic distance introduced by the subjunctive mood. Here again the hearer is faced with the challenge of figuring out why a cooperative speaker would use such a convoluted way of expressing him-/herself. If participants invest sufficient resources, they could generate the hypothesis that the speaker is using this complex construction as part of an argument that purports to dispel doubt about the antecedent. In the absence of alternative explanations for the patients' symptoms, this sub-argument could in turn be used as part of a larger argument to establish the truth of the antecedent, via an inference to the best explanation along the following lines:

> 'I think the patient took arsenic; for he has such-and-such symptoms; and these are the symptoms he would have if he had taken arsenic' (Edgington, 2008, p. 6)

In Zakkou (2019), it is argued that, contrary to appearances, Anderson conditionals do not provide a counterexample against a presupposition account. As part of her argument, Zakkou points out that a speaker, who first asserts 7a) and then 7b) need not contradict herself:

> 7a) "If Jones had taken arsenic, he would have shown the same symptoms he actually shows", 7b) "So he took arsenic"

The contradiction attributed to the presupposition account is removed, it is argued, if the speaker only accepts that Jones did not take arsenic for the purpose of the conversation in asserting 7a) and accepts that Jones did take arsenic, because she believes that he did, in asserting 7b). While this is certainly possible, it still needs to be established empirically that ordinary speakers are just as sophisticated in keeping track of different attitudes. The simpler explanation is that the speaker is cancelling a conversational implicature.

Similarly, Zakkou (2019) suggests that the speaker in (16) accepts for the purpose of conversation that Jones did not take arsenic and asserts his own belief to the contrary via a relative clause:

(16) If Jones had taken arsenic—which he did—he would have shown the same symptoms he actually shows.

A more straightforward account would be that the speaker cancels a commitment to the conversational implicature that Jones did not take arsenic through the relative clause.

In both cases, further empirical work is needed to distinguish between these possibilities. But it is worth highlighting that while it was found that participants have the same posterior probability of attributing belief and disbelief to the antecedent of an Anderson conditional in Experiment 1, negating an Anderson conditional shifts the modal tendency towards disbelief. So, it was not found that the belief state assumption concerning the antecedent of Anderson conditionals exhibit the standard behaviour of presuppositions.

Zakkou (2019) also dismisses an argument against the presupposition account based on Stalnaker's (1975, 2014) observation that the following modus tollens argument does not beg the question and presuppose what it is supposed to establish (i.e., the butler's innocence):

(8) "If the butler had done it, we would have found blood on the knife. The kitchen knife was clean; therefore the butler did not do it".

To make the case, Zakkou considers related examples in which the speaker may use presuppositions in the technical sense and anticipate the conclusion of a modus tollens argument, without begging the question by introducing the conclusion as a tacit premise. The discussion overlooks, however, that on a presuppositional account, the first premise of the modus tollens argument can only be true, if its presuppositions are satisfied; otherwise this premise is false or a truth-value gap (von Fintel, 2004). So, to have an argument with true premises, it is a requirement of an account that makes the falsity of the antecedent a presupposition of a subjunctive conditional that the conclusion is already true with the first premise, which is indeed question-begging.

In contrast, a conversational implicature account would fare better. For conversational implicatures are only plausible inferences about the speaker's mental states that the interlocutor is defeasibly warranted in making. This allows for the factual premises of the argument to be true irrespectively of the status of these inferences. Through the entailment, the modus tollens argument ensures that the premises cannot be true without the conclusion being true. So, whereas an uncancelled conversational implicature of the first premise at most establishes that it is reasonable for the interlocutor to assume that the speaker believes that the butler is innocent, the conclusion of the modus tollens argument shows that the butler *must* be innocent. The conversational implicature account, in other words, separates the truth and factual content of the premises from the conversational assumptions about the speaker's belief states and thereby avoids begging the question about the factual truth of the conclusion.

Mental Models Theory

Finally, we turn to the implications of our findings for Mental Models Theory (MMT). On the current revised version of MMT (Khemlani et al., 2018), the meaning of conditionals is explicated by Table 9:

Table 9. Mapping between indicative and counterfactuals, MMT

Row	**Partition**		**Factual:**	**Counterfactual:**
			If A then C	*If A had happened, then C would have happened*
1	A	C	Possibility	Counterfactual possibility
2	A	Not-C	Impossibility	Impossibility
3	Not-A	C	Possibility	Counterfactual possibility
4	Not-A	Not-C	Possibility	Fact

Note. Quelhas et al. (2018) call indicative conditionals "factual conditionals".

Conditionals are here interpreted as conjunctive assertions about possibilities (i. e. "*A*&*C* is possible and *A*&¬C is not possible..."). That not-*A* is possible is a shared presupposition of true and false conditionals; what matters for their truth evaluation is just that the first two rows get switched. In the case of counterfactual conditionals, the "¬*A*&¬C" possibility acquires the status of being *a fact* and the other possibilities change status to express "counterfactual possibilities", which were once possible but did not obtain. That the "¬*A*&¬C" possibility is a fact is rendered a presupposition when proponents of mental model speak of "the presupposed facts" (see, e.g., Byrne, 2005, 2016, 2017; Espino & Byrne, 2018).

If MMT adheres to a classical definition of presupposition (as suggested in Ragni & Johnson-Laird, 2020), we take the theory to hold that the presuppositions project under various operators and are not cancellable as long as the conditionals are unembedded. On this understanding, the theory therefore stands in tension with our findings, which suggest that the stances towards the antecedent do not project and are cancellable.

4.5 Conclusion

In this paper, we present new experimental evidence on the doxastic status of subjunctive conditionals. Previous theoretical papers in linguistics (e.g., Iatridou, 2000; Ippolito, 2003) have discussed the possibility of conversational implicature and presupposition accounts of the assumed falsity of subjunctive conditionals, but without presenting empirical data that could help decide the issue. To this end, we developed new stimulus materials to selectively manipulate the belief states of participants when evaluating indicative and subjunctive conditionals and probed the conversational implicature account and the presupposition account across two experiments. As part of these studies, we additionally investigated how participants assess so-called Anderson conditionals, where the falsity of the antecedent is bracketed in subjunctive conditionals. It was found in a family of sentences test that operators like negation, possibility modals, and interrogatives have an effect on participants' belief-state assumptions and that a presupposition hypothesis predicting that belief-state assumptions project past such operators could be rejected. Further, it was found in a cancellation task, that belief-state assumptions of indicative conditionals and subjunctive conditionals were either just as cancellable as scalar implicatures (subjunctive conditionals) or even more cancellable than scalar implicatures (indicative conditionals). This finding indicates that one of the central meaning differences between indicative and subjunctive conditionals can be attributed to a phenomenon which is uncontroversially pragmatic in nature; to wit, conversational implicatures.

References

Adams, E. (1970). Subjunctive and Indicative Conditionals. *Foundations of Language, 6*, 89-94.

Anderson, A. R. (1951). A note on subjunctive and counterfactual conditionals. *Analysis*, *12,* 35–38.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 340-412.

Baratgin, J.**,** Over, D. E., & Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking & Reasoning, 19,* 308-328.

Barrouillet, P., & Lecas, J.-F. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, *67*(3), 209–253. https://doi.org/10.1016/S0010-0277(98)00037-7

Beaver, D. I., & Geurts, B. (2014). Presupposition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. (Winter 2014 Edition). Retrieved from https://plato.stanford.edu/archives/win2014/entries/presupposition/

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Birner, B. J. (2013). *Introduction to Pragmatics*. Hoboken, N.J.: John Wiley & Sons.

Blome-Tillmann, M. (2013). Conversational Implicatures (and How to Spot Them). *Philosophy Compass*, *8*(2), 170–185. https://doi.org/10.1111/phc3.12003

Bradley, R. (2012). Multidimension Possible-world Semantics for Conditionals. *The Philosophical Review*, *121*(4), 539-71.

Bringsjord, S., & Govindarajulu, N. S. (2020). Rectifying the mischaracterization of logic by mental model theorists. *Cognitive Science*, *44*.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P., & Vuorre, M. (2019). Ordinal Regression Models in Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*. March 2019:77-101. doi:10.1177/2515245918823199

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press.

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Byrne, R. M. J. (2017). Counterfactual Thinking: From Logic to Morality. *Current Directions in Psychological Science*, *26*(4), 314–322. https://doi.org/10.1177/0963721417695617

Byrne, R. M. J., and Johnson-Laird, P. N. (2019). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. http://dx.doi.org/10.1037/xlm0000756

Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, *27*(4), 726–740. https://doi.org/10.3758/BF03211565

Chemla, E,, & Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes*, *38*, 241–260.

Declerck, R., & Reed, S. (2001). *Conditionals: A Comprehensive Empirical Analysis*. Berlin: Walter de Gruyter.

de Vega, M., Urrutia, M., & Riffo, B. (2007). Canceling updating in the comprehension of counterfactuals embedded in narratives. *Memory & Cognition*, *35*(6), 1410–1421. https://doi.org/10.3758/BF03193611

Edgington, D. (1995). On conditionals. *Mind*, *104*, 235-329.

Edgington, D. (2006). Conditionals. In: E.N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2008 edn.). Retrived from: http://plato.stanford.edu/archives/ win2008/entries/conditionals/.

Edgington, D. (2008). I-Counterfactuals. *Proceedings of the Aristotelian Society*, *108*(1), 1–21.

Espino, O., & Byrne, R. M. J. (2018). Thinking About the Opposite of What Is Said: Counterfactual Conditionals and Symbolic or Alternate Simulations of Negation. *Cognitive Science, 42*(8), 2459–2501. https://doi.org/10.1111/cogs.12677

Evans, J. St. B. T. & Over, D. (2004). *If*. Oxford: Oxford University Press.

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*(3), 609–626.

Fillenbaum, S. (1974). Information amplified: Memory for counterfactual conditionals. *Journal of Experimental Psychology*, *102*(1), 44–49. https://doi.org/10.1037/h0035693

Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How People Interpret Conditionals: Shifts Toward the Conditional Event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 635-648.

Gazdar, G. (1979). *Pragmatics: Implicature, Presuppositions, and Logical Form.* New York: Academic Press.

Geurts, B., Kissine, M., and van Tiel, B. (2020). Pragmatic reasoning in autism. In: Morsanyi, K. and Byrne, R. (Eds.), *Thinking, reasoning and decision making in autism* (pp. 113-134). London: Routledge.

Goodwin, G. P., and Johnson-Laird, P. N. (2018). The Truth of Conditional Assertions. *Cognitive Science*, *42*, 2502-2533.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Horn, L. R. (1984). Towards a New Taxonomy for Pragmatic Inference: Q-based and R-based Implicature. In: Schiffrin, D. (Eds), *Georgetown University Round Table on Languages and Linguistics 1984 (pp. 11-42).* Washington, DC: Georgetown University Press.

Huang, Y. (2007). *Pragmatics*. Oxford, England: Oxford University Press.

Huddleston, R. (2002). Content clauses and reported speech. In R. Huddleston & G. K. Pullum (Eds.), *The Cambridge Grammar of the English Language* (pp. 947–1030). Cambridge, England: Cambridge University Press.

Iatridou, S. (2000). The Grammatical Ingredients of Counterfactuality. *Linguistic Inquiry*, *31*(2), 231–270. https://doi.org/10.1162/002438900554352

Ippolito, M. (2003). Presuppositions and Implicatures in Counterfactuals. *Natural Language Semantics*, *11*(2), 145–186. https://doi.org/10.1023/A:1024411924818

Ippolito, M. (2016). How Similar Is Similar Enough? *Semantics and Pragmatics, 9*(6), 1–60.

Johnson-Laird, P.N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. and Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review 109*, 646-678.

Kadmon, N. (2001). *Formal Pragmatics*. Malden, MA: Blackwell Publishers.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science*, *42*(6), 1887–1924. https://doi.org/10.1111/cogs.12634

Kutschera, F. (1974). *Indicative Conditionals. Theoretical linguistics*, *1*, 257-269.

Leahy, B. (2011). Presuppositions and Antipresuppositions in Conditionals. *Semantics and Linguistic Theory*, *21*, 257. https://doi.org/10.3765/salt.v21i0.2613

Leahy, B. (2018). Counterfactual antecedent falsity and the epistemic sensitivity of counterfactuals. *Philosophical Studies*, *175*, 45-69.

Lecas, J.-F., & Barrouillet, P. (1999). Understanding conditional rules in childhood and adolescence: A mental models approach. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *18*(3), 363–396.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, England: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*, 297-315.

Mittwoch, A., Huddleston, R., & Collins, P. (2002). The clause: Adjuncts. In R. Huddleston & G. K. Pullum (Eds.), *The Cambridge Grammar of the English Language* (pp. 663–784). Cambridge, England: Cambridge University Press.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Over, D. E., & Baratgin, J. (2017). The "defective" truth table: Its past, present, and future. In N. Galbraith, E. Lucas, & D. E. Over (Eds.), *The Thinking Mind: A Festschrift for Ken Manktelow* (pp. 15-28). Abingdon, UK: Routledge.

Pfeifer, N. & Tulkki, L. (2017). Conditionals, Counterfactuals, and Rational Reasoning: An Experimental Study of Basic Principles. *Minds and Machines, 27*(1), 119-165.

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking and Reasoning, 16*(3), 172–197.

Potts, C. (2007). Into the Conventional-Implicature Dimension. *Philosophy Compass, 2*(4), 665-679.

Potts, C. (2015). Presupposition and implicature. In Shalom Lappin and Chris Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (pp. 168-202), 2nd edn,. Oxford: Wiley-Blackwell.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2018). The Relation Between Factual and Counterfactual Conditionals. *Cognitive Science*, *42*(7), 2205–2228. https://doi.org/10.1111/cogs.12663

Ragni, M., and Johnson-Laird, P. (2020). Reasoning about epistemic possibilities. *Acta Psychologica*, 208, 103081.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reips, U. D. (2002). *Standards for Internet-based experimenting. Experimental Psychology, 49* (4), *243*-256.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics*, *5*(6), 1-69.

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In: Schwarz (Ed.). *Experimental perspectives on presuppositions. Studies in Theoretical Psycholinguistics* (Vol. 45, pp. 215-240). Springer, Cham.

Santamaría, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and Semifactual Conditionals Prime Alternative Possibilities. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *31*(5), 1149–1154. https://doi.org/10.1037/0278-7393.31.5.1149

Schroyens, W. (2010). A meta-analytic review of thinking about what is true, possible, and irrelevant in reasoning from or reasoning about conditional propositions. *European Journal of Cognitive Psychology,* 22 (6), 897-921.

Simons, M. (2006). Foundational Issues in Presupposition. *Philosophy Compass, 1*(4), 357-372.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., and Klauer, K. C. (2019). Cancellation, Negation, and Rejection. *Cognitive Psychology, 108*, 42-71.

Smith, E. A. and Hall, K-C. 2011. Projection diversity: Experimental evidence. Workshop on Projective Meaning at ESLLI 2011.

Soames, S. (1982). How presuppositions are inherited: a solution to the Projection Problem. *Linguistic Inquiry, 13*, 483-545.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition* (Second Edition). Oxford, England: Blackwell.

Stalnaker, R. C. (1968). A Theory of Conditionals. In: Rescher, N. (Eds.), *Studies in Logical Theory (pp. 98-112)*. Oxford: Basil Blackwell.

Stalnaker, R. (1972). Pragmatics. In D. Davidson & G. Harman (Eds.), *Semantics of natural language* (pp. 389–408). Reidel: Dordrecht.

Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz & P. Unger (Eds.), *Semantics and philosophy* (pp. 197–214). New York, NY: New York University Press.

Stalnaker, R. C. (1975). Indicative conditionals. *Philosophia, 5*(3), 269-286.

Stalnaker, R. (2014). *Context*. Oxford: Oxford University Press.

Starr, W. B. (2014). A Uniform Theory of Conditionals. *Journal of Philosophical Logic*, *43*(6), 1019-1064.

Starr, W. (2019). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). Retrieved from forthcoming URL = <https://plato.stanford.edu/archives/fall2019/entries/counterfactuals/>.

Stewart, A. J., Haigh, M., & Kidd, E. (2009). An investigation into the online processing of counterfactual and indicative conditionals. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2113–2125.

Thompson, V. A., & Byrne, R. M. J. (2002). Reasoning counterfactually: Making inferences about things that didn’t happen. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 1154–1170.

von Fintel, K. (1997). The presupposition of subjunctive conditionals. In *MIT working papers in linguistics 25* (pp. 29–44). Cambridge, MA: MIT Working Papers in Linguistics.

von Fintel, K. (2004). Would you believe it? The king of France is back! Presuppositions and truth-value intuitions. In: Reimer, M. and Bezuidenhout, A. (eds.), *Descriptions and Beyond*. Oxford: Oxford University Press, 269–296.

von Fintel, K. (2012). Subjunctive conditionals. In G. Russell & D. Graff Fara (Eds.), *The Routledge companion to the philosophy of language* (pp. 466–477). New York, NY: Routledge.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev*, *25*, 35-57.

Zakou, J. (2019). Presupposing Counterfactuality. *Semantics & Pragmatics, 12*(21), 1-20. https://doi.org/10.3765/sp.12.21.

Chapter 5:
Possible World Truth-Table Task[56]

Niels Skovgaard-Olsen,
Peter Collins,
Karl Christoph Klauer

In this paper, a novel experimental task is developed for testing the highly influential, but experimentally underexplored, possible worlds account of conditionals (Stalnaker, 1968; Lewis, 1973). In Experiment 1, this new task is used to test both indicative and subjunctive conditionals. For indicative conditionals, five competing truth tables are compared, including the previously untested, multi-dimensional possible world semantics of Bradley (2012). In Experiment 2, individual variation in truth assignments of indicative conditionals is investigated via Bayesian mixture models that classify participants as following one of several competing models. As a novelty of this study, it is found that a possible worlds semantics of Lewis and Stalnaker is capable of accounting for participants' aggregate truth value assignments in this task. Applied to indicative conditionals, we show across two experiments, that the theory both captures participants' truth values at the aggregate level (Experiment 1) and that it makes up the largest subgroup in the analysis of individual variation in our experimental paradigm (Experiment 2).

[56] This chapter is under review:

Skovgaard-Olsen, N. Collins, P., and Klauer, K. C. (*in review, Cognition*). Possible World Truth Table Task.

Authors' Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).

Acknowledgement: We would like to thank Eric Raidl and David Over for comments and discussions.

5.1 Introduction

The sentences that follow illustrate a classic distinction between types of conditional:

(1) If Trump won the 2020 election, then the outcome was a fraud.

(2) If Trump had won the 2020 election, then the outcome would have been a fraud.

Sentence (1) is an indicative conditional; sentence (2) is a subjunctive conditional. It is intuitively clear that the conditionals mean quite different things. Suppose that a recipient of the sentences somehow has no knowledge of the outcome of the election. That recipient might well conclude from sentence (2), but likely not from sentence (1), that Trump did not win the 2020 election. These two conditionals have content which signifies a difference in which commitments the speaker adopts towards the outcome of the election. Accordingly, the official view of accepting the outcome of the election is reflected in endorsing the subjunctive over the indicative conditionals, and thus (2) might more reasonably be continued with "…but we all know, after more than 50 lawsuits, that the evidence for fraud did not hold up in court".

These differences in content ripple through further differences in background beliefs as those who accept sentence (2) will likely invoke different explanations for the outcome, and apportion credit and blame for the outcome differently, from those who do not accept it. Clear though such intuitions may be, there is no consensus theory of either indicative or subjunctive conditionals or, indeed, on the relationship between the two types of conditional.

The standard approach is to model the meaning of conditionals in terms of truth conditions: states of affairs that make the conditional true or false. This approach is by no means universal, with some arguing that only subjunctive conditionals have truth conditions (Bennett, 2004), and others that neither indicatives nor subjunctives have truth conditions (Edgington, 2008). But the approach is adopted by several key theories of conditionals, which argue for the truth conditions summarized in Table 1 for indicative conditionals (Evans & Over, 2004; Edgington, 2006; Over & Baratgin, 2017). For a conditional "If A, C", the "A" clause is the antecedent, and the "C" clause is the consequent. While the truth tables all agree in their first two rows, in which the antecedent is true, they differ markedly in the second two rows, in which the antecedent is false.

Table 1. Truth Tables for Indicative Conditionals

A	C	Material Implication	Possible Worlds	de Finetti	Jeffrey Table
T	T	T	T	T	T
T	F	F	F	F	F
F	T	T	T/F	NA	P(C|A)
F	F	T	T/F	NA	P(C|A)

Note. While the material implication makes the truth-value of the conditional a function of the truth values of the antecedent and the consequent, the possible world account of Stalnaker (1968) permits it to be either true or false, depending on whether the consequent is true in the most similar situation in which the antecedent is true. This feature of possible worlds conditionals means that their truth values are not truth-functional in contrast to the material implication, as we will return to in the General Discussion.

Historically, each truth table has had its supporters. The first of these truth tables, that of the material conditional, has lost favour in recent years and is rarely applied to counterfactuals (though see Williamson, 2020 for a recent defense). This loss of favour is in large part due to a commonly reported conflict with experimental data. In the truth table for the material conditional, a conditional is true whenever its antecedent is false. But classic studies suggested that experimental participants disagree with these truth values: that, when participants are asked if false-antecedent rows in the truth table are true, false or irrelevant, a substantial number of participants declare them irrelevant (e.g., Evans & Newstead, 1977; Evans et al., 2007; Johnson-Laird & Tagart, 1969).

This finding of frequent “irrelevant” responses lends support to the de Finetti truth table, on which account a conditional is neither true nor false when the antecedent is false. In such cases, the conditional can be viewed as “void” (for discussion, see Over & Cruz, 2018). That finding has, however, in turn been qualified through a recent meta-analysis (Schroyens, 2010), which reports that “irrelevant” is a minority response and only the modal response for the final row (the FF cases) - and then only with tasks which use implicit negation.

While the de Finetti Table takes the conditional to be void when its antecedent is false, the related Jeffrey Table replaces the “void” (above, “NA”) values with the conditional probability of the consequent given the antecedent, P(C|A). This amendment is interpreted as capturing the fact that one can be more or less confident in a conditional even when its antecedent is false (Over & Cruz, 2018). In our experiments below, we will test both the de Finetti and Jeffrey truth tables and return to the issues they raise in the context of our experimental task.

Last, but not least, is the possible-worlds account. Developed primarily for counterfactual conditionals (Lewis, 1973; Stalnaker, 1968), the account has been applied to both indicative and subjunctive conditionals (Stalnaker, 1968). We will provide a detailed summary of the possible-worlds account below. Here it suffices to note that, on the account of

Stalnaker (1968), when the antecedent is false, a conditional can be either true or false, depending on whether the consequent is true in the most similar situation in which the antecedent is true. In the psychology of reasoning, possible-worlds semantics is sometimes discussed as an important alternative (e.g., Evans & Over, 2004), and sometimes set aside as psychologically implausible (Johnson-Laird & Ragni, 2019). But the truth conditions of possible-worlds semantics are rarely tested.[57] Despite its roughly 50 years of existence, the theory has proved difficult to test empirically, which we hope to rectify in this paper.

It is a matter of much debate, then, which truth table best captures the meaning of conditionals. A related debate is whether indicative and subjunctive conditionals can be explained with a single theory, with theorists both advancing unified theories (e.g., Edgington, 2008; Stalnaker, 1968, 1975; Starr, 2014; Williamson, 2020) and opposing them (e.g., Bennett, 2003; Lewis, 1973, 1976). Within the psychology of reasoning, different camps have developed unified theories, including mental models theorists (Johnson-Laird & Byrne, 2002; Quelhas et al., 2018) and suppositional theorists (Over et al. 2007; Baratgin, Over et al., 2013; Pfeifer & Tulkki, 2017). In Over et al. (2007), this takes the form of interpreting counterfactuals as expressing an "epistemic past tense" (Adams, 1975), which is to be evaluated via conditional probabilities by considering a situation before the antecedent came to be disbelieved. Accordingly, the de Finetti and Jeffrey truth tables that have been applied to indicative conditionals (e.g., in Evans & Over, 2004; Baratgin, Over, & Politzer, 2013; Over & Baratgin, 2017) would also hold for counterfactual conditionals. Other related probabilistic approaches apply principles of causal reasoning to counterfactuals and emphasize that counterfactuals concern interventionalist probabilities when "rerunning history" in a causal model (Sloman & Lagnado, 2005; Rips, 2010; Pearl, 2013; Lassiter, 2017; Skovgaard-Olsen et al., 2021).

In deciding whether a unifying account of the two types of conditionals can be given, one question that deserves central attention is this: do indicative and subjunctive conditionals have the same type of truth conditions? In this paper, we present experiments which throw

[57] For some of the very few attempts to test other aspects of possible world semantics see: Cariani & Rips (2017), Johnson-Laird & Ragni (2019). A further exception is Wijnbergen-Huitink, Elqayam et al. (2015), which applied Lewis/Stalnaker's theory to right-nested conditionals 'if p, then if q, r' in *indicative mood* in a betting context. However, Wijnbergen-Huitink et al. do not discuss the following two complications that Lewis (1973, 1976) only applied this theory to *subjunctive* conditionals, and that Stalnaker (1975) only applies the theory to indicative conditionals under the assumption of the pragmatic principle explained below. A final exception is Douven et al. (2020), which applies one of Lewis's (1979) criteria for similarity (i.e. agreement with particular facts) to Stalnaker's theory of indicative conditionals.

new light on this question by presenting a novel experimental test of the truth conditions of one of the key theories of conditionals: the possible worlds account (Lewis, 1973; Stalnaker, 1968). The experiments allow us to distinguish the possible-worlds account from other key approaches, including the truth tables reviewed above.

Possible Worlds Semantics

Since this paper focuses in large part on testing the possible-worlds account, we provide a more detailed overview of the theory here. Possible-worlds semantics is one of the most widely used semantic frameworks for dealing with subjunctive conditionals in philosophy and linguistics (Stalnaker, 1968; Lewis, 1973; Bennett, 2003; Portner, 2009; Kratzer, 2012). Just like probability theory relies on a set of possible outcomes, possible-worlds semantics relies on a set of possible worlds specifying alternative ways the world could be. Sets of such possible worlds represent propositions, which correspond to events in probability theory. The account states, roughly, that:

> ‘If it had been the case that A, then it would have been the case that C’ (A > C) is true in the possible world, *w*, iff C is true in all the A-worlds that are most similar to *w*.

In Stalnaker (1968), this is explicated via a selection function, f(A, *w*), which selects the closest world (or, alternatively: the set of closest worlds) to *w* in which A is true. The conditional, [A > C], is then true iff the selected A-world(s) is a subset of the set of worlds in which C is true, [C]. For a less formal illustration consider (3) as uttered in 2019:

> (3) “If Hillary had won the 2016-election, the US relationship with the EU would have been better”

(3) is true roughly just in the following case. We consider the known facts in 2019 and (hypothetically) vary the actual circumstances such that Hillary won the 2016-election but keep as much else fixed as possible, such as Hillary’s disposition to govern and the previous US/EU relationship. And, in that hypothetical world, the consequent is true: the US relationship with the EU is better than in it was in 2019.

Pragmatics, indicatives, & subjunctives

While Lewis (1973, 1976) restricted possible-worlds semantics to subjunctives, Stalnaker (1975) extended the semantics to indicatives. In so doing he made use of pragmatic notions to account for the differences in meaning. To explicate the role of pragmatics, Stalnaker (1975) uses the notion of a common ground. The common ground between interlocutors can be understood as a context set of live possibilities that the interlocutors mutually presuppose. Through their assertions, interlocutors try to decide between the members of this *context set*. For instance, suppose that some colleagues are discussing, in hushed tones, recent budget cuts at ‘the Department’. For the conversation to proceed, the interlocutors must mutually presuppose much background information so that it is understood which department they are talking about. Moreover, the interlocutors also mutually presuppose a set of possibilities, which identify different candidates for deciding the latest budget cut.

On Stalnaker’s (1974, 2014) pragmatic notion of presuppositions, a sentence carrying presuppositions can only be felicitously uttered based on one of the following conditions: either its presuppositions are entailed by the common ground, or they are accepted by the hearer (“accommodated”) as an update to the common ground for the purposes of the conversation. According to Stalnaker (1975), there is a pragmatic difference between indicative and subjunctive conditionals in that whereas an indicative conditional focuses only on the set of possibilities in the shared common ground, subjunctive conditionals allow interlocutors to consider remote possibilities, outside the common ground, in their assessment. For instance, in example (3) from above, all interlocutors presuppose that Trump did indeed win the 2016 election: there are no other live possibilities. But the interlocutors can consider possible worlds outside of this common ground in evaluating the counterfactual stating what would have happened if Hillary had won.

For this theory then, the semantics specifies the *form* of truth conditions of conditionals by referring to selection functions (i.e. ‘$A > C$’ is true in w, iff C is true in the possible world/s returned by f(A, w)). *Which* proposition is expressed by the assertion of a conditional is, however, influenced by the context of utterance, given the pragmatic rule that f(A, w) outputs possible worlds *within* the context set for indicative conditionals, but is permitted to output possible words *outside* of the context set for subjunctive conditionals. As a result, the theory holds that it is *appropriate* to use an indicative conditional only in a context that is compatible with the antecedent (Stalnaker, 1975). Since counterfactuals cannot conform to the constraint of selecting worlds inside the context set, however, they must be expressed as subjunctive conditionals. This principle captures the intuition that while the use of indicative conditionals generally is understood under preservation of as many of the mutually shared presupposition of the conversation as possible, subjunctive conditionals are allowed to consider remote scenarios, where some of these assumptions are relaxed.

Subjunctive conditionals subtly cue their recipients through tense and mood that worlds may be selected outside the context set. As Iatridou (2000) has argued, subjunctives make use of an extra layer of past tense – the past perfect “had found” – as a “fake tense” to indicate the unlikeliness of the described situation rather than its temporal coordinates.

The Experiments

As we have outlined, while the possible-worlds account has proved highly influential, it has not proved easy to test, a point on which we elaborate below. This paper offers a new method for testing the truth evaluations of both indicative and subjunctive conditionals according to possible worlds semantics. Experiment 1 introduces the novel truth task for indicatives and subjunctives at the aggregate level; Experiment 2 uses this task to focus on indicative conditionals and study individual variation by modeling participants’ responses as coming from a mixture distribution of truth tables. Through both of our experiments, we contrast more than five different truth tables, including the novel account of Bradley (2012), which have not been tested previously, as far as we know.

5.2 Experiment 1: Truth Evaluations

The aim of Experiment 1 is to introduce an experimental setup for testing possible worlds semantics of subjunctive conditionals (Stalnaker, 1968; Lewis, 1973). To motivate it, we start out by explaining some of the difficulties with testing possible worlds semantics.

Difficulties in Testing Possible World Semantics

Possible worlds semantics for counterfactuals is widely used in linguistics and philosophy (Portner, 2009; Kratzer, 2012), because by placing weak constraints on the similarity relation—e.g., of all worlds, *w* is most similar to itself (*strong centering*)—it allows the formulation of several counterfactual logics. Yet, it turns out to be a very difficult task to explicate this similarity relation in a way that would permit the evaluation of ordinary counterfactuals in light of a set of strenuous counterexamples, and the inevitable context-sensitivity of such assessments (Rescher, 2007; Nickerson, 2015; Ippolito, 2016; Starr, 2019). To illustrate, a case can be made for either one of the following counterfactuals concerning the Korean War (Quine, 1960, p. 222; Spohn, 2013):

> If Caesar were in command, he would use the atomic bomb.
> If Caesar were in command, he would use catapults.

One of the central difficulties in testing possible worlds semantics is that participants are likely to consider very different possible worlds if left unconstrained due to differences in their background beliefs and in views on what counts as “similar”. For this reason, the experimental task which we introduce below creates a situation, where there is one salient hypothetical alternative to the present to the actual state of affairs to better control what counts as most similar in the task context.

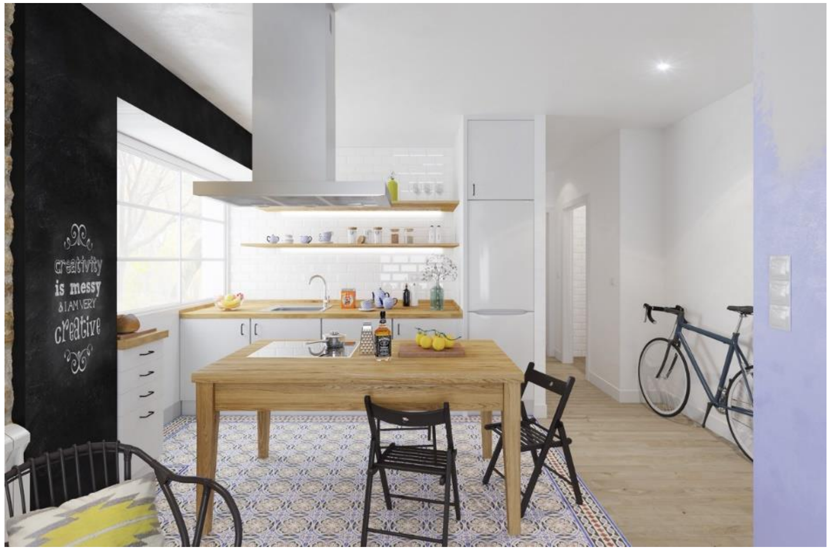
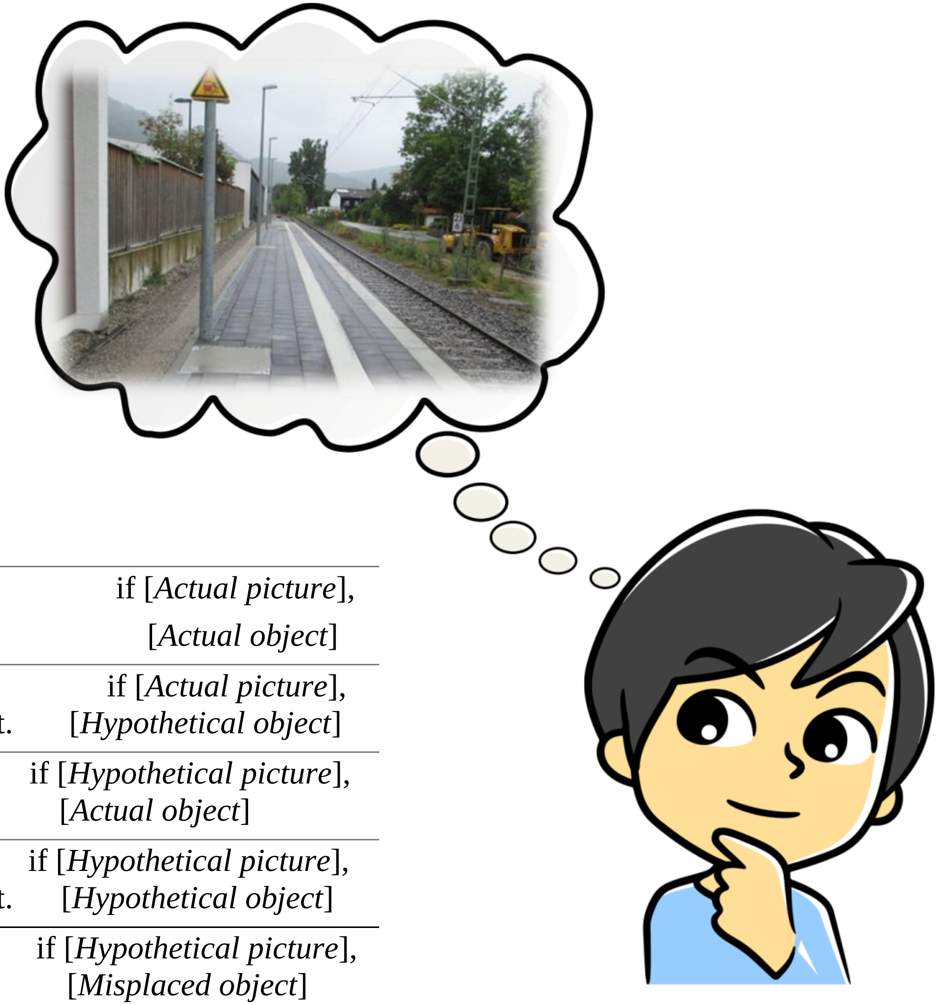
The Possible World Truth Table Task

In our experiments, participants were instructed that they were going to see a small sample of photos taken by Jack. The participants were then presented with different pairs of photos, which were later used to formulate the indicative and subjunctive conditionals, as illustrated in Table 2 below. In the example, the pair of pictures illustrate one of Jack's photos of a kitchen and of a railroad station. First, participants were asked a range of control questions concerning these pictures (explained below). Later, one of the pictures would be presented on the page and participants were asked to evaluate the truth value of the presented indicative/subjunctive conditional. For ease of illustration, we will refer to the picture presented as "the Actual Picture" and the other picture of the pair, which was not shown, as "the Hypothetical Picture".

We introduced this manipulation to investigate whether subjunctive conditionals with false antecedents would prompt participants to think about the counterfactual situation in which the Hypothetical Picture would be shown instead of the Actual Picture, as illustrated in the thought bubble in Table 2.

The function of the previous exposure to a pair of pictures is to constrain assessments of similarity by making salient one possibility of an alternative picture that could have been presented. This way, the task introduces an extra source of information (i.e., information about what is true/false in a salient possible A-world) beyond the manipulated truth table cells in the actual world (TT, TF, FT, FF). The goal is to probe whether participants selectively make use of this extra source of information when assigning truth values to subjunctive conditionals. For instance, when presented with the picture of a kitchen and the conditional "If this had been one of Jack's pictures of a railroad station, there would have been a warning sign about standing too close to the edge in it", do they consider what is true in the picture of the railroad station?

Table 2. Example of Stimulus Material

Actual Picture	*Hypothetical Picture*
	

TT: If this had been one of Jack’s photos of a kitchen, there would have been a fruit bowl in it.	if [*Actual picture*], [*Actual object*]
TF: If this had been one of Jack’s photos of a kitchen, there would have been a warning sign about standing too close to the edge in it.	if [*Actual picture*], [*Hypothetical object*]
FT: If this had been one of Jack’s photos of a railroad station, there would have been a fruit bowl in it.	if [*Hypothetical picture*], [*Actual object*]
FF: If this had been one of Jack’s photos of a railroad station, there would have been a warning sign about standing too close to the edge in it.	if [*Hypothetical picture*], [*Hypothetical object*]
FF$_{\text{misplaced}}$: If this had been one of Jack’s photos of a railroad station, there would have been a pair of bedlamps in it.	if [*Hypothetical picture*], [*Misplaced object*]

Note. The pictures were displayed in a size in which it was possible for the participants to discern the individual objects. *Actual Object*: bowl with fruit. *Hypothetical Object*: a warning sign about standing too close to the edge. *Misplaced Object*: a pair of bedlamps.

We further introduced a contrast between two ways of implementing the FF cell. We did this to probe whether participants were sensitive to variation in whether the objects talked about in the conditional sentences were present/absent on the Hypothetical Picture, although they would in each case be absent on the Actual Picture (thus giving rise to the FF cell, when evaluated relative to the Actual Picture). We label the first version of this truth table cell “FF” (where the object talked about is present in the Hypothetical Picture but absent in the Actual Picture). The second version where the object is also absent on the Hypothetical Picture, we label $\text{FF}_{\text{misplaced}}$.

To illustrate: whereas the Hypothetical Object (e.g., a warning sign) was present on the Hypothetical Picture and absent on the Actual Picture, a third Misplaced Object (e.g., a pair of bedlamps) was present on neither. The $\text{FF}_{\text{misplaced}}$ condition concerns this third misplaced object. If participants were evaluating the conditional based on the Actual Picture, then this contrast between two versions of the FF cell (FF vs. $\text{FF}_{\text{misplaced}}$) should not make a difference to their truth evaluations. If the conditionals were evaluated based on the salient Hypothetical Picture (e.g. Jack's picture of a railroad station), however, then the contrast between the Hypothetical Object and the Misplaced Object should play a role.

Thought of in terms of standard truth tables, one can think of this manipulation as implementing two versions, {TT, TF, FT, FF} and {TT, TF, FT, $\text{FF}_{\text{misplaced}}$}, and our interest is whether participants are systematically influenced in the truth evaluations of indicative and subjunctive conditionals by these two versions. When we discuss our findings below, we abbreviate the two versions as {TT, TF, FT, FF, $\text{FF}_{\text{misplaced}}$} and investigate how the various truth tables perform across all five types of truth table cells.

Predictions of Competing Truth Tables

While Experiment 1 is intended primarily to test the possible-worlds account, it also provides evidence that has a bearing on other truth tables of the conditional: namely, material implication, the de Finetti truth table, the Jeffrey Table (see Table 1), mental model theory, and the multi-dimensional approach of Bradley (2012). In Experiment 1, we attribute the predictions in Table 3 to the theories tested for indicative conditionals. The basis for these predictions is explained below.

Table 3. Predictions for Indicative Conditionals in Experiment 1

Cell	⊃	Possible Worlds	de Finetti	Bradley	Jeffrey	MMT$_{aux}$
TT	1	1	1	1	1	1
TF	0	0	0	0	0	0
FT	1	If (accom =1) 0; else 1	NA	0	0	0
FF	1	If (accom =1) 1; else 0	NA	1	1	1
FF$_{misplaced}$	1	0	NA	0	0	0
Model	$M_{Material}$	$M_{PossibleWorld}$	$M_{deFinetti}$	$M_{subjunctive}$		

Note. '⊃' = Material Implication. 'accom' = accommodate. The prediction for Bradley (2012) is determined by the task constraints of Experiment 1, since there is only one possible world that enters as a candidate for the counterworld, w_j, and this is the Hypothetical Picture. The truth value for the false antecedent cells (FT, FF, FF$_{misplaced}$) is therefore determined by the pair <actual picture, hypothetical picture>. The last row labels the truth tables by the multinominal processing tree models fitted in Experiment 1. The last row indicates which statistical model in Table 6 the truth table was mapped onto. The name '$M_{subjunctive}$' was chosen for the last three tables, because the indexed truth tables make the same predictions for indicative conditionals in this specific task as the modal values of the subjunctive conditionals.

As Edgington (2006) explains, the material implication and the possible world account in Stalnaker (1968) differ as follows. While the material implication holds that the indicative conditional is true when the antecedent is false, the possible world account in Stalnaker (1968) permits the conditional to be either true or false. For the possible worlds account, the conditional is, in other words, not truth-functional in that its truth value is not a function of the truth values of its clauses. Instead, the truth or falsity of the conditional in these cases is determined by whether the consequent is true in the most similar situation in which the antecedent is true (see Table 1).

The possible world semantics of Stalnaker (1975) applies the same truth conditions for indicatives and counterfactuals conditionals, but accounts for differences between the two through a pragmatic principle. If we are to fully understand the predictions of the possible-worlds account, we must briefly expand on this pragmatic principle.

As we have seen, Stalnaker (1975) invokes the notion of common ground in his account of indicative and counterfactual conditionals. Whereas an indicative conditional focuses only on the set of possibilities in the shared common ground, subjunctive conditionals allow interlocutors to consider remote possibilities, outside the common ground, in their assessment. Accordingly, it is appropriate to use an indicative only when the antecedent is not known to be false; otherwise, a subjunctive must be used. In Stalnaker (2011, 2014), this theory is extended using the notion of *accommodation*. Although presuppositions present information as taken-for-granted – as uncontroversially part of the common ground – they can be used to communicate novel information as well. But when presuppositions communicate novel information, they trigger a special mechanism.

To illustrate, suppose that a speaker says, "My sister is a schoolteacher" in a context, where it cannot be presupposed that the interlocutor even knew that the speaker had a sister. The speaker then goes beyond the common ground. But the utterance need not be infelicitous. The hearer can accept the presupposition as true – can *accommodate* it – and add the information to the common ground. Typically, this will occur automatically, unless someone objects (Lewis, 1979; Stalnaker, 2014). Accommodation thus has the discourse effect of making the hearer adjust her knowledge so that it entails the content that the speaker presupposes (Potts, 2005).

In (2011, 2014), Stalnaker extends the theory to permit accommodation of the common ground to make the context set compatible with the antecedent, if possible. The accommodation then creates a *posterior* context with respect to which the conditional is evaluated. Whereas the evaluation of counterfactuals permits the use of hypothetical assumptions that are not carried over to the main discourse, the accommodated presuppositions of this posterior context remain presuppositions in the subsequent discourse.[58] Accommodation is, thus, an important aspect of contemporary possible-worlds semantics, and it will be important to investigate whether participants perform such accommodation when faced with indicative conditionals with false antecedents.

If participants accommodate the common ground to make the context set compatible with the antecedent of the conditional, then the boundary of which possibilities count as being outside the context set shifts. Such shifts are predicted to affect the truth evaluation of indicative conditionals. For instance, when participants are presented with a picture of a kitchen and the conditional “If this is one of Jack’s pictures of a railroad station, then there is a warning sign about standing too close to the edge in it”, do they accommodate the antecedent and evaluate the condition with respect to the picture of the railroad station, or do they decline to accommodate and evaluate the conditional with respect to the presented picture of the kitchen? In our experiments, we test for the possible dependence of truth value judgments of indicative conditionals on accommodation. In Table 3 these shifts are illustrated through the “if … else …” clauses, which specify how the predicted truth values vary depending on whether participants accommodate the antecedent.

[58] A further possibility that Stalnaker (2011) discusses is to allow for truth-value gaps in case of false antecedents via a supervaluation approach for dealing with contextual ambiguity, just like the de Finetti table of Wijnbergen-Huitink et al. (2015) employs truth-value gaps. However, given our stimulus materials introduce a well-defined context without such ambiguity, this extension of the possible worlds account is set aside for present purposes (see Stalnaker, 2019, Chap. 11 for further discussion).

Bradley's (2012) account shares the non-truth functionality of possible worlds semantics but implements it differently. For Bradley (2012), the semantic content of conditionals is given by a set of ordered pairs of possible worlds: {<w_i, w_j>, …}. The first member (w_i) of such a pair is the actual world and the second world (w_j) is a counterworld in which the antecedent is true. If the consequent is true in w_j, then the conditional is true in <w_i, w_j>, otherwise it is false. Probabilities are assigned to conditionals by assigning probabilities to counterworlds. One of the great advantages of this technical refinement is that it allows Bradley (2012) to circumvent a problem known as the triviality results. The problem dates back to Lewis' (1976) demonstration that attempts to combine P(if A then C) = P(C|A) with classical truth conditions can only succeed for probability distributions subject to trivializing features that severely restrict their usefulness (for a review, see e.g., Bennett, 2003).

Through his constructions with truth conditions explicated via pairs of possible worlds, Bradley (2012) is able to show that he can combine the thesis, P(if A then C) = P(C|A), with truth conditions without being subject to the triviality results. This feature of Bradley's (2012) account is one of its main attractions. Via the exposure to participants of two pictures that Jack could have taken, only one of which is shown when evaluating the conditionals, our task allows us to experimentally fix the counterworld (w_j) and thus test Bradley's theory, possibly for the first time.

In contrast, the de Finetti truth table abandons classical truth conditions in a different way by holding that the indicative conditional is neither true nor false when the antecedent is false (Evans & Over, 2004; Fugard, Pfeifer et al., 2011; Baratgin, Over et al., 2013). A different,[59] but related approach, known as the Jeffrey-Table, holds that the semantic value of the conditional is given by P(C|A) (Over & Baratgin, 2017). As noted above, attempts have also been made to extend these truth tables to counterfactuals, so we will consider whether they can account for our data concerning subjunctive conditionals below.

Cruz and Over (2018) and Over (2020), have suggested that one can translate this third semantic value of the Jeffrey table into truth values supplied by participants via the auxiliary hypothesis that they assign the value 'true' when P(C|A) is "high" and 'false' when P(C|A) is

[59] Often these two approaches are taken to be similar, but Bradley (2012) shows that they differ on how to understand the probability of conditionals and that the semantic content attributed to indicative conditionals has a different meaning on the two accounts. Roughly, the de Finetti table avoids the triviality results by abandoning Bivalence and reinterpreting the probability of sentences as their probability of truth, *provided they are true or false*. In contrast, the Jeffrey table avoids the triviality results by dropping the independence of belief and the meaning of sentences by making the semantic values of indicative conditionals dependent on an agent's subjective degrees of belief.

"low". When outlining the predictions in Table 3, we rely on this empirical, auxiliary hypothesis. The resulting predictions are, in other words, specific to our experimental paradigm, and they presuppose that participants assign "low", "high", and "low" conditional probabilities for the FT, FF, and FF$_{\text{misplaced}}$ cells, respectively. In our experiments, we test this auxiliary hypothesis, and to anticipate, we obtain supporting evidence.

While mental model theory may earlier have adopted the material implication (Johnson-Laird & Byrne, 1991, 2002), the theory has since then been revised to reject such a commitment (Khemlani et al., 2018). On the revised version, indicative conditionals are viewed as conjunctive assertions about possibilities as shown in table 4.

Table 4. Mapping between indicative and counterfactuals, MMT

Row	**Partition**		**Factual:** *If A then C*	**Counterfactual:** *If A had happened then C would have happened*
1	A	C	Possibility	Counterfactual possibility
2	A	Not-C	Impossibility	Impossibility
3	Not-A	C	Possibility	Counterfactual possibility
4	Not-A	Not-C	Possibility	Fact

Note. Quelhas et al. (2018) call indicative conditionals "factual conditionals".

Whereas the indicative conditional asserts that only rows 1, 3, and 4 are possible, the counterfactual asserts that row 4 is a fact and that rows 1 and 3 are counterfactual possibilities which did not materialize. Since conditionals are viewed as conjunctive assertions about the rows of Table 4 on the revised mental model theory, there is a difficulty of how to assign truth values when participants are only presented with one of the rows, as in our experimental task. Quelhas et al. (2018) point out that the false antecedent cases (¬*A*&C and ¬*A*&¬C) are asserted to be possible for both true and false conditionals. So these are not diagnostic for the truth or falsity of the conditional. However, Goodwin and Johnson-Laird (2018, pp. 2529-2530) suggest a way of determining the truth of indicative conditionals when presented with only one of these cases. The idea is to make the truth of the indicative dependent on the truth of the corresponding counterfactual. In their example, if in fact Viv does not have shingles, then "If Viv has shingles, then she is in pain" is true provided that the following counterfactual is true: "If Viv had had shingles, then she would have been in pain". Accordingly, when one only possesses information about row 3 or 4 in Table 4, one can learn which of rows 1 and 2 are possible by considering a counterfactual version. Consequently, we attribute the prediction to mental model theory of lack of differences between indicative and counterfactual conditionals for the false antecedent cases as an auxiliary hypothesis, as outlined in Table 3.

5.2.1 Method

Participants, and sampling procedure shared for all experiments

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 292 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk from the USA, UK, Canada, and Australia. They were paid a small amount of money for their participation. The following *a priori* exclusion criteria were used: not having English as native language, completing the task in less than 240 seconds or in more than 3600 seconds, failing to answer at least one of two simple SAT comprehension questions correctly in a warm-up phase, and answering ‘not serious at all’ to the question 'how serious do you take your participation' at the beginning of the study. The final sample after applying the *a priori* exclusion criteria consisted of 211 participants.[60] Mean age was 44.06 years, ranging from 19 to 75. 42.65% of the participants identified as male; with the exception of 3, the rest identified themselves as female. 74.88 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. Applying the exclusion criteria had a minimal effect on the demographic variables.

Design

The experiment had a within-subject design with the following within-subject factors, which are explained below: Sentence (indicative vs. subjunctive) and Truth Table Cell (TT, TF, FT, FF, $FF_{misplaced}$). Since the experiment had three trial replications, participants saw a total of 30 within-subject conditions.

Procedure

To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002). Moreover, to ensure that the pictures were displayed properly if the participants completed the study on a smartphone, participants were asked to turn their smartphone in horizontal orientation, if they were using one.

The experiment was split into three blocks in random order. For each block, participants were instructed that they were going to see a small sample of photos taken by Jack. For each block, a pair of pictures was selected out of a pool of six possible pairs. The

[60] In addition to the common exclusion criteria for all the experiments, two participants were excluded in Experiment 1, because the javascript was not displaying correctly.

pairs were generated based on the stimulus materials used in a pilot study[61]. For instance, one pair of pictures might look as in Table 2, where the kitchen picture was selected as the Actual Picture shown.

For each of these pictures, participants were asked, in random order on separate pages, to answer whether it was true/false that three objects were on them, as control questions. In addition, participants were asked to evaluate conditional probabilities for fitting the Jeffrey table. Next, participants were asked to evaluate indicative and subjunctive conditionals for whether they were true, false, or neither. The conditional sentences were randomly assigned to the two members of the pair, so that participants would see one of each type (indicative vs. subjunctive) for each pair. On five separate pages, participants were then presented with the Actual Picture (e.g., the kitchen picture) and five conditionals (e.g., five subjunctive conditionals) implementing the truth table cells shown in Table 3 in random order and asked to evaluate whether the presented sentence was true (T), false (F), or neither true nor false (NN). Key words distinguishing indicative conditionals ("…is…, …is…") and subjunctives ("...had been …, …would ...") were highlighted in blue.

After completing the task, participants were presented with an accommodation task. The accommodation task featured three pages with pictures (e.g., Jack's picture of a kitchen) and indicative conditionals in the FF condition from the pairs displayed earlier in the study, that asked three times about what photo they took the sentence as referring to.

If participants accommodate the antecedent, which is incompatible with the picture shown (and the demonstrative reference: "If *this*..."), then they should select the hypothetical picture of a railroad station not shown. If, on the contrary, participants evaluate indicative conditionals with respect to the picture shown, they should select the displayed picture of a kitchen. Sample screenshots of these three accommodation items can be found on the osf project page: https://osf.io/6x7fb/?view_only=a2e50ef786e14fa99c883efa3e502af2.

5.2.2 Results

It was found across all pictures that participants correctly identified whether the Actual, Hypothetical, and Misplaced Objects were on the pictures in the initial control questions (median percentage of correct responses = 93 %, MAD = 2%). Participants' truth evaluations are displayed in Figure 1:

[61] Here is a link for preview: https://osf.io/6x7fb/?view_only=a2e50ef786e14fa99c883efa3e502af2

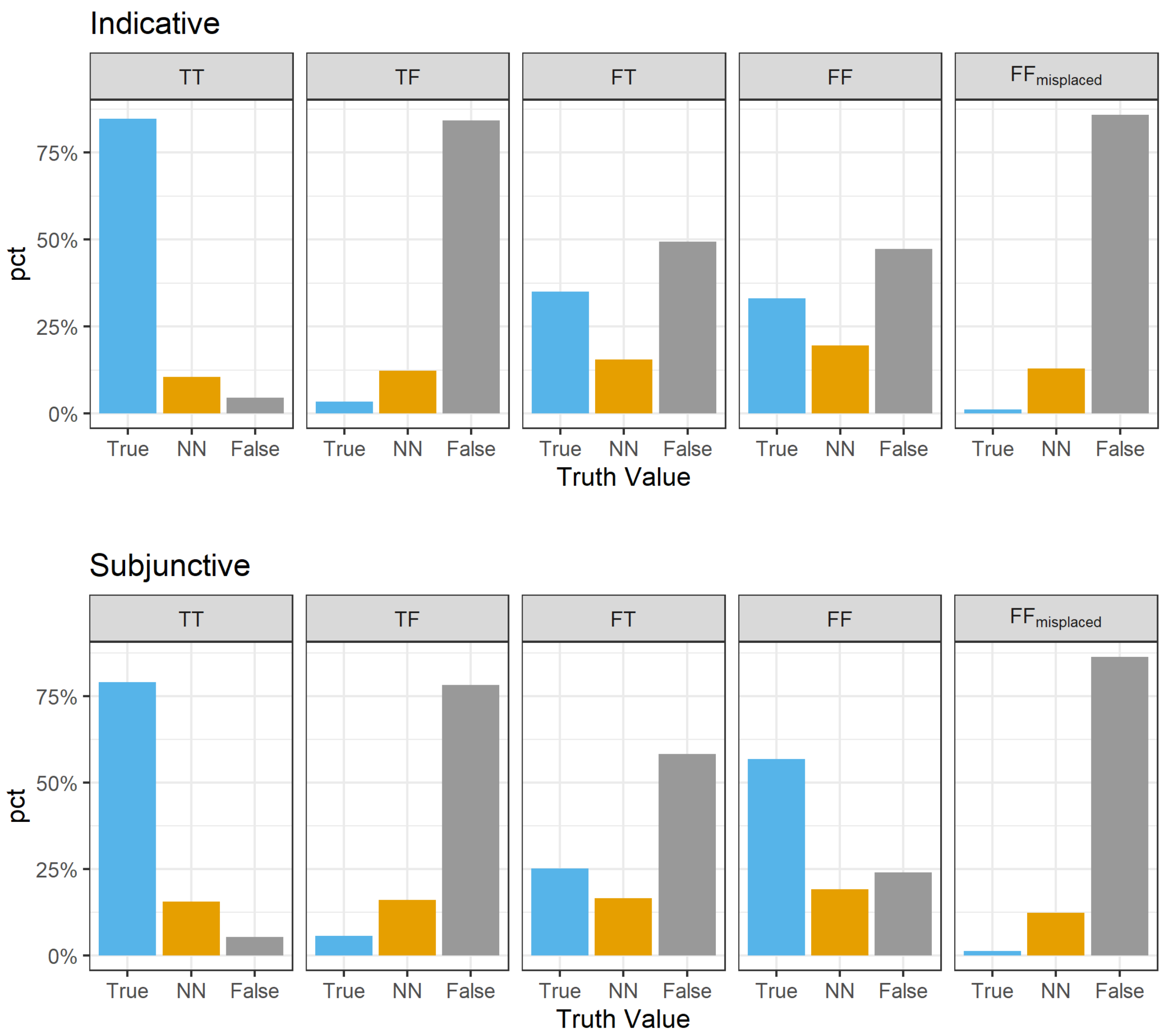


Figure 1. Truth Tables of Indicative and Subjunctive Conditionals. 'FF' = a conditional that is False False w.r.t. the Actual Picture shown but True True w.r.t. the salient Hypothetical Picture; 'FF$_{misplaced}$' = a conditional that is False False w.r.t. the Actual Picture shown but True False w.r.t. the salient Hypothetical Picture.

For the FT, FF, and FF$_{misplaced}$ cells, it was found across conditions that the average conditional probabilities for our task were .08 (SD = .09), .93 (SD = .09), and .04 (SD = .07). Accordingly, our results support assigning the truth values of Table 3 to the Jeffrey table by applying the auxiliary hypothesis of Cruz and Over (2018) and Over (2020) explained above.

Participants' truth evaluations were analyzed in two steps. First, differences between indicative and subjunctive conditionals were analyzed. Next, existing truth tables for indicative conditionals were fitted to the data. For both analyses, the observed response frequencies were analyzed with multinomial processing tree models (Riefer & Batchelder, 1988). This modeling framework is typically used to characterize the processes that underlie participants' categorical responses (see Batchelder & Riefer, 1999; Erdfelder et al., 2009). However, the framework can also be used to test hypotheses at the level of the observed response distributions through goodness of fit and model-selection statistics (e.g., Karabatsos, 2005; Klauer et al., 2015; Skovgaard-Olsen al., 2017). For a Bayesian implementation, we followed the hierarchical extension of multinominal processing trees in Klauer (2010), which was fitted via the R package TreeBUGS (Heck et al., 2018). Further technical details on the model can be found in Appendix B.[62]

Indicatives vs. Subjunctives

Using this framework, the following multinominal processing models were fitted to the data for both indicative and subjunctive conditionals to model the probabilities that a categorical response was selected:

$M_{\text{saturated}}$: model imposing no constraints. This model fits the data perfectly using one free parameter per degree of freedom provided by the data

M_{sentence}: model assuming that response probabilities are the same for indicatives and subjunctive conditionals

$M_{\text{FF,FFmis, ind}}$: model assuming no differences between the FF and $\text{FF}_{\text{misplaced}}$ truth table cells for indicative conditionals

$M_{\text{FT, FF, ind}}$: model assuming no differences between the FT and FF truth table cells for indicative conditionals

Table 5. Model-Comparison Results

Model	p_{T1}	p_{T2}	WAIC	Weight
$M_{\text{FT,FF,ind}}$	.37	.00	4683.8	0
$M_{\text{saturated}}$	.49	.11	4070.9	1
M_{sentence}	.00	.00	6705.4	0
$M_{\text{FF,FFmis,ind}}$	.00	.00	4599.7	0

Note. Note that the test statistics T_1 and T_2 represent Bayesian *p* values and are based on the posterior predictive model checks in Klauer (2010). WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC.

Model fit was assessed with the WAIC information criterion and posterior predicted *p* values based on T_1 and T_2 posterior model checks proposed in Klauer (2010). T_1 measures the adequacy of the models in capturing the mean observed truth value frequencies. T_2 measures the adequacy of the models in capturing the variability (variances and covariances) among the observed response frequencies. A small (Bayesian) *p* value for these test statistics indicates that the posterior predictive distribution of the model fails to capture an aspect of the data,

[62] R scripts will be made available on osf upon publication.

because this aspect of the actual observations is unlikely to be predicted by the model. When testing several of these posterior predictive checks, it is not uncommon for a model to be inadequate for some purposes but adequate for others, and the test statistics help us identify which aspects of the data are captured by a given model (Gelman et al., 2013: Ch. 6).

Based on these criteria, ($M_{\text{FT, FF, ind}}$) was the only model besides the saturated model (M_{sat}), which was able to capture the mean observed truth value frequencies (T_1). Yet, like the other non-saturated models, it accounted less well for their variances and covariances across individuals (T_2). Accordingly, the information criterion WAIC indicates that this model reduction was not justified compared to the saturated model. From the comparisons, we can infer that a) there is a difference in the truth value assignments of indicative and subjunctive conditionals, b) the difference between FF and $\text{FF}_{\text{misplaced}}$ plays a role even for indicative conditionals, but c) the difference between FT and FF need not be taken into account to account for the mean observed truth value frequencies of indicative conditionals; yet it plays a role for the variability in responses across individuals. Moreover, a glance at Figure 1 reveals that while the difference between FT and FF may not matter for indicative conditionals, it flips the modal value for subjunctive conditionals from false to true.

Truth Tables for Indicatives

To further investigate participants' truth value assignments for indicative conditionals, we fitted the truth tables from Table 3 to the data (see Table 6 below).

As indicated in Table 2, the truth table for the possible world semantic depended on whether participants accommodated the antecedent. As a result, the model implementing this truth table was constrained to predict the truth value assignment {T, F, F, T, F} for the participants (N = 95), who accommodated the reference of the antecedent to refer to the Hypothetical Picture. In contrast, the model was constrained to predict {T, F, T, F, F} for the participants (N = 116) who did not accommodate to follow the truth evaluation of the Actual Picture. This classification concerns participants who accommodated the reference of the false antecedent all three times when asked. Accordingly, the prior expectation of answering "yes" all three times by a random binominal process is that only 12.5% of the participants should have fallen in this group, instead 45% were found. Appendix A displays the bimodal pattern of these two different truth evaluations of indicative conditionals separately, which are merged in the aggregate results shown in Figure 1.

Following Skovgaard-Olsen et al. (2017), the most lenient criterion was applied in assigning a stochastic interpretation to the deterministic predictions of the truth tables using the order constraints in Klauer et al. (2015), as explained in Appendix B. In each case, only a relative majority of the predicted response was required of a given model (i.e., that the predicted response occurs at least as often as each of the other responses). In contrast, an absolute majority could also have been required. But the advantage of using this very lenient criterion lies in the diagnostic power associated with its failure, as any theory that fails under these minimal constraints should be seriously questioned.

Table 6. Model-Comparison Results for Indicatives

Model	p_{T1}	p_{T2}	WAIC	Weight
$M_{PossibleWorld}$	0.134	0.000	2258.9	1
$M_{Subjunctive}$	0.000	0.000	2549.7	0
$M_{Material}$	0.000	0.000	3405.8	0
$M_{deFinetti}$	0.000	0.000	3606.7	0

Note. The test statistics T_1 and T_2 represent Bayesian *p* values and are based on the posterior predictive model checks in Klauer (2010). WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of WAIC. In Table 3, the truth tables are displayed which correspond to these models.

Model fit was assessed with the WAIC information criterion and posterior-predicted *p* values based on T_1 and T_2 posterior model checks proposed in Klauer (2010). Of all the models tested, only $M_{PossibleWorld}$ had a satisfactory fit for the aggregate truth value frequencies (T_1). But like the other models, it accounted less well for their variances and covariances across individuals (T_2). The information criterion WAIC indicates a strong preference for $M_{PossibleWorld}$ in light of the parsimony vs. fit trade-off. We discuss why $M_{PossibleWorld}$ had this edge compared to the other models further below.

5.2.3 Discussion

As seen from Figure 3, participants' truth evaluations of subjunctive conditionals elicit the modal pattern {T, F, F, T, F} in spite of being shown a picture that would generate the following truth table cells: TT, TF, FT, FF, FF. This pattern fits with the following hypothesis. When participants evaluate subjunctive conditionals with false antecedents, they do not consider whether the sentence is true of the Actual Picture displayed (e.g., Jack's photo of a kitchen). Instead, they consider whether it is true of the Hypothetical Picture (e.g., Jack's photo of a railroad station) that the experiment made salient. This pattern corroborates the possible world account of subjunctive conditionals (Stalnaker, 1968; Lewis, 1973). Given the way the task was set up, participants were already familiar with a prototype that could make the antecedent true of the subjunctive conditionals. It was thereby possible for participants to solve the task of evaluating subjunctive conditionals at the closest possible-A worlds in a relatively uniform manner. These results thereby provide some of the first direct evidence that participants' truth evaluations coincide with a possible-world interpretation of subjunctive conditionals. Neither the material implication nor the de Finetti table can capture the modal pattern in the truth evaluations for subjunctive conditionals. Applying mental model theory to account for the same results is, however, possible but raises special theoretical issues, which we take up in the General Discussion.

The results moreover indicate that participants clearly distinguish between the truth evaluation of subjunctive and indicative conditionals. In contrast to subjunctives, the results for indicatives showed considerable individual variation. Given this variation, we see that the conjecture in Goodwin and Johnson-Laird (2018)—that indicative conditionals with false antecedents are true if the corresponding counterfactuals are true—significantly misfits the data. More specifically, the account encountered difficulties in the FF cell in which the subjunctive was evaluated as true yet the modal value for the matching indicative was false.

In general, it was found for indicative conditionals that influential truth tables, like the material implication (T, F, T, T) and the de Finetti table (T, F, NN, NN), severely misfit the data. In this, the present results are in line with the results of a recent meta-analysis (Schroyens, 2010). Some studies have recently reported stronger evidence in favor of the de Finetti truth table for indicative conditionals, however (e.g., Evans et al. 2007; Politzer et al., 2010; Wijnbergen-Huitink et al. 2015). But these studies have either relied on an experimental paradigm that asks whether a truth table cell “conforms to”/”contradicts” or “is irrelevant” to a conditional rule (for discussion, see Skovgaard-Olsen, 2020), or relied on a betting paradigm, which likewise requires participants to consider whether the evidence presented verifies a conditional rule. In contrast, the de Finetti truth table is much less well-supported when participants are asked to assign ternary truth values as here (Schroyens, 2010; Skovgaard-Olsen et al. 2017).

Proponents of the suppositional theory of conditionals (see e.g. Over & Baratgin, 2017) have emphasized the need for replacing the “void” values of the de Finetti table with the conditional probabilities of the Jeffrey table (see Table 1). To test this idea, we measured participants’ conditional probabilities. Based on the measured probabilities, we would have expected participants to give the same truth values to indicatives with false antecedents as the modal truth values for subjunctives. However, an inspection of Figure 3 shows why such a model did not fit the mean response frequencies. The main difficulty consists in accounting for the differences between indicative conditionals and subjunctives in the FT and FF cells. So, while a Jeffrey table is compatible with the truth values for subjunctive conditionals, it cannot account for the modal truth values for indicative conditionals in our experiments.

Of all the investigated alternatives, it was found that a possible worlds account was the only model which did not significantly misfit the data on the aggregate level. What permitted this account to outperform the other models was the following. The possible worlds model adjusted its predictions based on whether participants accommodated the reference of indicative conditionals with false antecedents to refer to the Hypothetical Picture not shown (despite the demonstrative reference to the Actual Picture shown). Ca. 55% of the participants chose to accommodate in this way across three trials. It turned out that this difference in whether participants accommodated was a factor in the truth evaluation of indicative conditionals, as shown in Appendix A. But given that the predictions of the truth tables used a stochastic representation that only required that the preferred option should be the modal response, there is scope for further investigations into factors that may influence individual variation in the truth evaluation of indicative conditionals. Indeed, the failure of all models to account for the variances and covariances across individuals indicate that further research into individual variation is needed.

5.3 Experiment 2: Individual Differences

Experiment 1 indicated that the possible world account was the only of the investigated theories that did not misfit the data at the aggregate level for indicative conditionals. At the same time, the posterior predictive checks indicated that none of the investigated theories were capable of accounting for the variances and covariances across individuals. For this reason, Experiment 2 applied a Bayesian mixture distribution analysis to investigate individual variation in participants' truth evaluations.

To increase the number of trial replications per participants, only indicative conditionals were investigated in Experiment 2. In addition, Experiment 2 also measured whether participants viewed the antecedent as a reason for or against the consequent. This dependent variable was included to investigate whether making reason relations, or inferential relations, part of the truth conditions would help account for participants' responses, as posited by truth-conditional inferentialism (Douven et al., 2018).

5.3.1 Method

Participants

288 people participated in the study. The final sample after applying the *a priori* exclusion criteria from Experiment 1 consisted of 211 participants. Mean age was 41.01 years, ranging from 19 to 73. 43.60% of the participants identified as male; the rest identified

themselves as female. 80.1 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. Applying the exclusion criteria had a minimal effect on the demographic variables.

Design

The experiment had a within-subject design with the following within-subject factors: Truth Table Cell (TT, TF, FT, FF, $FF_{misplaced}$). Since the experiment had six trial replications of indicative conditionals, participants saw a total of 30 within-subject conditions.

Procedure

Experiment 2 followed the procedure of Experiment 1 with one exception. In addition to the initial control questions concerning the truth/falsity of claims stating that three objects were on the pictures, and the conditional probability questions for the Jeffrey table, Experiment 2 included questions concerning reason relations. To illustrate, in one condition participants were asked to evaluate whether the statement "The picture shown is one of Jack's pictures of a study" is a reason for/against the statement "There is a shampoo in it"[/"There is a study lamp in it"/"There is a pair of bed lamps in it"]. Participants were provided with a five point labelled Likert scale {a strong reason against; a reason against; neither for nor against; a reason for; a strong reason for} to give their responses.

With this addition to the procedure of Experiment 1, the dependent variables for Experiment 2 were: (1) ternary truth evaluations, (2) conditional probabilities, (3), ordinal reason relation assessments, (4) three accommodation questions, (5) control questions ensuring that the participants understood the truth table cell by correctly identifying the presence/absence of named objects w.r.t. the displayed pictures.

5.3.2 Results

As shown in Figure 2, participants' ternary truth value judgments of indicative conditionals replicated those found in Experiment 1.

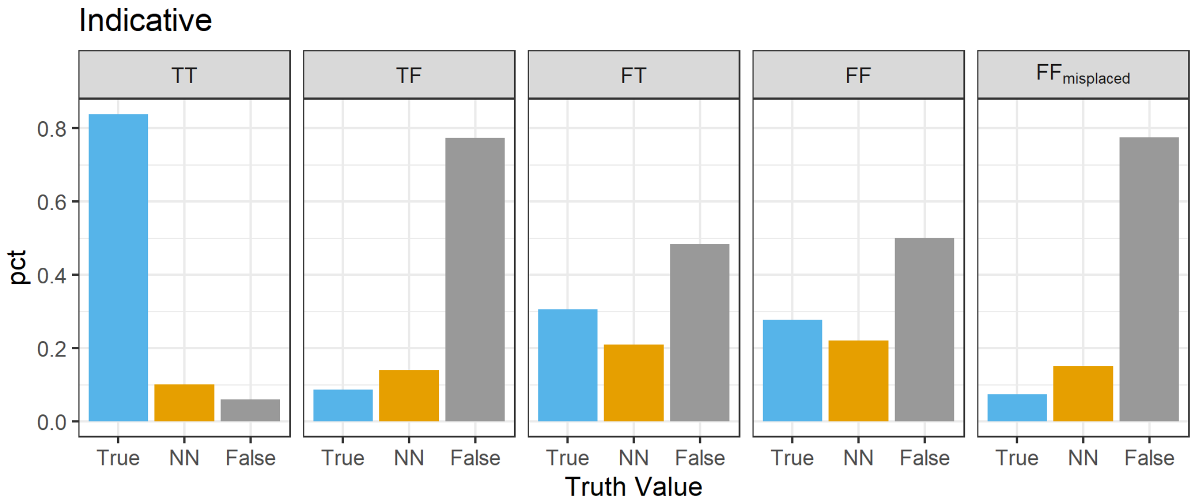


Figure 2. Truth Tables of Indicative. ‘FF’ = a conditional that is False False w.r.t. the Actual Picture shown but True True w.r.t. the salient Hypothetical Picture; ‘FF$_{misplaced}$’ = a conditional that is False False w.r.t. the Actual Picture shown but True False w.r.t. the salient Hypothetical Picture.

According to truth-conditional inferentialism advanced in Douven et al. (2018), the indicative conditional is true if the consequent can be inferred from the antecedent (possibly via background assumptions) and false if it cannot be inferred. To be able to test this theory in the present paradigm, participants were asked for the extent to which the antecedent provided a reason for/against the consequent on a five-point Likert scale. For each pair of pictures, these reason relation assessments were made by having the antecedent describe one of the pictures of the pair and having the consequent either describing a matching object (e.g., a fruit bowl in a picture of kitchen) or a mis-matching object (i.e., the warning sign of the hypothetical picture or the misplaced object, e.g., a pair of bed lamps). The results are shown below.

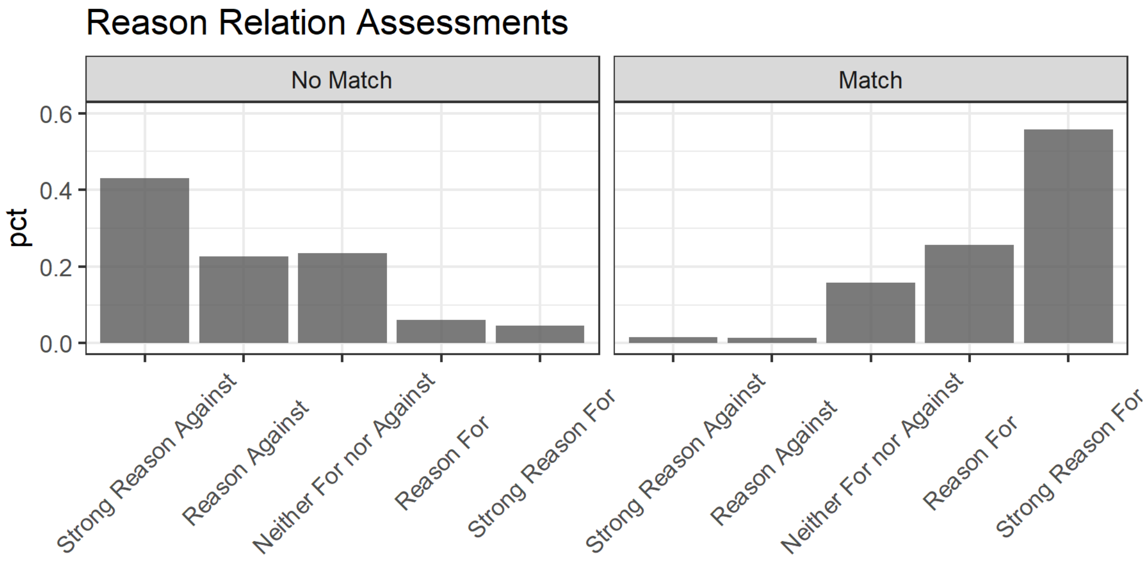


Figure 3. Reason relation assessments on a five-point Likert scale. The TT and FF cell correspond to the ‘matching cases’, where an object from the Actual/Hypothetical Picture was mentioned in the consequent and the corresponding Actual/Hypothetical Picture was mentioned in the antecedent. The TF, FT, and FF$_{Misplaced}$ cells correspond to the mismatching cases, where objects in the consequents were mentioned that violated the expectations concerning the pictures mentioned in the antecedent sentences.

The TT and FF cells correspond to the matching cases and the TF, FT, and FF$_{\text{misplaced}}$ cells correspond to the mismatching cases. Accordingly, truth-conditional inferentialism makes the same predictions as M$_{\text{Subjunctive}}$ in Table 3 for our experimental paradigm. Like in Experiment 1, conditional probabilities were measured to apply the Jeffrey table which holds that participants assess conditional probabilities in the false antecedent cells. When these conditional probabilities are high, participants are predicted to treat the conditional as true and when the conditional probabilities are low, participants are predicted to treat the conditional as false, based on the auxiliary assumptions of Over (2020) and Over and Cruz (2018).

Aside from a few outliers, a clear trend was recognizable with participants assigning high conditional probabilities in the FF cell (Mean = 90.15, SD = 11.78) and low conditional probabilities in the FT (Mean = 18.17, SD = 22.90) and FF$_{\text{Misplaced}}$ cells (Mean = 13.68, SD = 22.97), thus giving rise to the same predictions as M$_{\text{Subjunctive}}$ in Table 3.[63]

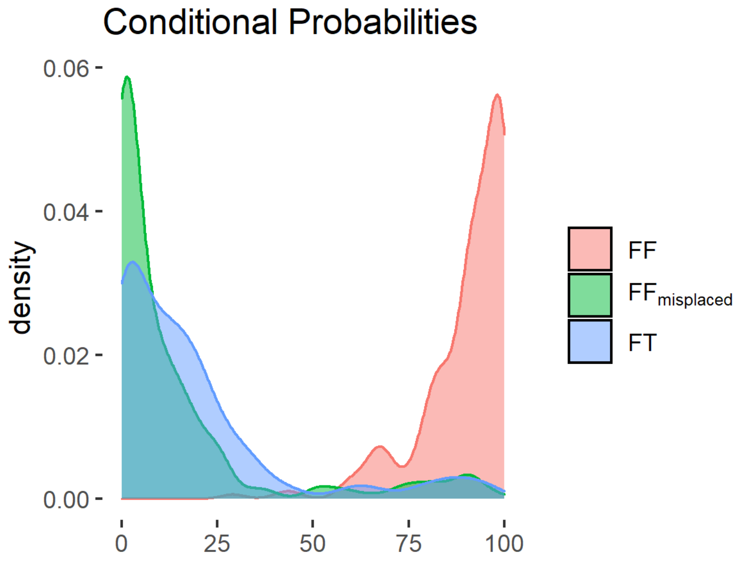


Figure 4. Conditional Probability distributions on a percentage scale between 0-100%.

Four Bayesian Mixture models were fitted in JAGS (Plummer, 2019) with a categorical variable deciding which of the mixture components an individual participant was assigned to. Three of these models included correlations among the MPT parameters in the hierarchical structure, following the latent trait model of Klauer (2010) applied in Experiment 1. The fourth model assumed that there were no correlations among these MPT parameters across

[63] Accordingly, for our experimental paradigm, this model encompasses truth-conditional inferentialism, the Jeffrey Table, MMT$_{\text{aux}}$, and Bradley's (2012) truth conditions. As we recall, the name 'M$_{\text{subjunctive}}$' was chosen for this model, because the indexed truth tables make the same predictions for indicative conditionals in this specific task as the modal values of the subjunctive conditionals in Experiment 1.

participants, following the beta-MPT approach of Smith and Batchelder (2010, see e.g., Heck et al. 2018). Further technical details can be found in Appendix B.

Due to the partial overlap in the predictions between the predictions of the truth tables, $M_{PossibleWorld}$ and $M_{Subjunctive}$, we fitted several versions of the latent trait model:

M1: Latent Trait model with three mixture components ($M_{Material}$, $M_{deFinetti}$, $M_{PossibleWorld}$).
M2: Latent Trait model with four mixture components ($M_{Material}$, $M_{deFinetti}$, $M_{PossibleWorld,}$ $M_{Subjunctive}$).
M3: Latent Trait model with five mixture components ($M_{Material}$, $M_{deFinetti}$, $M_{PossibleWorld,}$ $M_{Subjunctive,}$ $M_{conjunction}$).
M4: Beta-MPT model with four mixture components ($M_{Material}$, $M_{deFinetti}$, $M_{PossibleWorld,}$ $M_{Subjunctive}$).

In addition, all models featured a saturated mixture component to filter out noisy respondents, which did not fit any of the models in virtue of violating the shared predictions of ‘True’ and ‘False’ in the TT and TF cells.

The fifth mixture component in M3 was added, because previous research on individual variation with conditionals have shown that participants sometimes produce conjunctive responses (Evans et al., 2007). The fourth and final model was a beta-MPT version of M2, which differed from M2 by assuming that there were no correlations among these MPT parameters across participants in their *a priori* distributions (Smith & Batchelder, 2010). We then quantified the respective predictive performance of the four models by the leave-one-out cross validation criterion and WAIC.

Table 7. Model Comparison

	LOOIC	Δelpd	SE	WAIC	Weight	p_{T1}	p_{T2}
M3: Latent Trait 5	3892.82	0	0	3410.69	1.00	0.24	0.0005
M1: Latent Trait 3	4248.42	-177.80	24.32	3802.46	0.00	5e-06	0
M2: Latent Trait 4	4283.11	-195.15	24.85	3904.27	0.00	5.5e-05	0
M4: Beta-MPT 4	4958.75	-532.97	37.86	4679.38	0.00	1.5e-05	0

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = model averaging via staking of the predictive distributions. ‘elpd’ = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. The test statistics T_1 and T_2 represent Bayesian *p* values and are based on the posterior predictive model checks in Klauer (2010). ‘Beta-MPT 4’ refers to the beta-MPT approach of Smith and Batchelder (2010) with 4 mixture components. ‘Latent Trait 4’ refers to the hierarchical latent trait approach of Klauer (2010) with 4 mixture components. The two models differ on whether they permit correlations in the MPT parameters across participants.

The information criteria show a preference for the latent trait model with correlated MPT parameters across participants over the Beta-MPT model without these correlations. In addition, they indicate that the latent trait model with five mixture groups (M3) is preferred in light of the fit and parsimony trade-off. In addition, M3 was also the only model that showed a satisfactory fit for the aggregate truth value frequencies (T_1).

As Figure 5 shows, in general participants classified according to these three mixture groups had a relatively high posterior probability of agreeing in their truth value judgments with their assigned truth table.

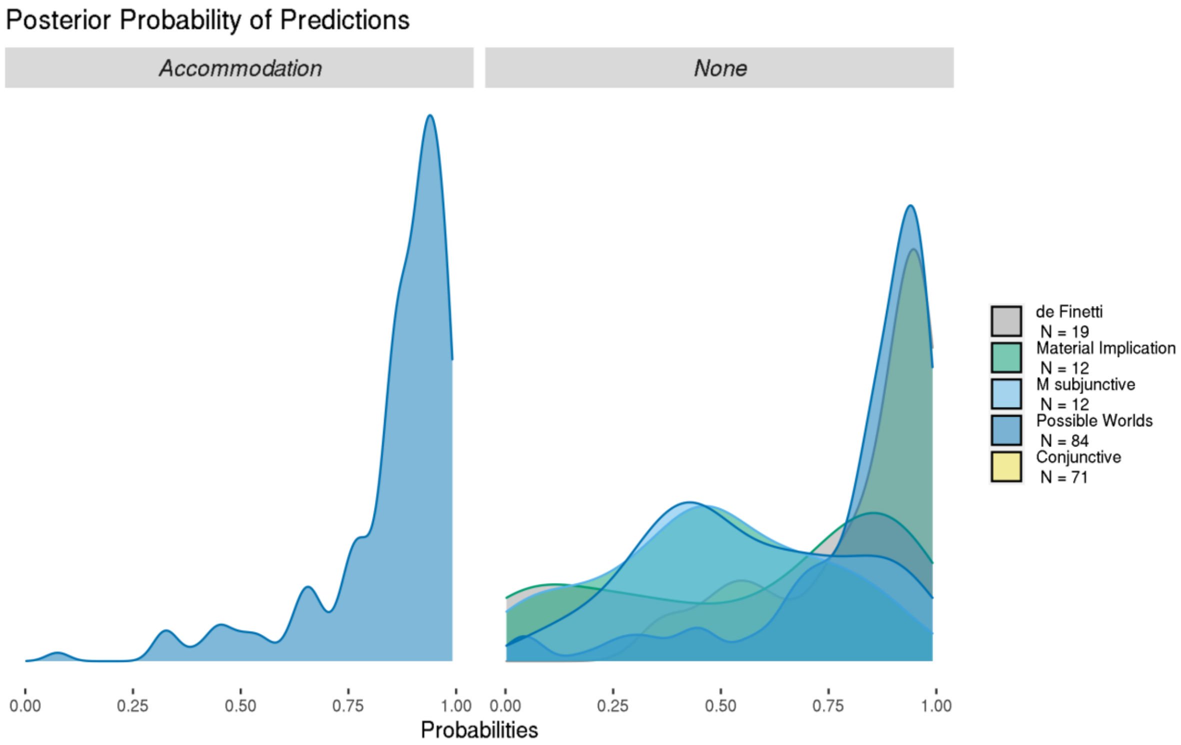


Figure 5. Posterior probability of the predictions of participants' truth value judgments agreeing with their predicted truth tables, for each of the three mixture groups of M3. 13 participants were captured by the saturated model used to filter out participants, who did not conform to the shared prediction of 'true' and 'false' in the TT and TF cells.

As the comparison shows, the largest group of participants were assigned to the possible world model. But individual variation was also found and so the model had to be supplemented with the other truth tables.

5.3.3 Discussion

As Figure 2 shows, Experiment 2 replicates the qualitative patterns of truth value judgments of indicatives from Experiment 1. Of the different models investigated in Experiment 1, $M_{\text{Subjunctive}}$ makes the same predictions for indicative conditionals as the modal truth value assignment for subjunctives in Experiment 1. It was shown in Table 3 that this model characterized several competing theories for our experimental paradigm. Based on the measured conditional probability judgments, it was found, like in Experiment 1, that $M_{\text{Subjunctive}}$ also captures the predictions of the Jeffrey truth table. In addition, based on the ordinal reason relation judgments of the participants, it was found in Experiment 2 that $M_{Subjunctive}$ also captures the predictions of truth-conditional inferentialism for our task.

In Experiment 1, $M_{PossibleWorld}$ was the only model not misfitting participants' responses at the group level. In contrast, Experiment 2 sought to investigate via mixture distributions whether subgroups of participants could be identified that followed competing truth tables.

Since, however, $M_{PossibleWorld}$ and $M_{Subjunctive}$ have overlapping predictions for the subset of participants who did not accommodate the antecedent, we did a model comparison between four models (M1, M2, M3, M4). It was found that a latent trait model (M3) that included five mixture groups ($M_{Material}$, $M_{deFinetti}$, $M_{PossibleWorld,}$ $M_{Subjunctive,}$ $M_{conjunction}$) performed the best in light of the fit vs. parsimony trade-off. In this model, it was found that overall participants had a high posterior probability of following the assigned truth table in their truth value judgments, across mixture groups (Figure 5). In agreement with Experiment 1, it was found that the majority of participants could be captured by $M_{PossibleWorld}$ at the individual level. In contrast, very few participants were captured by $M_{Subjunctive}$.

Like previous studies investigating individual variation (Evans, et al., 2007), it was found that there was a sizable minority of participants who produced a conjunctive pattern. In Experiment 2, this pattern was stronger than in Experiment 1, and it is thus possible that the within-subject comparison with subjunctive conditionals in Experiment 1 suppressed this response tendency.

5.4 General Discussion

Through our experiments, we have compared the possible worlds account to more than five competing truth tables and found that it could survive the competition. In Experiment 1, we investigated both indicative and subjunctive conditionals and found that the possible world account was the only one of the investigated theories that did not misfit the data at the aggregate level for indicative conditionals. In Experiment 2, we modeled participants' truth evaluations of indicative conditionals as a mixture distribution of competing truth tables and found that the possible world semantics accounted for the largest subgroup in the analysis of individual variation in our experimental paradigm. Collectively, both studies obtained evidence in favor of possible worlds semantics over a wide range of popular alternatives.

That indicative and subjunctive conditionals differ in content is illustrated by the following famous example (Adams, 1970), where most will accept the first indicative conditional while rejecting the second subjunctive:

(*indicative*) If Oswald did not shoot Kennedy, someone else did.

(*counterfactual*) If Oswald had not shot Kennedy, someone else would have.

The formulation of this minimal pair illustrating the indicative/subjunctive divide, has led to numerous attempts to either provide a unifying account (Stalnaker, 1975; Edgington, 2008; von Fintel, 2012; Spohn, 2013; Starr, 2014; Williamson, 2020), argue why disjunct accounts are needed (Lewis, 1973, 1976; Bennett, 2003), or argue for a unifying account by questioning that this indeed constitutes a minimal pair (Quelhas et al., 2018).

For proponents of the unifying account, it is tempting to formulate one semantics of conditionals, like possible worlds semantics, and look to linguistic phenomena closer to pragmatics to account for the differences in the sentences. For instance, in Skovgaard-Olsen and Collins (2021), evidence was found that at least the difference in the status of epistemic openness towards the antecedent of the indicative and disbelief towards the antecedent of the subjunctive could be attributed to the pragmatic phenomenon of conversational implicatures.

Below we are going to show that the strategy of Stalnaker (1968, 1974) of formulating one type of truth conditions in an abstract form that can be shared by indicatives and subjunctives, while supplying a pragmatic principle that account for their differences, is a fruitful strategy for accounting for our experimental data on truth evaluations. To make this argument, the discussion below starts out by considering how our results bear on the non-truth functionality of conditionals and Stalnaker's (1975) claim that indicative and subjunctive conditionals have truth conditions that share the same abstract form, but which diverge by applying his pragmatic principle. Next, we compare this approach with mental model theory as applied to our experiments. Finally, we contrast our own approach to making possible worlds semantics empirically testable with some alternative strategies.

Non-Truth Functionality

In Experiment 1, we presented a novel experimental task for probing the possible worlds semantics of Stalnaker (1968) and Lewis (1973). The idea was to experimentally constrain assessments of closest possible worlds by offering participants a prototype of what the closest possible-*A* world could look like (e.g., "one of Jack's photos of a railroad station"). We did this to probe whether participants selectively made use of this additional source of information, rather than the displayed image (e.g., "one of Jack's photos of a kitchen"), when assigning truth values to subjunctive conditionals. The results supported the constraint for a unifying account of indicative and subjunctive conditionals that their truth evaluations differ for conditionals with false antecedents.

An important aspect of the possible worlds account is the non-truth functionality of indicative and subjunctive conditionals (Edgington, 2006). On this account, the truth of a conditional is not only a function of the truth value of the antecedent and the consequent; rather it depends on what is true in the most similar antecedent situation.

Through the introduction of two types of FF cells with respect to the Actual Picture shown (the FF cells with the Hypothetical Objects and the $\text{FF}_{\text{misplaced}}$ with the Misplaced Objects), Experiment 1 was designed to test for non-truth functionality. Given the large differences between the FF and $\text{FF}_{\text{misplaced}}$ cells for both subjunctive and indicative conditionals, this feature of the possible worlds semantics is corroborated by the results. A constraint for a unifying account of indicative and subjunctive conditionals is therefore that our results suggest that they are both marked by the property of non-truth functionality. More generally, it was found in Experiment 1 that there was a clear tendency to not evaluate subjunctive conditionals truth functionally based on the Actual Picture shown (e.g., Jack's picture of a kitchen). Instead, most participants consider a hypothetical situation in which the Hypothetical Picture is displayed (e.g., Jack's picture of a railroad station), when the antecedent is false, and evaluate the subjunctive conditional based on it instead.

For indicative conditionals, participants were split between evaluating conditionals with false antecedents based on the Actual Picture displayed (T, F, T, F, F) or on the Hypothetical Picture (T, F, F, T, F), which the majority use for evaluating subjunctive conditionals (see Appendix A). In this task, there is a tension between the demonstrative reference to the Actual Picture shown (e.g., Jack's picture of a kitchen) and a description of the Hypothetical Picture (e.g., Jack's picture of a railroad station) in the false antecedents. As a result, participants were found to oscillate between the truth evaluations based on the Actual Picture and the Hypothetical Picture.

One way of interpreting this oscillation is based on the pragmatic principle of Stalnaker (1975). The principle constrains the selection function to consider possible situations within the context set for indicative conditionals while permitting the selection of possibilities outside the context set for subjunctive conditionals. The picture that is shown is clearly part of the context set. Yet, the indicative conditionals with the false antecedents violate Stalnaker's (1975) pragmatic principle when describing a hypothetical picture (e.g., Jack's picture of a railroad station) while demonstratively referring to the picture shown. Participants react differently to this: some are found to accommodate and take the indicative conditional as strictly referring to the Hypothetical Picture (despite the demonstrative reference); others evaluate the conditional based on the Actual Picture and ignore that the antecedent mischaracterizes it in terms of a feature of the Hypothetical Picture.

Experiment 1 thereby shows what happens when Stalnaker's pragmatic principle for indicative conditionals is violated. A bi-modal pattern emerges on the aggregate level which is not compatible with any of the other truth tables considered in Table 3. For subjunctive conditionals with false antecedents, in contrast, participants evaluate the conditionals based on the Hypothetical Picture, in accordance with Stalnaker's (1975) pragmatic principle, which permits them to consider possibilities outside the context set.

On Stalnaker's (1975) theory, the semantics specifies the *form* of truth conditions of conditionals by referring to selection functions (i.e. '$A > C$' is true in w, iff C is true in the possible world/s returned by f(A, w)). *Which* proposition is expressed by the assertion of a conditional is, however, influenced by the context of utterance, given the pragmatic rule that f(A, w) outputs possible worlds *within* the context set for indicative conditionals, and that this selection function can output possible worlds *outside* the context set for subjunctive conditionals. Our findings support this strategy of using the *same abstract form* of the truth conditions for indicative and subjunctive conditionals but allowing their truth conditions to diverge through the application of a pragmatic principle that constrains the selection of most similar situations.

In connecting the results on truth value assignments of Experiment 1 with the finding in Skovgaard-Olsen and Collins (2021) that the falsity of the antecedent expressed by a subjunctive conditional is a conversational implicature, the following observations can be made. If use of the subjunctive mood, and the fake past tense of subjunctive conditionals, generates the conversational implicature that the antecedent is false, then this has implications for the truth evaluations of subjunctive conditionals, on a possible worlds account. The conversational implicature that the antecedent is false warrants interlocutors to use possible worlds outside the common ground (in our case: the Hypothetical Picture) in assigning truth values to the conditional. The conversational implicature is, however, cancellable, which means that the search for the closest possible A-worlds outside the context-set can be overridden. One case in point, is the famous example of so-called Anderson conditionals:

(4) "If Jones had taken arsenic, he would have shown exactly those symptoms which he does in fact show" (Anderson, 1951, p. 37).

Since a speaker of this conditional could use (4) to argue that Jones had, in fact, taken arsenic, the conversational implicature of the subjunctive conditional that Jones did not take arsenic is cancelled in this case (von Fintel, 1997, 2012; Stalnaker, 1975, 2014).

Aside from Anderson conditionals, this type of cancellation takes place when participants are asked to evaluate subjunctive conditionals in TT and TF cells, where they need not go beyond the possible world offered by the displayed picture to identify the closest possible *A*-world. Through the constraint that the similarity relation is *centered*, so that the closest possible A-world to *w* is *w* itself, whenever A is true at *w*, this requirement is built into possible worlds semantics by fiat (Stalnaker, 1968; Lewis, 1973).

In the discussion sections of the individual experiments, we have already considered in detail what bearing our results have on a wide range of different accounts of indicative conditionals. Below we consider whether mental model theory would be able to account for our results concerning subjunctive conditionals in Experiment 1.

Mental Model Theory and Possible Worlds Semantics

On the revised mental model theory, conditionals are conjunctive assertions about possibilities (i.e., "*A*&*C* is possible and *A*&¬C is not possible..."). That not-*A* is possible is a shared presupposition of true and false conditionals. In the case of counterfactual conditionals, the "¬*A*&¬C" possibility acquires the status of being *a fact* and the other possibilities change status to express "counterfactual possibilities" (see Table 4). Counterfactual possibilities concern states that were once possible but did not obtain.

It is a well-known observation that participants often exhibit a biconditional interpretation of conditionals (see e.g., Goodwin & Johnson-Laird, 2018). If we consider only the standard truth table cells in Experiment 1 (TT, TF, FT, FF), then participants are found to interpret subjunctive conditionals bi-conditionally (T, F, F, T). The addition of the $\mathrm{FF}_{\mathrm{misplaced}}$ cell demonstrates, however, that it is not really a bi-conditional interpretation that is found, because the modal truth value makes a strong flip from T to F in what is an FF cell, when evaluated w.r.t. the Actual picture shown (see Figure 1).

Could the mental model account handle the results for subjunctive conditionals of Experiment 1? To be able to account for the full range of conditions (TT, TF, FT, FF, $\mathrm{FF}_{\mathrm{misplaced}}$), mental model theory would have to apply its possibility table (see Table 4) for both indicative and counterfactual conditionals. What enforces this is the constraint of counterfactuals that "¬*A*&¬*C* is a fact", which is only met in the FF and $\mathrm{FF}_{\mathrm{misplaced}}$ cells. That is to say, the subjunctive conditionals in the TT, TF, and FT cells are effectively treated as indicative conditionals, if the account is to be applicable. Under these assumptions, mental model theory is equipped to account for the modal pattern (T, F, F, T, F) in participants' responses, but problems emerge, as shown below.

A Technical Problem Concerning Subjunctives

The first problem is that mental model theory is forced to misrepresent the subjunctive conditional in the FT cell as making a (false) claim about a *real* possibility, instead of making a (false) claim about a *counterfactual* possibility (now that "¬*A*&*C*" is known as *a fact*). But treating the subjunctive conditional in the FT cell as an indicative conditional is problematic for other reasons as well. As we have seen, Goodwin and Johnson-Laird (2018) suggest that indicative conditionals are true in the false antecedent cases if the corresponding counterfactual conditionals are true. However, if we cannot evaluate the subjunctive conditional in the FT cell as a counterfactual on mental model theory (because "¬*A*&¬*C*" is *not* a fact), and we can only evaluate it as an indicative, if we already know the truth value of the corresponding counterfactual, then we have landed in a circle. It may thus prove difficult to derive a truth value in this case by following the proposed strategy.

In contrast, a selection function in possible worlds semantics can switch between the pictures more smoothly, and not misconstrue the modal status of the possibilities under evaluation, as shown in Table 7.

Table 7. Selection function applied to the subjunctives of Experiment 1

Cell	f(w_1, *A*) = w_i		w_1 = Actual P	Conditional	Response
TT	w_i = Actual P	TT	TT	Actual P > Actual object	T
TF	w_i = Actual P	TF	TF	Actual P > Hypothetical object	F
FT	w_i = Hypothetical P	TF	FT	Hypothetical P > Actual object	F
FF	w_i = Hypothetical P	TT	FF	Hypothetical P > Hypothetical object	T
FF$_{\text{mis}}$	w_i = Hypothetical P	TF	FF	Hypothetical P > Misplaced object	F

Note. "*A* > *B*" = "If *A* had been the case, then *B* would have been the case." 'Actual P' = the Actual Picture is displayed. 'Hypothetical P' = the Hypothetical Picture is displayed. 'FF$_{\text{mis}}$' = FF$_{\text{misplaced}}$. f(w_1, *A*) selects the closest possible *A*-world. For the TT, TF conditions, this is: a situation where the Actual Picture is shown. For the FT, FF, FF$_{\text{misplaced}}$ conditions, this is: an imagined situation where the Hypothetical Picture is shown. The conditional is then evaluated w.r.t. the selected world.

As seen, the selection function correctly selects the Actual Picture as the world of evaluation in the TT and TF conditions, and the Hypothetical Picture as the world of evaluation in the FT, FF, and FF$_{\text{misplaced}}$ conditions. A further problem for the mental model theory is that, strictly speaking, it is *both possible* that the Hypothetical Picture (i.e., "one of Jack's photos of a railroad station") has a warning sign on it *and possible* that it does *not* have a warning sign on it. Yet, to capture the participants' responses in Experiment 1, the second combination would have to be treated as *impossible*. It is unclear why any of the combinations of Jack's pictures would merit this description under the account by mental model theory.

It is easy to read this as the implausible suggestion that participants treat it as *epistemically impossible* that there would be no warning sign on one of Jack's photos of a railroad station. Goodwin and Johnson-Laird (2018, p. 2528) are aware of the difficulty and thus caution that they do not intend impossibility to be taken in an absolute sense but rather as that the conditional in a specific context is incompatible with a state of affairs.

However, in addition to understanding that a given conditional is *incompatible* with a specific possibility, accepting a conditional as true requires that that possibility is set aside as not pertinent in a given context. It is at this point that possible worlds semantics appeals to comparisons between which possibilities are more similar to the actual situation to allow for a *relative notion* of possibility.[64] Given the earlier encountered prototypes of Jack's photos, it would be more compatible with everything that is known to assume that one of Jack's photos of a railroad station would have had a warning sign on it than to assume that it would lack one. Most participants therefore converge on their truth evaluations of the subjunctive conditionals. But there would be nothing contradictory, or incoherent, about assuming that Jack could also have taken photos of railroad stations without warning signs. We therefore cannot exclude it across contexts, even if we treat it as a *more distant* possibility in this context. In contrast, while Goodwin and Johnson-Laird (2018) may stress that they do not mean 'impossibility' to be taken in an absolute sense, they have not, on the other hand, given the theoretical means for accounting for relative comparisons of possibility/impossibility.

Possible world semantics is thus better equipped to account for the results in Experiment 1 results concerning subjunctive conditionals. Johnson-Laird and Ragni (2019) argue against the psychological plausibility of possible worlds semantics on the grounds that it would require reasoners to consider an infinite number of possible worlds. It is, however, important to stress that participants need not consider an infinite number of possible worlds. One can in fact apply the semantics more locally based on a small number of situational models considered by the participants, as we have done in Experiments 1 and 2. When modal operators, like conditionals, are given a *global interpretation*, conditionals are assigned truth values *at all possible worlds*. But in natural-language semantics, the set of possible worlds may be restricted so that sentences containing conditionals are only assigned truth values at a small set of worlds that are deemed contextually relevant (Garson, 2013, p. 63, see further Portner, 2009). Formally, the set of possible worlds just needs to be non-empty (Garson,

[64] Formally, modal logic introduces a relative notion of relevant possibilities by introducing accessibility relations that constraint which possible worlds can be accessed from a given world for a specific modal operator (Garson, 2013; Ch. 5, 20). For this, Lewis (1973) uses spheres of similarity and Stalnaker (1968) applies selection functions.

2013, p. 93). Moreover, these situational models need not contain more structure than what is needed to address the current questions under discussion (Roberts, 1996; Stalnaker, 2019, p. 185, fn. 9).

Making Possible World Semantics Testable

Above, we outline some technical problems that arise for the revised mental model theory in attempting to account for the results for subjunctive conditionals in Experiment 1. We take this, as well as the parallel finding for indicative conditionals (see Table 3), as evidence that possible worlds semantics specifies an input-output function that fit the participants' mean response tendencies responses better than the alternatives we have looked at. However, the failure of all tested theories to account for the variances and covariances in assigned truth values to indicative conditionals across individuals suggests that there is further individual variation that is left unaccounted for by assuming any given truth table for all participants. In Experiment 2, we investigated this possibility via mixture distributions based on the truth tables used in Experiment 1. It was found that while the model implementing possible world semantics did not account for all the participants, it accounted for the majority. In addition, minorities following both the material implication and the de Finetti truth table could also be identified.

Before any conclusions can be reached about the psychological implementation, an important first step is to make the main theoretical alternatives *empirically testable* by designing new experimental tasks. Given its persistent popularity in other disciplines, in few other cases is this need arguably more urgent than for the possible worlds semantics of Stalnaker (1968) and Lewis (1973). It is to fill this gap between the types of theories considered in psychology and philosophy/linguistics that the present paper contributes.

Another way to accomplish this feat is to follow Wijnbergen-Huitink et al.'s (2015) strategy and apply notions of similarity from psychology to possible world semantics. For instance, Wijnbergen-Huitink et al. follow the approach to similarity of Tversky (1977) of counting the number of features that objects have in common. Relatedly, Pearce (1987, 1994) uses the following measure in his theory of configural learning to quantify the stimulus generalization from the stimulus (P) AB to (P′) ABC, based on their similarity, ${}_{P}S_{P'}$:

$${}_{P}S_{P'} = \frac{N_C^2}{N_P * N_{P'}}$$

In this case, P and P′ share two components ($N_C = 2$) and the number of elements of the stimuli is 2 ($N_P = 2$) and 3 ($N_{P'} = 3$), respectively. Further measures of similarity exist in spatial representations and semantic spaces based on distance measures in high-dimensional spaces (Markman, 1999, Ch. 2), structured representations (ibid., Ch. 5), models of categorization (ibid. Ch. 8), and relational representations (ibid., Ch. 10). Finally, Pearl (2009: Ch. 7) shows how to use an account of interventions in causal models to explicate the notion of similarity in Lewis (1973) and shows that it is possible to derive the same conditional logics as on Lewis' account.

Whether any of these various approaches to similarity fully captures what possible world semantics intends, we take to be an open and controversial question. In Lewis (1979), a system of weights and priorities was presented, which was supposed to explicate how to weight violations of natural laws and agreement on particular facts in the comparison of the similarity of worlds. Like Pearl (2009), this focuses more on underlying causal structure than on physical resemblance but is much broader. In linguistics and philosophy, endeavors to improve upon these qualitative criteria continue (Rescher, 2007; Ippolito, 2016; Starr, 2019). In Stalnaker (2019, Ch. 11) yet other another notion of similarity is explicated based on objective chances in branching tree representations of events unfolding over time.

Because these issues are so controversial, the strategy that we adopted in this paper was to attempt side-step these open questions by designing an experimental paradigm in which the structure of the task constraints what the most salient similar alternative is – and then to test whether participants selectively make use of this source of information when assigning truth values to indicative and subjunctive conditionals.

5.5 Conclusion

In this paper we presented a new experimental task for investigating the possible world account of subjunctive conditionals. This enabled us to present some of the first direct empirical corroboration for its truth conditions as capturing the mean response frequencies. In contrast, none of the salient alternatives, like the material implication, the de Finetti table, the Jeffrey Table, the multi-dimensional approach of Bradley (2012), nor the revised mental model theory, were able to capture these mean response frequencies in Experiment 1.

What enabled the possible worlds account to do better than the other accounts was its prediction of a bi-modal response pattern based on whether or not participants accommodated in the case of indicative conditionals with false antecedents. By incorporating this factor of whether participants accommodated the reference of the false antecedent, the possible world account was able to account for when participants evaluated indicative and subjunctive conditionals alike, and when their truth evaluations diverged, unlike the other accounts.

In Experiment 2, participants' truth evaluations of indicative conditionals were modeled as coming from a mixture distribution of competing truth tables. Through model comparisons, it was found that possible worlds semantics accounted for the truth evaluations of the largest subgroup of participants within our experimental task. At the same time, minorities of participants were classified as following competing truth tables, which could account for why the possible worlds account by itself only had a satisfactory fit for aggregate truth-value frequencies in Experiment 1 but did not account for the pattern of individual differences in the data.

References

Adams, E. (1970). Subjunctive and Indicative Conditionals. *Foundations of Language, 6*, 89-94.

Adams, E. (1975). The logic of conditionals: an application of probability to deductive logic. Dordrecht: Reidel.

Anderson, A. R. (1951). A note on subjunctive and counterfactual conditionals. *Analysis, 12*, 35–38.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 340-412.

Baratgin, J.**,** Over, D. E., & Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking & Reasoning, 19*, 308-328.

Barrouillet, P., & Lecas, J.-F. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, *67*(3), 209–253. https://doi.org/10.1016/S0010-0277(98)00037-7

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86.

Beaver, D. I., & Geurts, B. (2014). Presupposition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* (Winter 2014 Edition). Retrieved from https://plato.stanford.edu/archives/win2014/entries/presupposition/

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Birner, B. J. (2013). *Introduction to Pragmatics*. Hoboken, N.J.: John Wiley & Sons.

Blome-Tillmann, M. (2013). Conversational Implicatures (and How to Spot Them). *Philosophy Compass, 8*(2), 170–185. https://doi.org/10.1111/phc3.12003

Bradley, R. (2012). Multidimension Possible-world Semantics for Conditionals. *The Philosophical Review, 121*(4), 539-71.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P., & Vuorre, M. (2018, February 28). Ordinal Regression Models in Psychological Research: A Tutorial. Retrieved from http://doi.org/10.17605/OSF.IO/X8SWP

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press.

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Byrne, R. M. J. (2017). Counterfactual Thinking: From Logic to Morality. *Current Directions in Psychological Science*, *26*(4), 314–322. https://doi.org/10.1177/0963721417695617

Byrne, R. M. J., and Johnson-Laird, P. N. (2019). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. http://dx.doi.org/10.1037/xlm0000756

Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition*, *27*(4), 726–740. https://doi.org/10.3758/BF03211565

Cariani, F. and Rips, L. J. (2017). Conditionals, Context, and the Suppression Effect. *Cognitive Science*, *41*, 540-589.

Chemla, E,, & Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes*, *38*, 241–260.

Declerck, R., & Reed, S. (2001). *Conditionals: A Comprehensive Empirical Analysis*. Berlin: Walter de Gruyter.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen- Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology, 101*, 50– 81.

Douven, I., Elqayam, S., Singmann, H., Wijnbergen-Huitink, JV., (2020). Conditionals and inferential connections: toward a new semantics. *Thinking and Reasoning*, *26*(3), 311-351.

de Vega, M., Urrutia, M., & Riffo, B. (2007). Canceling updating in the comprehension of counterfactuals embedded in narratives. *Memory & Cognition*, *35*(6), 1410–1421. https://doi.org/10.3758/BF03193611

Edgington, D. (1995). On conditionals. *Mind*, *104*, 235-329.

Edgington, D. (2006). Conditionals. In: E.N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2008 edn.). Retrived from: http://plato.stanford.edu/archives/win2008/entries/conditionals/.

Edgington, D. (2008). I-Counterfactuals. *Proceedings of the Aristotelian Society, 108*(1), 1–21.

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift fur Psychologie / Journal of Psychology, 217*, 108–124.

Espino, O., & Byrne, R. M. J. (2018). Thinking About the Opposite of What Is Said: Counterfactual Conditionals and Symbolic or Alternate Simulations of Negation. *Cognitive Science, 42*(8), 2459–2501. https://doi.org/10.1111/cogs.12677

Evans, J. S. B., & Newstead, S. E. (1977). Language and reasoning: A study of temporal factors. *Cognition, 5*(3), 265-283.

Evans, J. St. B. T., Handley, S. J., Neilens, H., and Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition*, *35*(7), 1772-84.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, England: Lawrence Erlbaum Associates.

Evans, J. St. B. T. & Over, D. (2004). *If*. Oxford: Oxford University Press.

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*(3), 609–626.

Fillenbaum, S. (1974). Information amplified: Memory for counterfactual conditionals. *Journal of Experimental Psychology*, *102*(1), 44–49. https://doi.org/10.1037/h0035693

Foley, R. (1992). The Epistemology of Belief and the Epistemology of Degrees of Belief. *American Philosophical Quarterly*, 29, 111–121.

Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How People Interpret Conditionals: Shifts Toward the Conditional Event. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 635-648.

Garson, J. W. (2013). *Modal Logic for Philosophers.* (Second Edition) New York: Cambridge University Pres.

Gazdar, G. (1979). *Pragmatics: Implicature, Presuppositions, and Logical Form.* New York: Academic Press.

Gelman, A., Carlin, J. B., Stern, H. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis* (Third Edition). Boca Raton: Chapman & Hall/CRC.

Geurts, B., Kissine, M., and van Tiel, B. (2020). Pragmatic reasoning in autism. In: Morsanyi, K. and Byrne, R. (Eds.), *Thinking, reasoning and decision making in autism* (pp. 113-134). London: Routledge.

Goodwin, G. P., and Johnson-Laird, P. N. (2018). The Truth of Conditional Assertions. *Cognitive Science*, *42*, 2502-2533.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grünwald, P. (2007). The minimum description length principle. Cambridge, MA: MIT Press.

Heck, D. W., Arnold, N. R., and Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modelling. *Behavior Research Methods, 50*, 264-84.

Horn, L. R. (1984). Towards a New Taxonomy for Pragmatic Inference: Q-based and R-based Implicature. In: Schiffrin, D. (Eds), *Georgetown University Round Table on Languages and Linguistics 1984 (pp. 11-42).* Washington, DC: Georgetown University Press.

Huang, Y. (2007). *Pragmatics*. Oxford, England: Oxford University Press.

Huddleston, R. (2002). Content clauses and reported speech. In R. Huddleston & G. K. Pullum (Eds.), *The Cambridge Grammar of the English Language* (pp. 947–1030). Cambridge, England: Cambridge University Press.

Iatridou, S. (2000). The Grammatical Ingredients of Counterfactuality. *Linguistic Inquiry*, *31*(2), 231–270. https://doi.org/10.1162/002438900554352

Ippolito, M. (2003). Presuppositions and Implicatures in Counterfactuals. *Natural Language Semantics*, *11*(2), 145–186. https://doi.org/10.1023/A:1024411924818

Ippolito, M. (2016). How Similar Is Similar Enough? *Semantics and Pragmatics*, *9*(6), 1–60.

Johnson-Laird, P.N., & Tagart, J. (1969. How implication is understood. *American Journal of Psychology, 2*, 367-373.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. and Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review 109*, 646-678.

Johnson-Laird, P.N., Khemlani, S. S., and Goodwin, G. P. (2015). Response to Baratgin et al.: Mental Models Integrate Probability and Deduction. *Trends in Cognitive Sciences, 19*(10), 548–549.

Johnson-Laird, P. N. and Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition, 193*, 103950. doi: 10.1016/j.cognition.2019.04.019

Johnson-Laird, P. N., & Savary, F. (1995). How to make the impossible seem probabile. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 381–384). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kadmon, N. (2001). *Formal Pragmatics*. Malden, MA: Blackwell Publishers.

Karabatsos, G. (2005). The exchangeable multinominal model as an approach for testing axioms of choice and measurement. *Journal of Mathematical Psychology, 49*, 51-69.

Karttunen, L., & Peters, S. (1979). Conventional implicature. In C.-K. Oh & D. Dinneen (Eds.), *Presupposition* (pp. 1–56). New York, NY: Academic Press.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science, 42*(6), 1887–1924. https://doi.org/10.1111/cogs.12634

Klauer, K. C. (2010). Hierarchical Multinominal Processing Tree Models: A Latent-Trait Approach. *Psychometrika, 75*(1), 70-98.

Klauer, K. C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in Multinominal Processing Tree Models: An extension of Knapp & Batchelder (2004). *Journal of Mathematical Psychology*, 64, 1-5.

Kratzer, A. (2012). *Modals and Conditionals: New and Revised Perspectives*. New York: Oxford University Press.

Krzyżanowska, K., Collins, P. J., & Hahn, U. (2020). *True clauses and false connections*. Manuscript submitted for publication.

Kutschera, F. (1974). *Indicative Conditionals. Theoretical linguistics, 1*, 257-269.

Lassiter, D. (2017). Probabilistic language in indicative and counterfactual conditionals. *Proceedings of SALT, 27*, 525-546.

Leahy, B. (2011). Presuppositions and Antipresuppositions in Conditionals. *Semantics and Linguistic Theory*, *21*, 257. https://doi.org/10.3765/salt.v21i0.2613

Leahy, B. (2018). Counterfactual antecedent falsity and the epistemic sensitivity of counterfactuals. *Philosophical Studies, 175*, 45-69.

Lecas, J.-F., & Barrouillet, P. (1999). Understanding conditional rules in childhood and adolescence: A mental models approach. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *18*(3), 363–396.

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, England: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*, 297-315.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*, 339-59.

Markman, A. B. (1999). *Knowledge Representations*. New York: Psychology Press.

Mittwoch, A., Huddleston, R., & Collins, P. (2002). The clause: Adjuncts. In R. Huddleston & G. K. Pullum (Eds.), *The Cambridge Grammar of the English Language* (pp. 663–784). Cambridge, England: Cambridge University Press.

Nickerson, R. (2015). *Conditional Reasoning. The Unruly Syntactics, Semantics, Thematics, and Pragmatics of "If"*. Oxford: Oxford University Press.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Over, D. E., & Baratgin, J. (2017). The “defective” truth table: Its past, present, and future. In N. Galbraith, E. Lucas, & D. E. Over (Eds.), *The Thinking Mind: A Festschrift for Ken Manktelow* (pp. 15-28). Abingdon, UK: Routledge.

Over, D. E. (2020). The development of the new paradigm in the psychology of reasoning. In S. Elqayam, I. Douven, J. St. B. T. Evans, & N. Cruz (Eds.), Logic and uncertainty in the human mind (pp. 243-263). Abingdon: Routledge.

Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In Linden J. Ball and Valerie A. Thompson (Eds.), International handbook of thinking and reasoning (pp. 434-450). Hove: Psychology Press.

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review*, *94*, 61-73.

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*, 587-607.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2th Ed.). Cambridge:

Cambridge University Press.

Pearl, J. (2013). Structural Counterfactuals: A Brief Introduction. *Cognitive Science, 37*(6), 977-985.

Pfeifer, N. & Tulkki, L. (2017). Conditionals, Counterfactuals, and Rational Reasoning: An Experimental Study of Basic Principles. *Minds and Machines, 27*(1), 119-165.

Plummer, M. (2019). rjags: Bayesian Graphical Models using MCMC. R package version 4-10. https://CRAN.R-project.org/package=rjags

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking and Reasoning, 16*(3), 172–197.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Potts, C. (2005). *The Logic of Conventional Implicatures.* Oxford: Oxford University Press.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *The Quarterly Journal of Experimental Psychology, 63*(9), 1716–1739. https://doi.org/10.1080/17470210903536902

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2018). The Relation Between Factual and Counterfactual Conditionals. *Cognitive Science*, *42*(7), 2205–2228. https://doi.org/10.1111/cogs.12663

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Read, T., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Reips, U. D. (2002). *Standards for Internet-based experimenting. Experimental Psychology, 49* (4), *243*-256.

Rescher, N. (2007). *Conditionals*. Cambridge, Massachusetts: The MIT Press.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339.

Rips, L. J. (2010). Two Causal Theories of Counterfactual Conditionals. *Cognitive Science, 34*(2), 175-221.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics, 5*(6), 1-69.

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In: Schwarz (Ed.). *Experimental perspectives on presuppositions. Studies in Theoretical Psycholinguistics* (Vol. 45, pp. 215-240). Springer, Cham.

Santamaría, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and Semifactual Conditionals Prime Alternative Possibilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1149–1154. https://doi.org/10.1037/0278-7393.31.5.1149

Schroyens, W. (2010). A meta-analytic review of thinking about what is true, possible, and irrelevant in reasoning from or reasoning about conditional propositions. *European Journal of Cognitive Psychology*, 22 (6), 897-921.

Skovgaard-Olsen, N. (2021). Relevance and Conditionals: A Synopsis. In: Elqayam, S., Douven, I., Cruz, N., Evans, J. (Eds.), *Festschrift for David Over*. Routledge (pp. 192-206).

Skovgaard-Olsen, N. and Collins, P. (2021). Indicatives, Subjunctives, and the Falsity of the Antecedent. *Cognitive Science*. *Cognitive Science, 45*(11), https://doi.org/10.1111/cogs.13058.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., and Klauer, K. C. (2019). Cancellation, Negation, and Rejection. *Cognitive Psychology, 108*, 42-71.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., and Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of 'And', 'But', 'Therefore', and 'If Then'. *Thinking and Reasoning, 23*(4), 449-482.

Skovgaard-Olsen, N., Stephan, S., and Waldmann, M. (2021). Conditionals and the Hierarchy of Causal Queries. *Journal of Experimental Psychology: General*, *150*(12), 2472-2505.

Sloman, S. A. and Lagnado, D. A. (2005). Do We "do"? *Cognitve Science*, *29*, 5-39.

Smith, E. A. and Hall, K-C. 2011. Projection diversity: Experimental evidence. Workshop on Projective Meaning at ESLLI 2011.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183.

Soames, S. (1982). How presuppositions are inherited: a solution to the Projection Problem. *Linguistic Inquiry*, *13*, 483-545.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition* (Second Edition). Oxford, England: Blackwell.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, *37*, 1074–1106.

Stalnaker, R. C. (1968). A Theory of Conditionals. In: Rescher, N. (Eds.), *Studies in Logical Theory (pp. 98-112)*. Oxford: Basil Blackwell.

Stalnaker, R. (1972). Pragmatics. In D. Davidson & G. Harman (Eds.), *Semantics of natural language* (pp. 389–408). Reidel: Dordrecht.

Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz & P. Unger (Eds.), *Semantics and philosophy* (pp. 197–214). New York, NY: New York University Press.

Stalnaker, R. C. (1975). Indicative conditionals. *Philosophia, 5*(3), 269-286.

Stalnaker, R. (2011). Conditional propositions and conditional assertions. In: Egan, A. and Weatherson, B. (Eds.), *Epistemic Modality* (pp. 227-48). Oxford: Oxford University Press.

Stalnaker, R. (2014). *Context*. Oxford: Oxford University Press.

Stalnaker, R. (2019). *Knowledge and Conditionals*. Oxford: Oxford University Press.

Starr, W. B. (2014). A Uniform Theory of Conditionals. *Journal of Philosophical Logic*, *43*(6), 1019-1064.

Starr, W. (2019). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). Retrieved from forthcoming URL = <https://plato.stanford.edu/archives/fall2019/entries/counterfactuals/>.

Stewart, A. J., Haigh, M., & Kidd, E. (2009). An investigation into the online processing of counterfactual and indicative conditionals. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2113–2125.

Thompson, V. A., & Byrne, R. M. J. (2002). Reasoning counterfactually: Making inferences about things that didn’t happen. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 1154–1170.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

van Wijnbergen-Huitink, Elqayam, S. and Over, D. E. (2015). The probability of iterated conditionals. *Cognitive Science, 39*(4), 788-803.

von Fintel, K. (1997). The presupposition of subjunctive conditionals. In *MIT working papers in linguistics 25* (pp. 29–44). Cambridge, MA: MIT Working Papers in Linguistics.

von Fintel, K. (2012). Subjunctive conditionals. In G. Russell & D. Graff Fara (Eds.), *The Routledge companion to the philosophy of language* (pp. 466–477). New York, NY: Routledge.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev, 25*, 35-57.

Williamson, T. (2020). *Suppose and Tell. The Semantics and Heuristics of Conditionals.* Oxford: Oxford University Press.

Appendix A: Accommodation

The figures below illustrate the bimodal pattern in the truth evaluation of indicative conditionals, which arose depending on whether participants accommodated the reference of the antecedent to refer to the Hypothetical Picture across three trials.

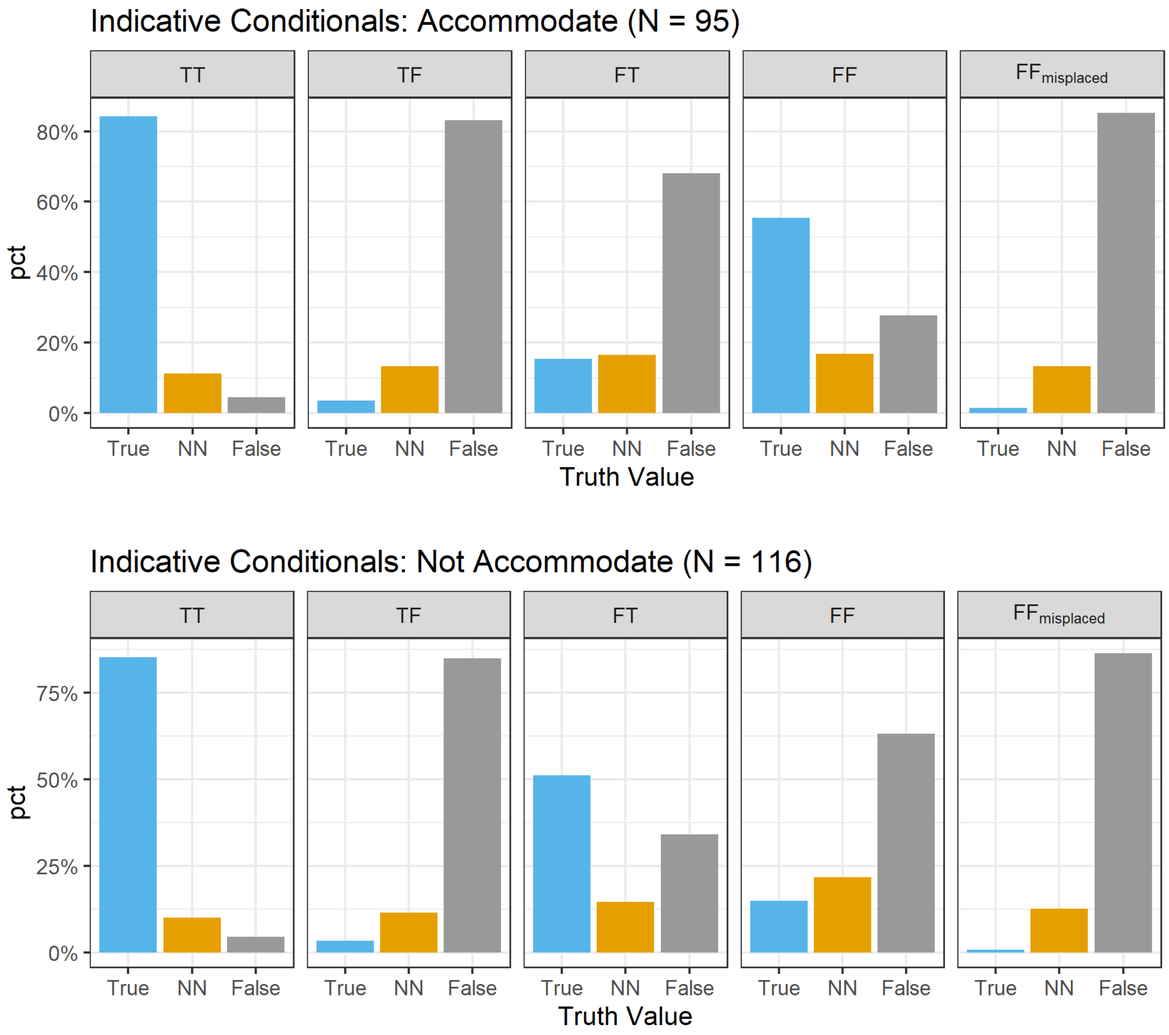


Figure A1. Truth Tables of Indicative Conditionals as a Function of Accommodation. 'FF' = a conditional that is False False w.r.t. the Actual Picture shown but True True w.r.t. the salient Hypothetical Picture; '$\text{FF}_{\text{misplaced}}$' = a conditional that is False False w.r.t. the Actual Picture shown but True False w.r.t. the salient Hypothetical Picture.

As shown, for the 95 participants who did accommodate, the indicative conditionals were evaluated corresponding to the modal values for subjunctive conditionals (see Figure 1). In contrast, the modal truth values flipped in the FT and FF cells for the 116 participants who chose not to accommodate. When the truth evaluations for these two groups are merged, the aggregate pattern for indicative conditionals displayed in Figure 1 arises.

As a result, it is found that participants' decisions of whether to accommodate the antecedent across three trials are predictive of the qualitative differences in truth evaluations of indicative conditionals displayed in Figure A1.

Appendix B: Hierarchical MPT Models

Multinominal processing trees models model response frequencies for a set of mutually exclusive categorical response outcomes (Riefer & Batchelder, 1988). In this paper, multinomial processing tree models are used to analyze the observed response frequencies in truth value assignments (T, F, NA). To implement the various truth tables, inequality constraints are introduced for the MPT parameters following the analytic approach of Klauer et al. (2015, Appendix A). For example, if response "T" is predicted to be the modal response, the response probabilities $\eta_1, \eta_{2,}$ and η_3 for the responses "T", "F", and "NA", respectively, should satisfy the inequalities $\eta_1 \geq \eta_2$ and $\eta_1 \geq \eta_3$. This is guaranteed by parameterizing the three response probabilities as follows (Eq. 1):

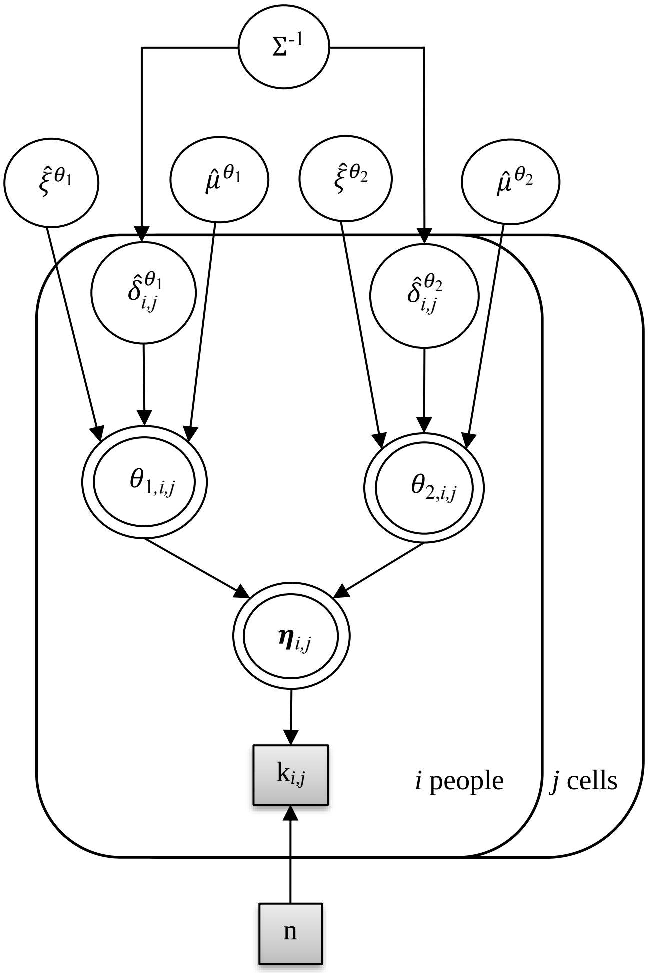
$$\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix} = (1-\theta_1)(1-\theta_2)\begin{pmatrix}1\\0\\0\end{pmatrix} + \theta_1(1-\theta_2)\begin{pmatrix}1/2\\1/2\\0\end{pmatrix} + (1-\theta_1)\theta_2\begin{pmatrix}1/2\\0\\1/2\end{pmatrix} + \theta_1\theta_2\begin{pmatrix}1/3\\1/3\\1/3\end{pmatrix}$$

Instead of aggregating the categorical outcomes across participants, the hierarchical latent trait approach of Klauer (2010) is followed in Experiment 1, which adds a hierarchical structure in which the participants' MPT parameters are constrained to be samples from a population-level probability distribution.

On this approach, a probit link function is used to transform MPT parameters (representing probabilities between 0 and 1) to the real line, $\Phi^{-1}(\theta)$. The transformed parameters are then modelled via a multivariate normal distribution while estimating mean, μ, and covariance matrix, Σ, from the data. The advantage of this approach is that heterogeneity in parameter estimates across participants and correlations among MPT parameters can be accommodated while allowing for partial aggregation of statistical information across participants in the posterior parameters of the multivariate normal distribution (Klauer, 2010). Accordingly, for each participant, i, the probit-transformed parameters are additively decomposed into a group mean, μ, and a random effect, $\Phi^{-1}(\theta) = \mu + \delta_i$.

In Experiment 1, a set of hierarchical multinominal models following this approach with different order constraints implementing competing truth tables were contrasted in a model-comparison exercise. Table 1B illustrates the general form of these models.

Table 1B. Hierarchical Latent Trait MPT Model with Inequality Constraints



$$\hat{\mu}^{\theta_1}, \hat{\mu}^{\theta_2} \sim \text{Gaussian}(0,1)$$
$$\hat{\xi}^{\theta_1}, \hat{\xi}^{\theta_2} \sim \text{Uniform}(0,10)$$
$$\Sigma^{-1} \sim \text{Wishart}(\mathbf{I}_{10\times 10}, 11)$$
$$\hat{\delta}^{\theta_1}_{i,j}, \hat{\delta}^{\theta_2}_{i,j} \sim \text{MvGaussian}(\mathbf{0}, \Sigma^{-1})$$
$$\theta_{1,i,j} \leftarrow \Phi\left(\hat{\mu}^{\theta_1} + \hat{\xi}^{\theta_1}\hat{\delta}^{\theta_1}_{i,j}\right)$$
$$\theta_{2,i,j} \leftarrow \Phi\left(\hat{\mu}^{\theta_2} + \hat{\xi}^{\theta_2}\hat{\delta}^{\theta_2}_{i,j}\right)$$
$$\eta_{1,i,j} \leftarrow \ldots$$
$$\eta_{2,i,j} \leftarrow \ldots$$
$$\eta_{3,i,j} \leftarrow \ldots$$
$$\mathrm{k}_{i,j} \sim \text{Multinomial}\left(\boldsymbol{\eta}_{i,j}, n\right)$$

Note. $\boldsymbol{\eta}_{i,j}$ is a vector of $\eta_{1,i,j}$, $\eta_{2,i,j}$, and $\eta_{3,i,j}$. See Eq. 1 above for the parameterization of these parameters in terms of $\theta_{1,i,j}$ and $\theta_{2,i,j}$. The *j* cells correspond to {TT, TF, FT, FF, $\text{FF}_{\text{misplaced}}$}. Since there are five truth table cells, with three categorical responses parameterized in terms of 2 theta parameters each, there are 5 × 2 = 10 theta parameters in total and 10+1 degrees of freedom of the inverse Wishart distribution together with a 10×10 identity matrix.

In Experiment 2, these hierarchical multinominal processing trees were extended by representing several competing truth tables as mixture components of a mixture distribution. A categorical variable with a Dirichlet prior assigned participants to each of the different mixture components and thereby permitted the assignment of individual truth tables to participants to examine individual differences.

In Experiment 2, three implementations of the hierarchical mixture distributions were contrasted in a model comparison (containing 3, 4, or 5 mixture components, respectively), which used a scaled inverse-Wishart prior for modelling the covariance matrix, following the the hierarchical latent trait approach of Klauer (2010) from Experiment 1, illustrated in Table 1B. In addition, a fourth model with 4 mixture components were fitted which followed the approach of Smith and Batchelder (2010) of using independent beta distributions for the different MPT parameters. Of the two, the latter assumes independence of MPT parameters in the hierarchical prior distribution while the former explicitly accommodates correlations among the MPT parameters already in the prior distribution.

In addition, all models featured a saturated mixture component to filter out noisy respondents, which did not fit any of the models in virtue of violating the shared predictions of ‘True’ and ‘False’ in the TT and TF cells.

The models were fitted in a Bayesian framework through the Gibbs sampler implemented in JAGS (Plummer, 2019), which estimates the posterior distributions of model parameters by means of Monte Carlo-Markov chains.

Chapter 6:
Conditionals and the Hierarchy of Causal Queries[65]

Niels Skovgaard-Olsen,
Simon Stephan,
Michael R. Waldmann

Recent studies indicate that indicative conditionals like "If people wear masks, the spread of Covid-19 will be diminished" require a probabilistic dependency between their antecedents and consequents to be acceptable (Skovgaard-Olsen et al., 2016). But it is easy to make the slip from this claim to the thesis that indicative conditionals are acceptable only if this probabilistic dependency results from a causal relation between antecedent and consequent. According to Pearl (2009), understanding a causal relation involves multiple, hierarchically organized conceptual dimensions: *prediction*, *intervention*, and *counterfactual dependence*. In a series of experiments, we test the hypothesis that these conceptual dimensions are differentially encoded in indicative and counterfactual conditionals. If this hypothesis holds, then there are limits as to how much of a causal relation is captured by indicative conditionals alone. Our results show that the acceptance of indicative and counterfactual conditionals can become dissociated. Furthermore, it is found that the acceptance of both is needed for accepting a causal relation between two co-occurring events. The implications that these findings have for the hypothesis above, and for recent debates at the intersection of the psychology of reasoning and causal judgment, are critically discussed. Our findings are consistent with viewing indicative conditionals as answering *predictive queries*

[65] This chapter has been published as:

Skovgaard-Olsen, N., Stephan, S., and Waldmann, M. (2021). Conditionals and the Hierarchy of Causal Queries. *Journal of Experimental Psychology: General*, *150*(12), 2472–2505.

Authors' Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com). Supplementary Materials: https://osf.io/fa9rj/

requiring *evidential relevance* (even in the absence of direct causal relations). Counterfactual conditionals in contrast target *causal relevance*, specifically. Finally, we discuss the implications our results have for the yet unsolved question of how reasoners succeed in constructing causal models from verbal descriptions.

6.1 Introduction[66]

There is wide agreement that conditional statements of the type "if A, then C" play a central role in reasoning and argumentation (where 'A' refers to the antecedent and 'C' to the consequent). For instance, in 2019 much political discussion centered around the statement "If Trump is impeached, then it will affect the 2020 election". At the same time, conditionals pose many unsolved theoretical problems that have kept researchers busy, despite continuous, multidisciplinary efforts (Bennett, 2003; Kern-Isberner, 2001; Kratzer, 2012; Nickerson, 2015; Oaksford & Chater, 2010a; Spohn, 2013).

One of the reasons why conditionals are thought to be so central in our cognitive lives is due to their relationship with causal knowledge (Oaksford & Chater, 2010b). The linguistic encoding of knowledge about causal relations in conditionals plays a vital role for the cultural transfer of causal knowledge across generations. For causal knowledge about objects that are not in our immediate vicinity, we rely on culturally transferred causal knowledge. The same goes for objects that are governed by mechanisms, which we do not fully understand, like artifacts designed by engineers. In addition, the acquisition of causal knowledge through observed covariances and interventions dealing with the objects that *are* in our direct vicinity is often guided by linguistically acquired causal schemes (Gopnik et al., 2004). Various authors have emphasized that probably most of our causal knowledge comes through this linguistic source (e.g., Pearl, 2009, Ch. 7). But according to Danks (2014, Ch. 4), it is also the one that is the least investigated empirically.

The relationship between conditionals and causal relations has, however, been the focus of much theoretical discussion. The importance of this issue is highlighted by counterfactual approaches to causation coming from philosophy (Goodman, 1947; Lewis, 1973; Collins, Hall, & Paul 2004), computer science (Pearl, 2009), and statistics (Morgan & Winship, 2018; VanderWeele, 2015). Recently, various authors in psychology and philosophy

[66] **Acknowledgement:** We would like to thank Dominik Glandorf, Louisa Reins, and Maike Holland-Letz for their help in coding responses and setting up experiments. We also thank audiences at talks at London Reasoning Workshop (2019), EuroCogSci (2019), Regensburg University (2020), and the Reviewers and our Editor, Pierre Barrouillet, for valuable feedback.

have also made a case for causal interpretations of indicative conditionals (e.g. Oaksford & Chater, 2017; Andreas & Günther, 2018; van Rooij & Schulz, 2019; Vandenburgh, 2020).

In this paper, we investigate whether indicative conditionals by themselves suffice to express causal relations or whether there are aspects of causal relations that are not captured by indicatives.[67] We will rely on Pearl's (2009) theory of causality and his idea of *a hierarchy of causal queries*. Through our experiments, we present new evidence in support of this framework and investigate its relations to natural language conditionals. Before we turn to our research questions, we first sketch some recent developments in the psychology of reasoning, which have kindled a renewed debate about the causal interpretation of indicative conditionals. Secondly, we outline Pearl's theory of a hierarchy of causal queries and discuss its critical potential vis-à-vis this debate.

Indicative Conditionals and Probabilities

Building on the work of Adams (1975), Edgington (1995), and Bennett (2003), psychologists have found support for the hypothesis that:

$$[\text{Eq1.}] \qquad P(\text{if } A, \text{ then } C) = P(C|A)$$

which goes by the name of "the Equation" or "the conditional probability hypothesis" (Evans, Handley, & Over, 2003; Oberauer & Wilhelm, 2003; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Pfeifer & Kleiter, 2009). Recently, these results were challenged, however. It has been found that the relationship between P(if A, then C) and P(C|A) is moderated by *relevance effects* of the probabilistic dependency between A and C (Skovgaard-Olsen, Collins, et al., 2019; Skovgaard-Olsen, Kellen, et al., 2017; Skovgaard-Olsen, Singmann, & Klauer, 2016; Vidal & Baratgin, 2017). This type of probabilistic dependency can be captured by ΔP as a measure of the extent to which A changes the probability of C:

$$[\text{Eq2.}] \qquad \Delta P = P(C|A) - P(C|\neg A)$$

These studies have found that in the case of Positive Relevance, ($\Delta P > 0$), the conditional probability remained a good predictor of both the acceptance and probability of

[67] As a short-form, we refer to indicative conditionals, like "If A, then C", as 'indicatives', and to counterfactual conditionals, like "If A had not been the case, then C would not have been the case", as 'counterfactuals'. Our focus will be on paradigmatic cases of indicative conditionals, like the examples provided in the main text. Other controversial examples like non-interference conditionals ("If Trump won the 2020-election, then pigs can fly!") are not treated here but see Douven (2016) and Skovgaard-Olsen (2016) for further discussion.

indicative conditionals. An example would be “If Paul pushes down the gas pedal, then the car will speed up” in the context of a scenario describing Paul driving in his car and running late for work. For cases of Negative Relevance (ΔP < 0) and Irrelevance (ΔP = 0) this relationship was disrupted, however. Two examples would be “If Paul pushes down the gas pedal, then the car will slow down” (Negative Relevance) and “If Paul is wearing a shirt, then his car will suddenly break down” (Irrelevance).

These findings suggest that participants tend to view indicative conditionals as defective if their antecedents fail to raise the probability of their consequents. In such cases, their antecedents fail to provide a reason *for* the consequent (Douven, 2016; Krzyżanowska, Collins, et al., 2017; Skovgaard-Olsen, 2016; Spohn, 2013). Drawing on the literature on confirmation measures, the notion of A being a reason *for* or *against* C is here explicated in terms of its *evidential relevance*, or the difference in degrees of beliefs that A makes to C (Spohn 2012, Ch. 6). If A raises the probability of C (ΔP > 0), then A is said to be a reason *for* C, or *positively relevant* to C. If A lowers the probability of C (ΔP < 0), then A is said to be a reason *against* C, or *negatively relevant* to C. If A leaves the probability of C unchanged (ΔP = 0), then A is said to be *irrelevant* to C, or neither a reason *for* nor *against* C. Indicative conditionals are said to express such qualitative reason relation assessments on this account (Brandom, 1994; Spohn, 2013; Skovgaard-Olsen, 2016; see also Rott, 1986; Krzyżanowska, Wenmackers, et al., 2013; Douven, 2016). Throughout the paper, we will measure qualitative assessments of the extent to which A is a reason for/against C on an ordinal scale and refer to them as ‘ordinal reason relation assessments’.

As a psychological construct, it is possible that multiple factors influence the assessment of relevance and reason relations including topical relevance, processing effort, and goals in a dialogue (Walton, 2004; Wilson & Sperber, 2004). Potentially, such factors influence the categorization of variables as *capable* or *incapable* of affecting the probability of the consequent. Variables that are categorized as *incapable* get ignored. This makes it seem defective to find such variables in the antecedent of conditionals, where one expects to find a *reason for* the consequent (Skovgaard-Olsen, Collins et al., 2019). As a measure of the cognitive effects of a variable, we rely on the notion of probabilistic difference-making from above but note that there is a discussion with mixed evidence concerning further factors influencing the perceived relevance.[68] The data pattern described above constitutes the Relevance Effect as an interaction effect (see Figure 1).

[68] See e.g. Cruz et al. (2016), Skovgaard-Olsen, Singmann et al. (2017, supplementary materials), Vidal and Baratgin (2017), Krzyżanowska et al. (2017).

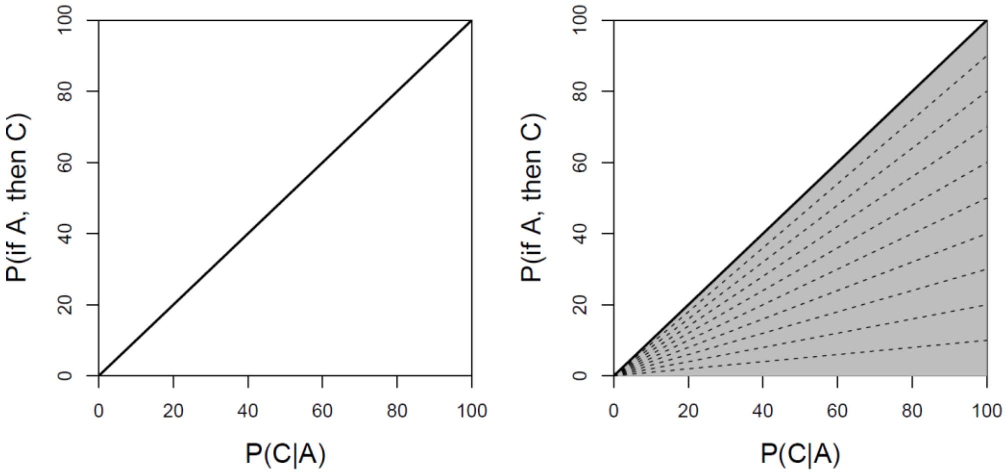


Figure 1. The left panel illustrates relationship predicted by [Eq1.]. The right panel illustrates the Relevance Effect, i.e. the moderation of the slope by relevance, in case of irrelevance ($\Delta P = 0$) or negative relevance ($\Delta P < 0$), after Skovgaard-Olsen, Kellen et al. (2019).

Accounts differ on whether this finding is to be given a semantic or pragmatic interpretation (see, e.g., Skovgaard-Olsen, Collins, et al., 2019 for a review), but here we focus on a different issue. It has recently been suggested (e.g. in Oasksford & Chater, 2020a, 2020b; van Rooij & Schulz, 2019) that relevance effects of this kind need to be given a causal interpretation. One of the goals of the present paper is to systematically explore this link through a series of experiments.

As we will explain further below, these experiments have a bearing on whether (1) P(C|A) is a good predictor of P(if A, then C) as predicted by [Eq1.] (Evans & Over, 2004; Oaksford & Chater, 2017), (2) whether a causal interpretation (van Rooij & Schulz, 2019; Oaksford & Chater, 2020a, 2020b) or (3) an evidential relevance interpretation of P(if A, then C) is needed (Skovgaard-Olsen, Singmann, & Klauer, 2016). According to Evans and Over (2004), people assess P(C|A) via the Ramsey Test:

> RAMSEY TEST: to evaluate 'if A, then C' add the antecedent (i.e. A) to the set of background beliefs, make minimal adjustments to secure consistency, and evaluate the consequent (i.e. C) on the basis of this temporarily augmented set.

Using the Ramsey Test as a basis of explicating the relationship between conditionals and suppositional reasoning has been influential in at least three competing research programs in logic (Horacio, 2007). However, in and of itself it is an abstract description of a mental algorithm which needs to be fleshed out in terms of psychological processes to be of use for cognitive scientists. As Over et al. (2007) have noted:

> Explaining how the Ramsey Test is actually implemented—by means of deduction, induction, heuristics, causal models, and other processes—is a major challenge, in our view, in the psychology of reasoning. (p. 63)

In the past decade, psychologists have made extensive use of the Ramsey Test (for a review, see Oaksford & Chater, 2020a). But the fundamental problem that Over et al. (2007) pointed to remains. Resolving this issue is important, because [Eq1.] and the abovementioned probabilistic view on conditionals has not just been taken to be one view on conditional reasoning among others. Rather, it has been treated as "one of the defining features of what has come to be referred to as the *new paradigm* in cognitive psychology" (Nickerson, 2015, p. 199) and been said to be "at the heart of the probabilistic *new paradigm* in reasoning" (Oaksford & Chater, 2017, p. 330; see also Vance & Oaksford, 2020).

One of the processes for implementing the Ramsey Test that Over et al. (2007) consider is the use of *causal models*. In line with this, Fernbach, Darlow, et al. (2011) and others have argued that causal beliefs are used as a guide for estimating subjective probabilities. The notion that conditional probabilities are assessed based on causal models via the Ramsey Test is interesting. If it can be corroborated, then this would have implications for which of the previously mentioned interpretations relating P(C|A) and P(if A, then C) is correct. For if the conditional probabilities estimated via the Ramsey Test were to rely on causal models, then P(C|A) would not be independent of a causal interpretation. In that case, P(if A, then C) would also not be independent of causal considerations given [Eq1.].

In addition, recent work on causal power suggests another possible connection between indicative conditionals and causality, which we will now turn to, because it will figure centrally in our later experiments.

Causal Power and Alternative Causes

On Cheng's (1997) account of causal power, the generative power of a cause to produce its effect is explicated by a scaled version of ΔP, where the causal contribution of alternative causes is shielded off:

[Eq3.] $$W_{\text{Cause}} = \frac{\Delta P}{1\text{-}P(\text{effect}|\neg\text{cause})} \text{ , } \Delta P = P(\text{effect}|\text{cause}) - P(\text{effect}|\neg\text{cause})$$

Causal power (W_{Cause}) is here understood as the probability with which a target cause generates its effect[69] independently of alternative causes: P(effect|cause,¬alternatives). [Eq3.] measures this quantity by determining how much the candidate cause contributes to raising the probability of the effect, while bracketing the influence of alternative causes. Following Glymour (2001), causal power has been used to parameterize Bayes nets (see, e.g., Griffiths & Tenenbaum, 2005; Fernbach, Darlow, et al., 2010, 2011; Fernbach & Erb 2013; Cummins, 2014; Meder, Mayrhofer, et al., 2014; Aßfalg & Klauer, 2019; Stephan & Waldmann, 2018), as illustrated in Figure 2:

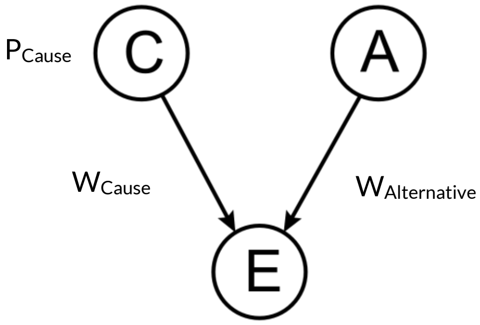


Figure 2. Common-effects Bayes Net, parameterized by the base-rate (P_{Cause}) of the cause, C, its causal power (W_{Cause}), and the combined base-rate and causal power ($W_{Alternatives}$) of the alternative cause(s), A. 'E' = effect.

Here 'C' refers to the cause and 'A' refers to alternative causes. Throughout this paper, we follow, however, the convention of using 'A' and 'C' to refer to the antecedent and consequent of conditionals, whether or not they are related as cause and effect. Based on this parametrization and other assumptions (discussed in Luhmann & Ahn, 2005), conditional probabilities have been explicated as follows, with 'W' representing the causal powers of the respective causes:

$$[\text{Eq4.}] \qquad P(\text{effect}|\text{cause}) = W_{cause} + W_{alternative} - W_{cause} * W_{alternative}$$

Notice how conditional probabilities are here explicated in terms of causal power parameters, which in turn are defined via conditional probabilities. There is accordingly a choice as to which of these constructs (i.e. conditional, subjective degrees of belief or mental representations of causal powers) is to be treated as psychologically primitive. For example, for Cheng (1997) causal powers represent latent, causal capacities of distal objects. On this view, the relative frequencies encoded in conditional probabilities are merely the manifestations of these latent capacities. But this is not the only position possible and the

[69] For preventive causes, a separate equation was given by Cheng (1997), which we return to in Experiment 1 (see [Eq5.]).

answer to the question of psychological primacy will have repercussions for the relationships between conditionals, conditional probabilities, and causality.

Oaksford and Chater (2017) have suggested that a causal interpretation of indicative conditionals can be combined with work in probabilistic treatments of conditionals based on the Ramsey Test (e.g., Adams, 1975; Edgington, 1995; Bennett, 2003; Evans & Over, 2004; Oaksford & Chater, 2007). Oaksford and Chater (2017) do this by combining the thesis P(if A, then C) = P(C|A) [Eq1.] with a causal power explication of conditional probabilities (see [Eq4.]). Making this move allows Oaksford and Chater (2017) to emphasize that there is an inferential dependency between antecedents and consequents of indicative conditionals (in line with, e.g., Douven, 2016; Krzyżanowska, Collins, et al., 2017; Skovgaard-Olsen, Singmann, et al., 2016; Spohn, 2013). At the same time, it allows Oaksford and Chater (2017) to build on the work on probability logic of Adams (1975), which has been applied to the psychology of reasoning (e.g., in Evans & Over 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009).

One challenge to this account, however, is that the Relevance Effect (Skovgaard-Olsen, Singmann, et al., 2016) identifies boundary conditions on P(C|A) as a predictor of P(if A, then C). As a consequence, if probabilistic dependency is factored into the account through a causal power explication of conditional probabilities, then we are left without an account of why relevance moderates the relationship between P(C|A) and P(if A, then C) in violation of [Eq1.]. The interaction effect depicted in Figure 1 shows that P(if A, then C) can vary due to the influence of relevance even when P(C|A) is held constant.

Accordingly, Oaksford and Chater (2020b) discuss the different possibility where the Relevance Effect is itself an indicator of a causal interpretation of indicative conditionals. But this amounts to abandoning [Eq1.] in its full generality.

As noted by van Rooij and Schulz (2019), there is, however, also a different possibility for interpreting the relationship between conditional probabilities, causal power, and P(if A, then C). The general account relies on interpreting the acceptability of indicative conditionals in terms of causal power. But by introducing this conjecture, van Rooij and Schulz rely on the auxiliary hypothesis that participants tend to ignore alternative causes. The motivation for this auxiliary hypothesis is that the equation for causal power [Eq3.] shows that causal power coincides with the conditional probability of the effect given the cause when there are no alternative causes:

$$\{\text{x: x is an alternative cause of } E\} = \emptyset \quad \Longrightarrow \quad \text{W}_{\text{Cause}} = \text{P(effect}|\text{cause)}$$

If participants ignore alternative causes and by mistake treat P(effect|¬cause) as 0, then they should also underestimate P(effect|cause) by evaluating it as P(effect|cause, ¬alternatives). Their estimate of P(effect|cause) will then coincide with the value of causal power, which would explain the studies corroborating [Eq1.]. In van Rooij and Schulz (2019), an formal analysis of such limiting cases was used to propose a causal power measure of the acceptability of conditionals by arguing that it is the presence of causal power that makes indicative conditionals acceptable.

Studies in the psychology of causal judgments have shown that reasoners often tend to neglect alternative causes (see, e.g., Rottman & Hastie, 2014, for an overview). These findings, in turn, fit with well-known effects from the psychology of reasoning concerning inferences like denial of the antecedent (*If A, C; ¬A, therefore ¬C*) and affirmation of the consequent (*If A, C; C, therefore A*). Indeed, a neglect of alternative antecedents (e.g., "If B, then C") has long been suspected as being part of the explanation why participants would endorse these logically fallacious inferences (Cummins, 1995; Politzer & Bonnefon, 2006). In linguistics, there is a convergent body of research studying conditional perfection (for review, see Liu, 2019), which describes the tendency to strengthen an indicative conditional into a bi-conditional that suppresses alternative antecedents. Moreover, the tendency to suppress the impact of alternative hypotheses has long been suspected of playing a role in the confirmation bias (Nickerson, 1998).

According to Fernbach, Darlow, et al. (2010, 2011), participants who are asked for conditional probabilities report them but are biased by their neglect of alternative causes. Alternatively, one may hold that participants who are asked for conditional probabilities construe the task differently and give causal power estimates instead (but see Aßfalg & Klauer, 2019). For our purposes, it is, however, interesting to note that if participants tend to ignore alternative causes, then the causal power interpretation of indicative conditionals in van Rooij and Schulz (2019) can be used to account for the Relevance Effect.

Accordingly, van Rooij and Schulz conjecture that what explains when P(C|A) is and when it is not a good predictor of P(if A, then C) in studies like Skovgaard-Olsen, Singmann, et al. (2016) is exactly whether participants take alternative causes into account. Participants are thereby portrayed as ignoring alternative causes when processing positive relevance conditionals, like "If Paul pushes down the gas pedal, then the car will speed up". In contrast, participants are predicted to take alternative causes into account when processing irrelevance items, like "If Paul is wearing a shirt, then his car will function normally", where the antecedent is obviously not an appropriate cause.

In Experiment 1, we test whether participants' tendency to ignore alternative causes makes them estimate P(C|A) as causal power in scenarios that can be interpreted causally. Experiment 1 thereby provides a critical test of the following hypotheses based on van Rooij and Schulz's (2019) work:

> (H_1) causal power [Eq. 3] accounts for the acceptance of indicative conditionals.
>
> (H_2) participants' tendency to ignore alternative causes is part of the explanation of the Relevance Effect.

We now turn to Pearl's (2009) theory of causality, which we will use to reconceptualize the relationship between indicative conditionals and causal relations. Of central importance in this context is the following observation. While Oaksford and Chater (2017) and van Rooij and Schulz (2019) argue for a causal interpretation of indicative conditionals, Pearl's idea of a hierarchy of causal queries invites a more complex picture in which indicative conditionals only play a partial role.

Pearl's Hierarchical Theory of Causality

According to Pearl (2009) and Pearl and Mackenzie (2018), there are three conceptual layers of causality: *prediction, intervention,* and *counterfactual dependency*. An understanding of these three conceptual layers is manifested by the ability to answer three different types of queries concerning the relationship between two variables, X and Y. In Pearl and Mackenzie (2018), these queries take, roughly, the following form:

Table 1. The Hierarchy of Causal Queries

Query Type	Natural Language Query	Computational Model
Predictive	"What happens to my belief in Y if I see X?"	Bayes net, SEM
Interventional	"What happens to Y if I do X?"	causal Bayes net, SEM
Counterfactual	"Would Y not have occurred if X had not occurred?"	SEM

Note. SEM = Structural Equation Modelling (see Appendix A). The distinction between 'Bayes nets' and 'causal Bayes nets' is made to emphasize that Bayes nets exist with both undirected edges representing symmetrical relations of evidential relevance, as well Bayes nets that encode directed edges used for representing assymmetrical relations of causal relevance (Højsgaard, Edwards, et al., 2012; Danks, 2014).

In Pearl (2009, p. 29), the following examples are given: 1) "would the pavement be slippery if we *find* the sprinkler off" (*prediction*), 2) "would the pavement be slippery if we *make sure* that the sprinkler is off" (*intervention*), and 3) "would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?" (*counterfactual*). As a normative competence model of causal inference, Pearl (2009) presents a theory of causal Bayes nets augmented by structural equation modelling (SEM). For Pearl (2009), it is important to emphasize that there are three irreducible layers of conceptual understanding of causal relations: 1) statistical associations for predictive inference (which can be computed by conditionalization, e.g., via Bayes nets), 2) predictions based on interventions (which are observed through manipulations in randomized, experimental studies),[70] and 3) counterfactual inferences (which can only be computed based on structural equation models of the data generating processes). In Appendix A, we illustrate the distinction between these computational models via one of Pearl's examples.

Several aspects of Pearl's theory have been investigated in psychological studies. For instance, whether reasoners differentiate between observational probabilities and interventional probabilities (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). Similarly, studies have looked at participants' understanding of the Markov assumption and the implied conditional independencies (Rehder, 2014; Rottman & Hastie, 2014; Mayrhofer & Waldmann, 2015). But whereas the causal Bayes net component of the theory has received extensive attention, the structural equation component has received less attention in psychology. Yet, some exceptions like Lagnado, Gerstenberg, et al. (2014) do exist. In Appendix A, we explain why it is important for psychology to focus more on SEM.

Research Questions Motivating this Investigation

The central question motivating the present inquiry is this: what role do conditionals as linguistic expressions play in representing causal information? Or: by accepting a conditional statement in a causal scenario, which of the three aspects of the causal relation highlighted by Pearl does a reasoner thereby accept, if any? Looking back at Table 1, answering the first two types of queries seems[71] equivalent to processing indicatives ("will the pavement be slippery, if we see/make sure that the sprinkler is off?"). Moreover, answering the third type of query is naturally taken to involve processing counterfactuals ("would the pavement have been slippery, if the sprinkler had been off?"). It is then natural to formulate the following hypothesis based on Pearl's view:

> (H_3) causal relations encode multiple layers, some of which *can* be expressed by indicatives (i.e. predictive queries), whereas the most advanced one requires the use of counterfactuals (i.e. counterfactual queries).

[70] In addition, these interventions can now also be computed by applying Pearl's (2009) do-calculus to observational studies (see also Morgan & Winship, 2018).

[71] Note that Pearl (2009, p. 29) uses "would" instead of "will" in the consequents of the observational and interventionist queries. However, the resulting conditional questions are closer in meaning to the indicatives above given the indicative antecedents than the corresponding counterfactuals. When Pearl wants to stress a counterfactual interpretation, he often uses "would have" (see, e.g., Pearl & Mackenzie, 2018, p. 320).

This, in turn, makes it natural to conjecture that:

> (H_4) indicatives that *support* and indicatives that *do not support* counterfactuals can be empirically distinguished (see also Lassiter, 2017).
> (H_5) the use of indicatives and the acceptance of causal relations can be dissociated even in causal scenarios.

To illustrate (H_5), indicatives based on spurious correlations can be used to answer predictive queries, but they do not express direct causal links between their antecedents and consequents. A well-known example is “If the barometer falls, then bad weather is coming”. According to (H_4), we would expect that a characteristic of such indicative conditionals expressing spurious correlations is that they do not support counterfactuals.

Depending on the query, the intervention might represent a natural continuation expressed in the indicative mood (e.g., “the cappuccino will taste better, if I use espresso beans”). Alternatively, the intervention might represent an unlikely continuation expressed in the subjunctive mood (e.g., “the cappuccino would taste better, if I bought an espresso machine for 10.000 €”). In our experiments, we are less concerned with interventions, however. Instead, we focus instead on different aspects of the distinction between predictive use of indicative conditionals for expressing statistical associations of *evidential relevance* and use of counterfactuals to answer queries that target *causal relevance*. For a psychological theory of probabilistic reasoning, ΔP is often used to represent evidential relevance and causal power can be used to represent causal relevance.

Overview of the Experiments

To address the above research questions, we conducted experiments that contrast a situation in which participants are provided a detailed representation of a mechanism linking inputs and outputs with observations of blackbox trials in which the mechanism was covered. The animations were inspired by the 1993 computer game, “The Incredible Machine”. Illustrations of the trials are shown in Figures 3 and 4 below:

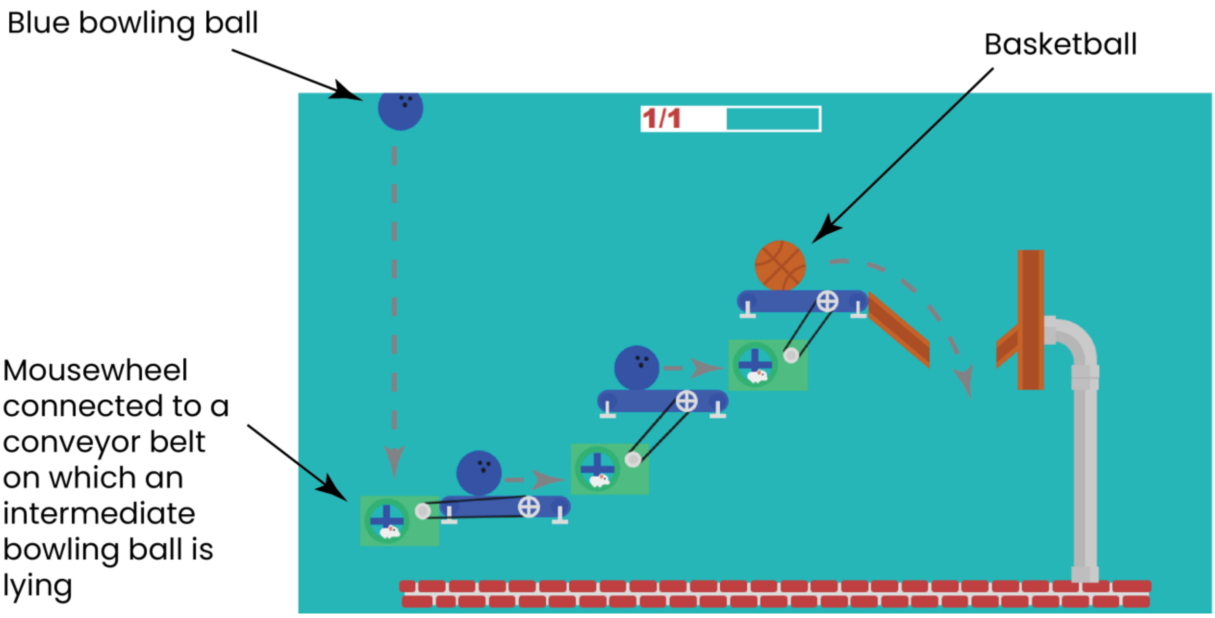

Blue bowling ball
Basketball
1/1
Mousewheel connected to a conveyor belt on which an intermediate bowling ball is lying

Figure 3. Annotated illustration of a Machine Trial in which the whole mechanism is visible. See https://osf.io/fa9rj/ for a video illustration.

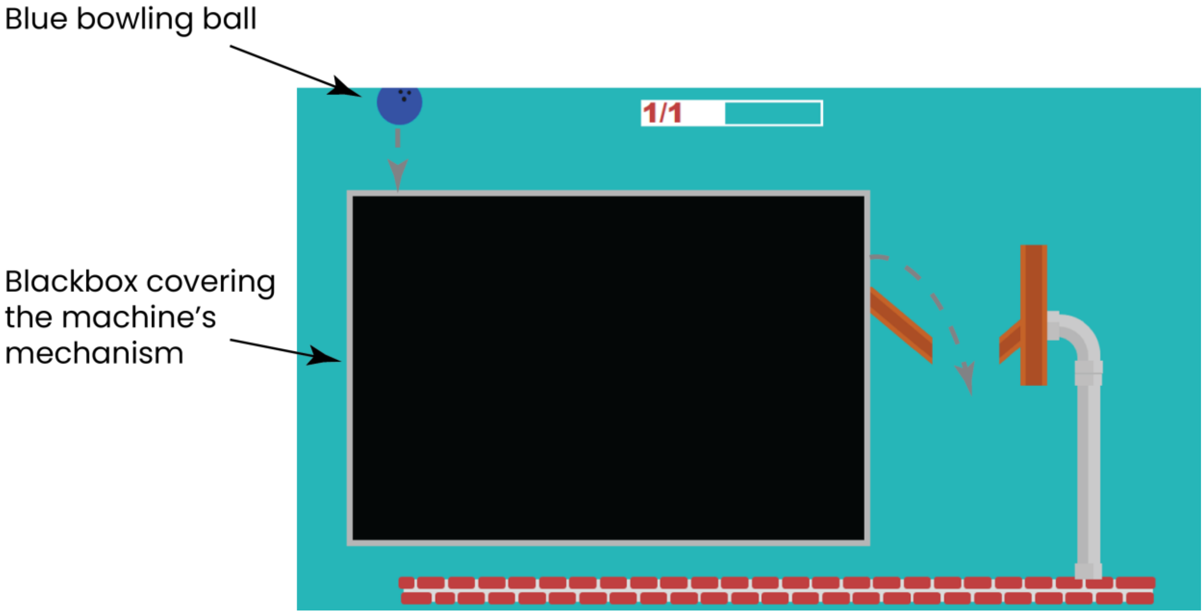


Figure 4. Annotated Illustration of a Blackbox Trial in which the mechanism is covered. See https://osf.io/fa9rj/ for a video illustration.

Figures 3 and 4 show annotated snapshots of the animations. Figure 3 depicts the Machine condition in which a causal chain unfolds when a blue bowling ball (root cause) falls onto a mouse wheel connected to a conveyor belt. This chain of events ends with the basketball dropping into the basket. In Figure 4, the mechanism is concealed. Note that this system is not deterministic because the mice can start to run on their own and they may sometimes not run even if a bowling ball hits their cage. We adopted this format as a way of manipulating the depth of participants' understanding of a causal relation in light of long-standing debates in the psychology of causal judgment about possession of structural knowledge that goes beyond associative learning (Waldmann, 1996; Waldmann & Hagmayer, 2005; Pelley, Griffiths, et al., 2017). The animations that we used conveyed the information in a trial-by-trial format. Usually, the psychology of reasoning (Manktelow, 2012) follows the research tradition on cognitive illusions (Kahneman, Slovic, et al., 1982) in studying reasoning problems via verbal scenarios. However, trial-by-trial learning paradigms are common in areas such as the psychology of learning (Bouton, 2016) and causal reasoning (Waldmann, 2017). The finding of the description-experience gap (Hertwig & Erev, 2009; Rehder & Waldmann, 2017) shows that the two paradigms can lead to different results. There is therefore a need for applying trial-by-trial learning paradigms to problems in the psychology of reasoning (Vance & Oaksford, 2020).

In our experiments, we manipulated different levels of contingency (ΔP), conditional probability (P(C|A)), and causal power (W_{Cause}). A trial-by-trial learning paradigm with the animated mouse-wheel machine was used in Experiments 2-6. Table 2 provides a brief overview of the experiments:

Table 2. Overview of the Experiments

Exp	Purpose	Method	Hypothesis
1	Critical test of assumptions needed to account for the Relevance Effect based on van Rooij and Schulz (2019).	Verbal scenarios, test of causal power as a predictor of P(if A, then C) and the influence of alternative causes on the Relevance Effect.	H_1, H_2
2	Replication of the Relevance Effect in a trial-by-trial learning paradigm.	Animations with the mouse-wheel machine task in a causal chain structure.	See below.
3	Investigate the relationship between judgments of causal power, indicatives, counterfactuals, and singular causation.	Animations with the mouse-wheel machine task in a causal chain structure with a blackbox condition.	H_3
4	Test of the acceptance of indicatives and counterfactuals as predictors of singular causation judgments.	"	H_3
5	Test of dissociation between the acceptance of indicatives and counterfactuals.	Animations with the mouse-wheel machine task in a common cause structure with a blackbox condition.	H_4, H_5
6	Replicating Experiment 4 while controlling for the influence of tense and the order of events.	"	H_4, H_5

Using the verbal stimulus materials used to originally document the Relevance Effect in Skovgaard-Olsen, Singmann, et al. (2016), Experiment 1 aimed at providing a critical test of assumptions in van Rooij and Schulz (2019). Experiment 1 thereby probed a causal power account of the acceptance of indicative conditionals (H_1) and whether participants' tendency to ignore alternative causes accounts for the Relevance Effect (H_2). The goal of Experiment 2 was to test whether the Relevance Effect could be replicated in a trial-by-trial learning task.

The next two experiments involved singular causation judgments. Singular causation judgments typically concern situations in which both the potential cause and effect are known to have co-occurred and reasoners have to establish whether the former actually caused the effect on this specific occasion. Our interest in these types of judgments originates in their role in testing (H_3) – with its claim of multiple conceptual layers in the understanding of causal relations. Moreover, we investigated singular causation judgments to ensure that participants were making the causal attributions intended by our experimental designs.

Experiment 3 investigated whether the four central constructs of 1) causal power, 2) indicative conditionals, 3) counterfactual conditionals, and 4) singular causation are influenced by the same factors in a large between-subjects experiment. The motivation for this comparison was that according to a causal interpretation of conditionals, one would expect conditionals to be affected by manipulations that influence causal judgments.

The purpose of Experiment 4 was to investigate whether singular causation judgments could be predicted by the acceptance of indicative and counterfactual conditionals. In line with the hierarchy of causal queries, Pearl (2009, Ch. 10) and Halpern (2019) build in explicit counterfactual conditions in their accounts of singular causation. Experiment 4 therefore tests whether the acceptance of counterfactual conditionals plays a role for singular causation.

Experiments 5 and 6 compared the acceptance of indicative and counterfactual conditionals in a common-cause version of the trial-by-trial learning paradigm. The goal was to investigate whether the acceptance of indicatives and counterfactuals would become dissociated for diagnostic and common-cause conditionals to test (H_4) and (H_5). The investigation of common-cause and diagnostic reasoning scenarios is crucial because they exemplify cases, where the answers to predictive queries need not represent relations of direct causal impact. For instance, measurements on a barometer are diagnostic for the coming weather conditions and can be used to answer predictive queries (e.g., “Can we expect bad weather, if the barometer falls?”). But the common cause of both are changes in atmospheric pressure.

6.2 Experiment 1: Alternative Causes

According to van Rooij and Schulz (2019), the acceptability of indicative conditionals is determined by causal power (H_1). Based on this account, it is natural to conjecture that participants assign probabilities to indicative conditionals, ‘if A, then C’, based on causal power.[72] On the auxiliary assumption that participants ignore alternative causes, causal power

[72] Note that van Rooij and Schulz (2019) are careful in stating their theory only in terms of categorical acceptance of indicative conditionals. But they indicate an extension of it to

would coincide with the conditional probability, as we have seen. van Rooij and Schulz (2019) suggest (H_2) that we can use this observation to account for the Relevance Effect in Skovgaard-Olsen et al. (2016). To do so, one would have to conjecture that participants' tendency to ignore alternative causes makes P(C|A) a good predictor of P(if A, then C) for Positive Relevance ($\Delta p > 0$) items. In contrast, the lack of causal dependence of consequent on the antecedent would make P(C|A) overestimate P(if A, then C) for Irrelevance items ($\Delta p = 0$). In addition, we probe whether we can replicate the Relevance Effect in a situation, where it is difficult to ignore alternative causes by using a task that builds on Byrne (1989). The purpose of this was to provide a critical test of (H_2) as an auxiliary assumption of van Rooij and Schulz (2019), however.

In a much-discussed study, Byrne (1989) presented participants with conditional inference problems like, e.g., “If Lisa has an essay to write, then Lisa will study late in the library”, along with an additional premise presenting an alternative antecedent, e.g., “If Lisa has some textbooks to read, then Lisa will study late in the library”. Applying this idea to our context, we asked participants for probability evaluations in the presence of alternative causes. We did this by first obtaining alternative causes generated by other participants from a pilot study. We then displayed these above the test questions in the present study for participants in the Alternative-Causes condition. The goal was to see whether the Relevance Effect could be replicated even under full knowledge of alternative causes, when the potential cognitive effort of generating alternative causes had been removed.

Experiment 1 thus provides a critical test of the assumptions needed to account for the Relevance Effect based on van Rooij and Schulz' (2019) causal power account of indicative conditionals.

6.2.1 Method

Procedures shared by all Experiments

Experiment 1, like all the other experiments reported in this paper, was conducted as an online study testing a large and demographically diverse sample. Participants were sampled over the Internet (via Mechanical Turk) from the USA, UK, Canada, and Australia. Subjects received a monetary compensation for their participation. The following exclusion criteria were used: 1) not having English as native language, completing the task in less than

account for degrees of acceptability as here and explicitly apply their theory to data from psychological experiments that asked for degrees of acceptability in the form of subjective probabilities. For this reason, we empirically test such an extension of their theory.

min seconds or in more than *max* seconds,[73] 2) failing to answer two simple SAT comprehension questions correctly in a warm-up phase, 3) answering 'not serious at all' to the question 'how serious do you take your participation' at the beginning of the study, and 4) answering "yes" to whether they recognized the animation from the computer game "Incredible Machines".[74] For each experiment, it was found that these exclusion criteria had a minimal effect on the demographic variables.

To reduce the dropout rate during the experiment, participants first went through three pages in all the experiments. These three pages stated our academic affiliations, posed the two SAT comprehension questions in a warm-up phase, and presented a seriousness check asking how careful the participants would be in their responses (Reips, 2002). Participants were also shown two dummy probability questions to familiarize them with the use of a slider.

Participants

A total of 1004 people completed Experiment 1. After applying the *a priori* exclusion criteria the final sample consisted of 681 participants. Mean age was 39.82 years, ranging from 18 to 79.[75] 46.1 % of the participants were male. 72.39 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a between-subjects design with three factors. The first was Relevance (with two levels: Positive Relevance (PO) vs. Irrelevance (IR)). The second was Priors (with four levels: HH vs. HL vs. LH vs. LL; for example, HL means that P(A) = high and P(C) = low). The third was Group (with two levels: Alternative-Causes vs. Control). Thus, there were 16 between-subjects conditions in total.

We will abbreviate the 2 Relevance × 4 Prior conditions as follows below: POHH, POHL, POLH, POLL, IRHH, IRHL, IRLH, IRLL. The Relevance and Prior factors were combined factorially to ensure that the examined relationship generalize across a wide range of different probabilities. This ensures that our results do not merely pertain, e.g., to conditionals with high antecedent and consequent probabilities, which tend to sound more plausible, but generalize across a wider spectrum.

[73] Due to differences among the tasks, the min and max varied between experiments: Experiment 1 = [60s, 1800s], Experiments 2, 3 = [240s, 3600s], Experiment 4 = [120s, 1800s], Experiments 5, 6 = [240s, 1800s].

[74] This last exclusion criterion was used only in Experiments 3-6, which introduced a blackbox condition that required controlling the background knowledge of the participants.

[75] One participant indicated the age of '14' but given Amazon's regulations we doubt this value.

Materials and Procedures

Each of the 16 between-subjects conditions was randomly assigned to one of 12 scenarios. Random assignment was performed with replacement, such that each participant saw a different scenario for each condition. This ensured that the mapping of condition to scenario was counterbalanced across participants. One of the 16 between-subjects conditions was randomly assigned to a participant within a block. The block consisted of one page displaying a scenario and three pages presenting the dependent variables (see below). As a reminder, the scenario was presented in grey on the top of these three pages. These scenario texts have been found in previous experiments (Skovgaard-Olsen, Singmann, et al., 2016, 2017) to reliably induce assumptions about relevance and prior probabilities of the antecedent and the consequent that implement our experimental conditions. Table 3 displays sample items of the Paul scenario for Positive Relevance ($\Delta p > 0$), and Irrelevance ($\Delta p = 0$).

Table 3. Stimulus Materials of the Paul Scenario

Scenario	Paul is driving on a straight road with hardly any traffic ahead. He is on his way to work in an investment bank and is running late. At this point the drive will take about one hour and he is supposed to arrive in 40 minutes.	
	Positive Relevance	**Irrelevance**
HH	If Paul pushes down the gas pedal, then the car will speed up.	If Paul is wearing a shirt, then his car will function normally.
HL	If Paul drives fast, then he will be there in time for work.	If Paul is wearing a shirt, then his car will suddenly break down.
LH	If Paul's car suddenly breaks down, then he will be late for work.	If Paul is wearing shorts, then his car will function normally.
LL	If Paul pushes down the brake pedal, then the car will slow down.	If Paul is wearing shorts, then his car will suddenly break down.
	Positive relevance (PO): mean ΔP = .32 Irrelevance (IR) mean ΔP = -.01	High antecedent: mean P(A) = .70 Low antecedent: mean P(A) = .15 High consequent: mean P(C) = .77 Low consequent: mean P(C) = .27

Note. HL: P(A) = High, P(C) = low; LH: P(A) = low, P(C) = high. The bottom rows display the mean values for all 12 scenarios pretested in (Skovgaard-Olsen, Singmann, et al., 2017). $\Delta p = P(C \mid A) - P(C \mid \neg A)$

For the Paul scenario text in Table 3, participants assume that the event "Paul pushes down the gas pedal" raises the probability of the event "the car will speed up". They moreover assume that both sentences have a high prior probability (Positive Relevance, HH). Conversely, participants assume that the event "Paul is wearing a shirt" is irrelevant for whether "his car will function normally", and that both have a high prior (Irrelevance, HH). Previous studies have moreover confirmed that participants view "Paul pushes down the gas pedal" as a *reason for* the event "the car will speed up" and "Paul is wearing a shirt" as neither a reason for nor against "his car will function normally". The full list of scenarios can be found at: https://osf.io/j4swp/.

On the three randomly ordered pages following the initial scenario, participants were asked to provide estimates of conditional probabilities (P(C|A), P(C|¬A)) via the Ramsey Test. They were thus asked to suppose that the antecedent is the case and evaluate the probability of the consequent under this assumption on a scale from 0-100%. In addition, participants were asked to assign probabilities on the same scale to conditional statements across relevance conditions, e.g.: "IF Paul pushes down the gas pedal, THEN the car will speed up".

In a pilot study, we had participants generate alternative causes for the Positive Relevance and Irrelevance items. Two independent raters coded how many independent and plausible causes the participants listed (see https://osf.io/fa9rj/ for the coding instructions). It was found that the rank order $|\text{Alternatives}_{\text{PO}}| > |\text{Alternatives}_{\text{IR}}|$ obtained not only for the averaged ratings across conditions ($\overline{Alternatıves}_{PO}$= 3.13, $\overline{Alternatıves}_{IR}$= 2.06), *t*(3.56) = 3.39, *p* = 0.033, but also for each condition and each rater within each condition. For participants in the Alternative-Causes Group, an alternative cause generated by the participants in the pilot study was selected for each of the 96 Relevance × Prior × Scenario combinations. This alternative cause was presented to participants as the antecedent of a conditional. For instance, some participants in the Alternative-Causes Group were shown the following conditional presenting an alternative antecedent for the item above:

"IF Paul is driving down a hill, THEN Paul's car will speed up."

This conditional was displayed on a separate page after the scenario and repeated on every page above the test question for participants in the Alternative-Causes Group. In contrast, participants in the Control Group were presented with the three dependent variables without alternative antecedents.

6.2.2 Results and Discussion

Causal power was calculated based on participants' responses to the conditional probability questions through calculations of ΔP and the following formulas:

[Eq5.] $$power = \begin{cases} \frac{\Delta P}{1-P(C|\neg A)} & if\ \Delta P \geq 0 \\ \frac{-\Delta P}{P(C|\neg A)} & if\ \Delta P < 0 \end{cases}$$

The formulas calculate causal power for generative and preventive causes, respectively.[76] The first goal of the analysis was to establish whether the contrast between the Alternative-Causes and the Control Group influenced the Relevance Effect.

A mixed ANOVA was first conducted using the R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020). The Condition factor (POHH vs. POHL vs. POLH vs. POLL vs. IRHH vs. IRHL vs. IRLH vs. IRLL) and Alternatives factor (Alternative-Causes vs. Control Group) were specified as varying between-subjects. The DV factor (P(C|A) vs. P(C|¬A) vs. P(if A, then C) vs. ΔP vs. power) was specified as a within-subject factor. Through this model, we tested the impact of the Alternative-Causes vs. Control Group contrast on both the three measured (P(C|A), P(C|¬A), P(if A, then C)) and the two calculated dependent variables (ΔP, power) across the between-subjects conditions.

Table 4. ANOVA Table for Experiment 1

Effect	*df*	MSE	*F*	η^2_G	*p*
Condition	7, 665	0.19	73.58	.23	< .0001
Alternatives	1, 665	0.19	2.62	.002	ns
Condition:Alternatives	7, 665	0.19	1.55	.006	ns
DV	2.49, 1655.62	0.12	192.73	.15	< .0001
Condition:DV	17.43, 1655.62	0.12	22.24	.13	< .0001
Alternatives:DV	2.49, 1655.62	0.12	0.79	.0007	ns
Condition:DV:Alternatives	17.43, 1655.62	0.12	0.89	.006	ns

Note. η^2_G is generalized eta squared, which is an effect size measure that is recommended for repeated measures ANOVA in Bakeman (2005). The Alternatives factor encodes the contrast between the Alternative-Causes Group and the Control Group (alternative causes absent).

Given that the contrast between the Alternatives-Causes and the Control Group was neither involved in a simple effect nor in any statistically significant interactions (Table 4), Figure 5 displays the results without this factor:

[76] When ΔP = 0 causal power was stipulated to be zero to avoid the problem of undefined values for cases when P(C|¬A) = 1. Removing the 41 participants with undefined values does not change the relative fit of the models, however. For the purpose of predicting P(if A, then C) by causal power (see M1 below), it would also have been possible to only apply the causal power formula to the subset of cases where ΔP ≥ 0. Figure 5 reveals, however, that the fit of M1 would not have improved by predicting P(if A, then C) = 0 in such cases due to zero generative, causal power.

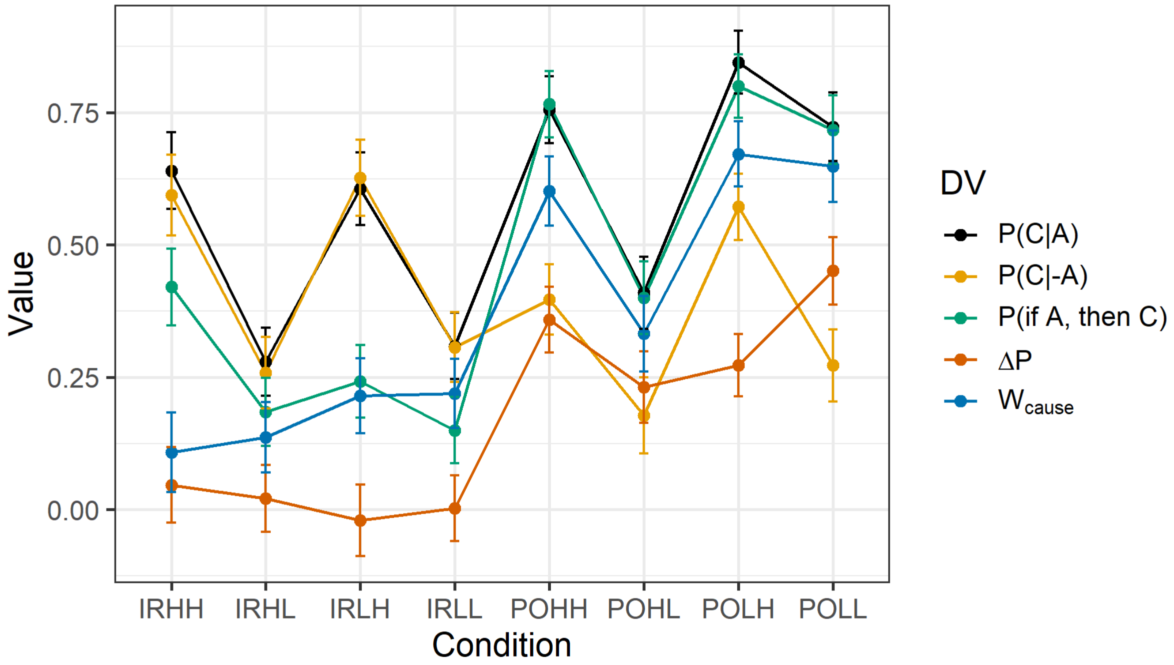


Figure 5. The measured and calculated mean estimates of the five DVs are displayed across the eight Relevance × Priors conditions. The error-bars represent 95% CI intervals.

The systematic differences between P(C|A) and P(if A, then C) for the IR items are noteworthy in Figure 5, because they violate [Eq.1]. At the same time, the two constructs are nearly identical for the PO items. Both findings are in line with the predictions of the Relevance Effect shown in Figure 1. The lack of coincidence of W_{cause} and P(C|A), and the finding that P(C|¬A) estimates are consistently above 0, is also noteworthy, because it casts doubt on van Rooij and Schulz's (2019) auxiliary assumption.

The goal of the second analysis was to test whether causal power predicted P(if A, then C) better than other models. Three mixed linear models were contrasted for modelling P(if A, then C), with random intercepts for scenarios using the R-package `lme4` (Bates et al., 2015):

(**M1**) a model that predicts P(if A, then C) based on causal power (van Rooij & Schulz, 2019), which measures P(effect|cause,¬alternatives).

(**M2**) a model that predicts P(if A, then C) by P(C|A) as measured by the Ramsey Test, which corresponds to the suppositional theory of conditionals (Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009).

(**M3**) a model that predicts P(if A, then C) based on an interaction between P(C|A) and the Relevance Condition factor (Positive Relevance vs. Irrelevance), which corresponds to the model used by Skovgaard-Olsen, Singmann, et al. (2016).

The outcome of the model comparison is displayed in Table 5:

Table 5. Model Comparison for Indicative Conditionals

Model		χ^2	df	*p*	AIC	BIC
M1	Causal Power	241.52	1	< .0001	481.70	499.80
M2	P(C|A)	652.26	1	< .0001	232.43	250.52
M3	P(C|A)	515.81	1	< .0001	38.87	66.01
	Relevance Condition	200.67	1	< .0001		
	P(C|A): Relevance Condition	28.72	1	< .0001		

Note. The lower AIC and BIC values indicate that M3 is superior to M1-M2 in light of the parsimoni vs. fit trade-off. 'Relevance' is a categorical factor encoding 'Positive Relevance' vs. 'Irrelevance'.

The information criteria clearly converge on M3. This model permits an interaction between P(C|A) and the Relevance Condition factor such that a lower slope of P(C|A) is expected in the Irrelevance Condition.

In this experiment, the Relevance Effect reported in Skovgaard-Olsen et al. (2016) was replicated both in the Alternative-Causes and the Control Group. It was thereby found that there was no significant effect of explicitly presenting alternative causes to the participants in the manner of Byrne (1989) for the Relevance Effect. This finding, in turn, challenges the auxiliary assumption (H_2) in van Rooij and Schulz (2019) that participants' tendency to ignore alternative causes accounts for the Relevance Effect.

6.2.3 Summary

Based on the pilot study, we know that participants *can* generate alternative causes for both the positive relevance and irrelevance items. Hence, the stimuli in Skovgaard-Olsen, Singmann, et al. (2016) implicitly manipulate the presence of alternative causes. When comparing participants' probability assignments when the presence of alternative causes is implicitly manipulated (the Control Group) and when it is explicitly manipulated (the Alternative-Causes Group), we find no significant differences (see Table 4).[77]

In a direct model comparison, it was found that when comparing the Suppositional Theory of Conditionals (Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009), the causal power theory of the acceptability of indicative conditionals (van Rooij & Schulz, 2019), and the model used in Skovgaard-Olsen, Singmann, et al. (2016), the latter turned out to be the best fitting model. What allowed this model to outperform the other models was that it includes a simple effect of Relevance and an interaction between P(C|A)

[77] As such, the relationship between the Alternative-Causes Group and the Control Group can be viewed as resembling the relationship between the so-called *explicit* paradigm in Byrne (1989) and the *implicit* paradigm in Cummins, et al. (1991). These two paradigms also led to similar results.

and the Relevance Condition factor. This interaction term expects a lower slope of P(C|A) in the Irrelevance Condition, where indicative conditionals are predicted to appear defective. At the same time, it allows the use of P(C|A) as a predictor of P(if A, then C), which is especially well-supported in the Positive Relevance Condition. In Appendix B, we further investigate the issue of why causal power theories could not account for our findings through a simulation analysis.

6.3 Experiment 2: The Relevance Effect

Beginning with Experiment 2, we used the animated mouse-wheel-machine paradigm.
In Experiment 1, the Relevance Effect was replicated with verbal scenarios. The purpose of Experiment 2 was to replicate this effect using a trial-by-trial learning paradigm involving mechanistic knowledge for the first time.

6.3.1 Method

Participants

A total of 350 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 1. The final sample after applying the *a priori* exclusion criteria consisted of 221 participants. Mean age was 40.27 years, ranging from 20 to 74. 38.91 % of the participants were male. 69.23 % indicated that their highest level of education was an undergraduate degree.

Design

The experiment had a within-subject design with Relevance as a within-subject factor (with three levels: Positive Relevance (PO) vs. Negative Relevance (NE) vs. Irrelevance (IR)), which refers to three types of items explained below. In total, 20 trials were shown which implemented the following conditions:

Table 6. Experimental Design

	P(C|A)	P(C|¬A)	ΔP
PO	0.83	0.75	0.08
IR	0.80	0.80	0.00
NE	0.75	0.83	-0.08

Note. Contingencies calculated based on the initial trial, where the mc questions were presented, and the subsequent 19 randomly ordered machine trials.

A pilot study[78] had found that although ΔP in the trials shown differed modestly, participants were able to arrive at stronger ΔP differences across conditions when processing the items introduced below. Their background knowledge and the evidence presented concerning the mechanism permitted them to arrive at stronger subjective ΔP values than what was displayed in the trials. These subjective ΔP values correlated with participants' ordinal reason relation assessments, $r_{\text{polyserial}}(97) = .73$, $p < .0001$. The pilot study thus showed that we could use a single contingency condition to reliably manipulate the differences Positive Relevance, Negative Relevance, and Irrelevance using the items introduced below.

Materials and Procedure

To ensure that the animations were displayed properly, participants were instructed to adjust their browser so that they would see the whole box in which the animation was presented. We first presented one trial with three multiple-choice questions. After the display of a fixation cross in the upper left corner, participants saw an animation with the mechanistic set-up depicted in Figure 3. In the animation, a blue bowling ball fell down on a mouse-wheel, connected to a conveyor belt, which set a chain of events in action that eventually resulted in a red basketball falling down the basket on the right side of the screen. Participants were instructed that the animations would always start with the display of a white fixation cross in the upper left corner (the position in which the blue bowling ball occurred). Secondly, participants learned that there was a process bar in the middle of the screen that visualizes when the animations would stop. Thirdly, they were asked to pay attention to the animation in all trials, and that they could not press "continue" until all animations had been shown.

In the first trial, the animation paused several times to pose multiple-choice questions to ensure that participants had understood what they had seen. After this trial, participants were given the following instruction:

> As you will see, sometimes the mice can be sleepy ("ZzzZZZz") and fail to run despite being prompted. The mice can also be excited ("Wo hoo!") and start to run without being prompted.

This information was given to make participants aware that (1) the effect could occur in the absence of the target cause and (2) sometimes the effect could remain absent even in the presence of the target cause. The next page informed participants about the change of an irrelevant feature of the machine to implement the Irrelevance condition:

[78] https://osf.io/fa9rj/

Sometimes, the bricks also look a bit brighter due to small random shifts in the lights.

Participants then saw 19 further trials implementing the conditions outlined in Table 6. An illustration of the trials can be found on: https://osf.io/fa9rj/.

Following these further animations, three blocks of items were displayed in random order containing several randomly ordered questions. These three blocks implemented the within-subject Relevance factor by presenting participants with the following Positive Relevance (PO), Negative Relevance (NE), and Irrelevance (IR) items, which all concerned properties of the machine shown:

PO: IF the blue bowling ball falls down, THEN the red basketball drops down in the basket.

IR: IF the lights make the bricks in the machine look brighter, THEN the red basketball drops down in the basket.

NE: IF none of the blue bowling balls are moving, THEN the red basketball drops down in the basket.

Within each block, participants were asked to evaluate the probability of these conditionals and the conditional probability of the consequent given the antecedent via the Ramsey Test on a scale from 0% to 100%.

Finally, participants were asked whether they recognized the animation as originating from the computer game "The Incredible Machine", and a list of demographic questions.

6.3.2 Results and Discussion

As a manipulation check, it was found that the following percentages of the participants answered the initial MC question correctly: 81.45%, 96.38%, 94.57%. Regressing P(if A, then C) on P(C|A), the following differences across Relevance conditions were found:

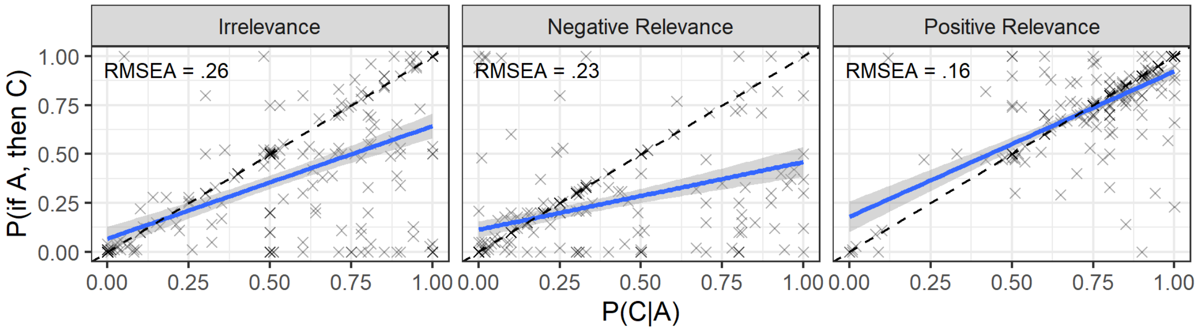


Figure 6. The figure displays predictions of their ratings of P(if A, then C) by their P(C|A) responses. Both variables were rescaled by dividing by 100. The dashed lines indicate the predictions by [Eq. 1]. The root mean square error (RMSEA) values displayed were calculated based on fitting separate least square linear regressions to the PO, NE, and IR conditions.

To test for the influence of Relevance on P(C|A) as a predictor of P(if A, then C), three mixed linear models were contrasted, with random intercepts for participants using the R-package `lme4` (Bates et al., 2015), as shown in Table 7:

Table 7. Model Comparison for Indicative Conditionals

Model		χ^2	df	*p*	AIC	BIC
M1	P(C|A)	637.15	1	< .0001	57.03	75.01
M2	P(C|A)	291.66	1	< .0001	-72.58	-45.60
	Relevance Condition	166.98	2	< .0001		
M3	P(C|A)	302.04	1	< .0001	-86.08	-50.11
	Relevance Condition	172.46	2	< .0001		
	P(C|A): Relevance Condition	24.81	2	< .0001		

Note. Note that 'P(C|A)' here refers to the values measured by the Ramsey Test. The lower AIC and BIC values indicate that M3 is superior to M1 and M2 in light of the parsimoni vs. fit trade-off.

The information criteria favor M3. The results thus indicate that there was both a simple effect of Relevance on P(If A, then C) and an interaction between P(C|A) and Relevance.

For the PO item, the estimated marginal means of the P(if A, then C) ratings were 0.55, 95% CI [0.51, 0.60], 0.74, 95% CI [0.71, 0.77], and 0.92, 95% CI [0.88, 0.97], when P(C|A) was held fixed as 0.50, 0.75, and 1.00, respectively. In contrast, when P(C|A) was held fixed at the same values for the IR item, the estimated marginal means of the P(if A, then C) rating were 0.36, 95% CI [0.33, 0.39], 0.50, 95% CI [0.46, 0.54], and 0.64, 95% CI [0.59, 0.70], respectively. For the NE item, the corresponding values were 0.29, 95% CI [0.25, 0.32], 0.37, 95% CI [0.32, 0.42], and 0.46, 95% CI [0.39, 0.53].

There is a striking match between the data pattern in Figure 6 and the pattern outlined in Figure 1. The results indicate that while participants' responses are well described by [Eq. 1] for the Positive Relevance item, substantial divergences are found for the NE and IR items. Previously, this effect has only been reported using verbal scenarios (Skovgaard-Olsen, Singmann, et al., 2016; Skovgaard-Olsen, Kellen, et al., 2019; Vidal & Baratgin, 2017), which was replicated in Experiment 1. Now we show that this Relevance Effect can also be found in a trial-by-trial learning paradigm in the presence of mechanistic knowledge for the first time.

6.4 Experiment 3: Black Box vs. Mechanism

To investigate the impact of mechanistic knowledge, Experiment 3 introduced a contrast between one group of participants seeing the underlying mechanism (as in Experiment 2) and another group of participants seeing the same setting covered by a blackbox. The black box concealed the underlying mechanism of the events participants saw (see Figure 4). Our experiments thus allowed us to investigate the effects of knowledge about the operation of a machine, compared with when one can only form associations based on observed covariances in blackbox trials. Experiment 3 used this blackbox manipulation to investigate the impact of participants' causal knowledge on estimates of conditional probabilities and conditional reasoning. Because participants in the blackbox condition only had observed covariances to rely on, we will refer to this group as 'the Regularity Group'.

While other studies have investigated indicative conditionals and singular causation judgments in the same experiment (e.g., Sikorski et al., 2019), we decided to additionally have participants provide counterfactual conditionals and causal power judgments. To investigate the relationship between mechanistic knowledge, causality, conditionals, and contingency, a large online study was therefore conducted with 32 between-subjects conditions that factorially varied these factors.

According to (H_3), causal relations encode multiple conceptual layers, some of which require answers to queries that go beyond what is expressed by indicative conditionals. On the opposing view, indicative conditionals themselves express causal relations. To corroborate (H_3), it would have to be shown that there are aspects of causal relations that go beyond the acceptance of indicative conditionals. Experiments 4-6 were devoted to this aim. In contrast, evidence against (H_3) would have to show that participants evaluate indicative conditionals equivalently to explicit causal notions like singular causation and causal power. Experiment 3 tested this hypothesis.

Experiment 3 therefore investigated whether experimental manipulations known to influence causal reasoning (i.e., contingency conditions and the Machine vs. Blackbox contrast) had a similar impact on four outcome variables of theoretical interest (the probability of indicative conditionals, counterfactual conditionals, singular causation, and causal power). Secondly, Experiment 3 investigated whether participants evaluated these four variables equivalently, or whether differences between them emerged in support of (H_3). To test this, SEM models were fitted to the data across all 32 conditions. A comparison of these models revealed whether it was possible to constraint the four main DVs to be identical. Of interest for these comparisons was whether indicative conditionals were evaluated as explicit causal constructs such as causal power and singular causation in a between-subjects comparison. Thirdly, Experiment 3 was designed to investigate whether the influence of our experimental manipulations on the four main DVs was mediated by participants' estimations of Ramsey Test conditional probabilities. Fourthly, it was investigated whether this mediational relationship in turn was moderated by reason relation assessments.

6.4.1 Method

Participants

A total of 2211 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 1 with one addition. Experiment 3 additionally excluded participants who recognized the set-up from the computer game "The Incredible Machine", because such participants will know the mechanism of the machine even in the blackbox condition. The final sample after applying the *a priori* exclusion criteria consisted of 1472 participants. Mean age was 38.94 years, ranging from 18 to 81.[79] 40.42 % of the participants were male. 70.72 % indicated that their highest level of education was an undergraduate degree.

Design

The experiment had a between-subjects design with three factors: DV_{type} (with four levels: indicative conditional vs. singular causation vs. counterfactual conditional vs. causal power), Contingency (with four levels outlined in Table 8 below: a vs. b vs. c vs. d), and Group (with two levels: Machine vs. Regularity, which differed on whether participants saw the underlying mechanism as in Figure 3 or only the blackbox trials as in Figure 4).

Table 8. Experimental design, contingency conditions

	$P(C\|A)$	$P(C\|\neg A)$	ΔP	$W_{Antecedent}$
a	0.75	0.50	0.25	0.50
b	0.25	0.00	0.25	0.25
c	0.25	0.25	0.00	0.00
d	0.75	0.75	0.00	0.00

Note. The contingency conditions were introduced through the first initial trial and consecutive 15 randomly ordered blackbox trials. These were subject to the constraint that the last trial displayed was a <bowling ball, basketball> trial. This was done to enable, e.g., participants to make singular causation judgments about whether the bowling ball caused the basketball to fall down the basket. '$W_{Antecedent}$' = the causal power of the antecedent of the conditionals.

Materials and Procedure

Participants were randomly assigned to one of these 32 between-subjects conditions. To investigate the impact of mechanistic knowledge, we first presented one group of participants (those in the Machine condition) with a trial showing the mechanistic set-up from Experiment 2. Participants in the Regularity Group, by contrast, only saw a blackbox trial.

[79] One participant answered '5'. This answer was excluded from the reported age range.

In the first trial, the animation was paused several times to pose multiple-choice questions to ensure that participants had understood what they had seen.

For the 15 trials that followed, all participants saw 15 blackbox trials (Figure 4) conveying the different contingencies listed in Table 8. Participants in the Machine Group were instructed that the blackbox covered most of the animation with the machine that they had seen on the first trial.[80] Following these trials, participants were shown a block with three dependent variables in random order. Two of the dependent variables were shown to all participants. One of these was the following Ramsey Test question:

> Suppose that the blue bowling ball falls down. [highlighted in blue]
> Under this assumption, how probable is the following statement on a scale from 0 to 100%:
> The red basketball drops down in the basket. [highlighted in blue]

The second question was an ordinal reason relation assessment on a five-point Likert-scale, where the quoted sentences were highlighted in blue:

> Please indicate the extent to which "the blue bowling ball falls down" is a reason for/against "the red basketball falls into the basket":
> A strong reason against; a reason against; neutral; a reason for; a strong reason for.

The third dependent variable was a probability judgment on a scale from 0 to 100% with an item determined randomly based on the chosen between-subjects condition from the following list:

> **Singular causation:**
> The blue bowling ball caused the red basketball to drop down in the basket.
>
> **Indicative Conditional:**
> IF the blue bowling ball falls down, THEN the red basketball drops down in the basket.
>
> **Counterfactual Conditional:**
> IF the blue bowling ball had NOT fallen down, THEN the red basketball would NOT have dropped down in the basket.

[80] Since the mechanism was covered for these trials, participants were never exposed to animations of sleepy or excited mice as disablers and alternative antecedents like in Experiment 2. Moreover, since the IR item from Experiment 2 was not used, the colour of the bricks remained constant throughout.

> **Causal Power:**
>
> Some instances of the red basketball dropping down in the basket are due to hidden alternative causes. Imagine there are 100 runs of the animation in which no alternative causes are present. Suppose that the blue bowling ball falls down in all of these 100 runs. In how many of them would the red basketball drop down in the basket?

The formulation of the causal power question followed a standard formulation found in the literature on causal judgment (see e.g., Cheng & Lu, 2017; Liljeholm & Cheng, 2009).

Finally, participants were asked whether they recognized the animation as originating from the computer game "The Incredible Machine", and a list of demographic questions.

6.4.2 Results and Discussion

Pilot Study for Experiment 3

We first conducted a pilot study. We here summarize some of its results, because they concern the issue of whether participants ignore alternative causes in our experimental paradigm, which was the auxiliary hypothesis used to explain the Relevance Effect in van Rooij and Schulz (2019). Further results concerning the impact of mechanistic knowledge on changes to contingencies are reported on: https://osf.io/fa9rj.

The pilot study presented participants with two open-ended questions, where participants were requested to list up to seven other alternative causes of the basketball dropping into the basket than the blue bowling ball falling down. An acceptable answer to this question might be that one of the mice started to run on their own volition. Secondly, participants were asked to explain the mechanism in the black box which makes the basketball fall into the basket. To analyze participants' open-ended responses, we had two raters classify the number of alternative causes to the blue bowling ball falling down. As a proxy for the complexity of the explanations, the two raters also classified the number of functional units in participants' explanations of why the red basketball dropped into the basket. Details on the classification can be found on: https://osf.io/fa9rj.

The Machine Group (*M* = 4.35, *SD* = 2.25) produced significantly more functional units in their explanations than the Regularity Group (*M* = 1.84, *SD* = 1.13), *t*(127.18) = 9.17, *p* < .0001. Moreover, it was found that the Machine Group (*M* = 1.21, *SD* = 1.4) produced significantly more alternative causes than the Regularity Group (*M* = 0.82, *SD* = 0.99), *t*(153.42) = 2.054, *p* = .042. In the Machine condition, 39.54% produced zero (plausible) alternative causes. In the Regularity condition, 47.5% of the participants produced zero (plausible) alternative causes. However, these proportions did not differ significantly, $\chi^2(1)$ =

0.77, *p* = 0.38. In sum, it was found that the explanations of the Machine Group were more complex, as measured by the number of functional units used in their explanations. Moreover, the Machine Group tended to list more alternative causes than the Regularity Group. However, the two groups did not differ in the frequency with which zero physically plausible, alternative causes were listed, which was found to be high (>39%) in both groups.

Main Study

Participants in the Machine Group were asked three MC questions. In the Regularity Group, two MC questions were presented. As a manipulation check, it was found that the following percentages of participants answered the initial mc question correctly: (Machine Group) 83.70%, 98.10%, 96.20%, (Regularity Group) 88.18%, 88.45%.

Structural Equation Model

To analyze all 32 between-subjects conditions, a structural equation model with four groups (one for each of the main dependent variables, y_j) and moderated mediation was fitted to the data of all 1472 participants (see Figures 7, 8). Structural equation modelling (SEM) is a generalization of regression models used for causal inference in statistics, which is based on modelling the covariance matrix. SEM moreover permits the estimation of direct and indirect effects of explanatory variables as well as imposing conditional independence constraints from a causal model (Kline, 2016; Shipley, 2016). For our purposes, SEM is suited for identifying the sensitivity of our four main outcome variables to the experimental manipulations while holding other factors fixed. Moreover, we use the SEM model for testing the indirect effects of the experimental manipulations through mediating variables.

Due to the theoretical importance of Ramsey Test conditional probabilities, they were considered as a mediator of our manipulations. In line with previous reseach, the indirect paths through the Ramsey test ($P(C|A)_{DV}$) were furthermore moderated by a qualitative reason relation assessment, $Reason_{DV}$. Across the four groups, the two mediators, P(C|A) and Reason, were modeled in the same way. But the model allowed for differential influence of these on the main outcome variable across the four different types.

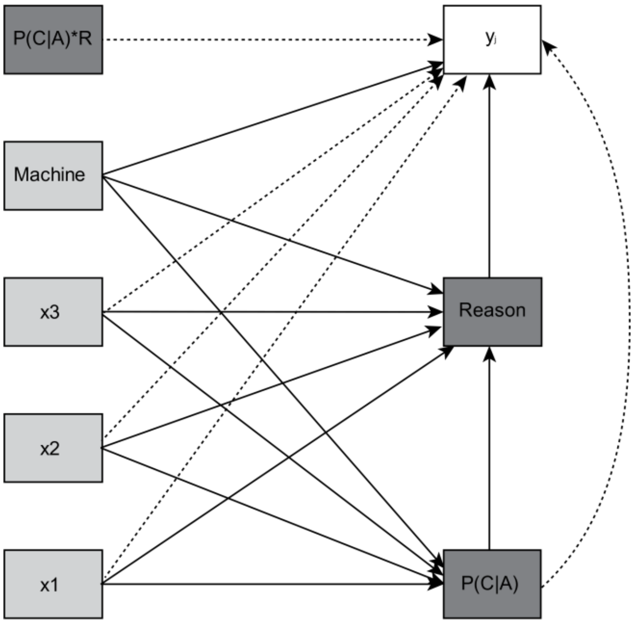
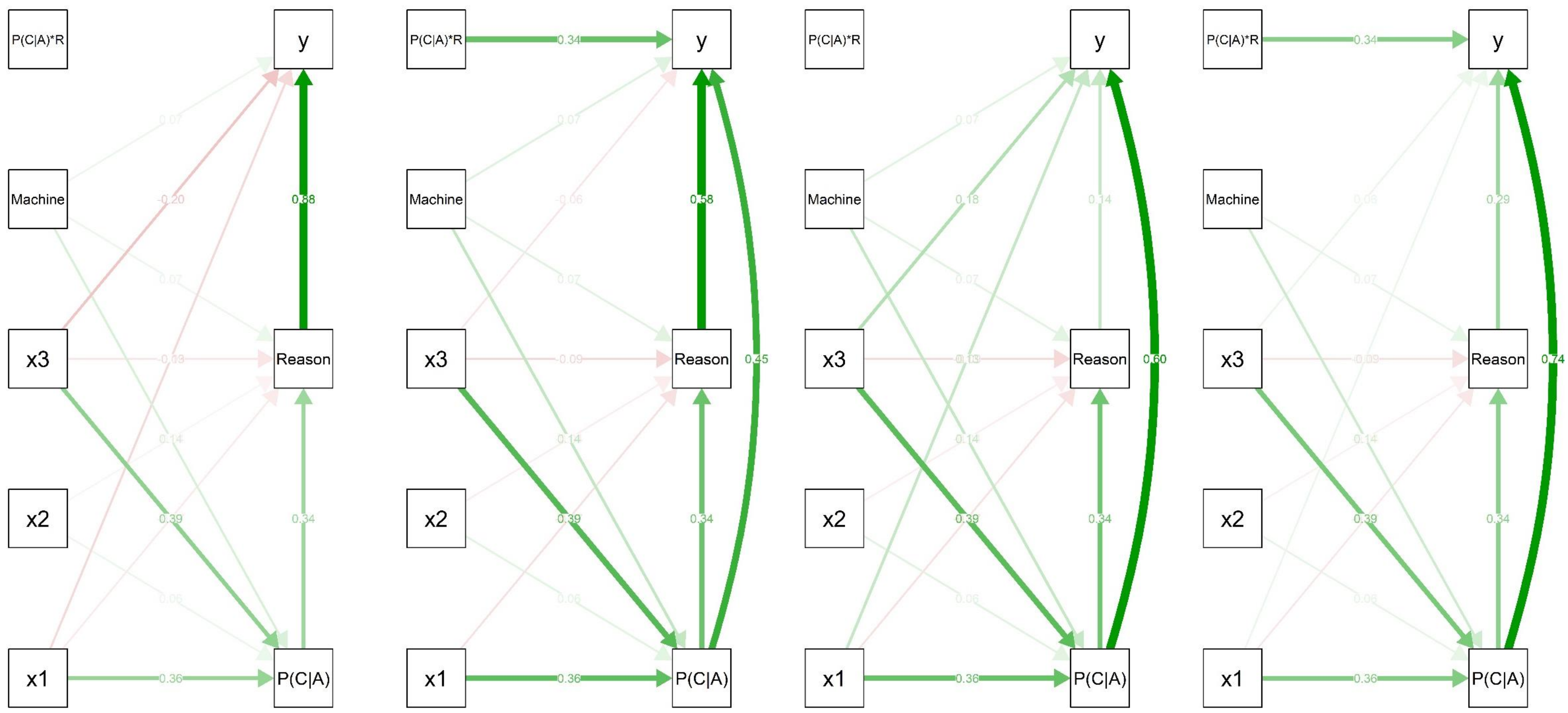


Figure 7. *Conceptual Diagram*. The dashed edges could vary between the four main dependent variables; the solid lines were fixed for all. ‘Contingency’ (a, b, c, d) was coded into three contrasts: x1, x2, and x3 (see below). A mean structure and covariances (not displayed here) were also added to the SEM model: see https://osf.io/fa9rj for further details. ‘P(C|A)*R’ = interaction between P(C|A) and Reason.

The model permits the experimental conditions to influence the four main outcomes variables via two causal chains: 1) through the direct effects of the objective input (i.e. the experimental conditions) on the subjectively evaluated DVs, and 2) through indirect effects, where the objective input affects subjective evaluations of P(C|A) and reason relations, which in turn influence the subjectively evaluated DVs. On the hypothesis of a causal interpretation of indicatives, similar psychological processes should be involved in evaluating the four central DVs. The model implements this by allowing the same structure across all four DVs. In addition, the model permits the rejection of this hypothesis by allowing the dashed edges to differ across the four DVs. Comparing models that set the dashed edges equal for some of the four main DVs thus provides a test of differences between these psychological constructs.

In the following, P(C|A) and the four main outcome variables were divided by 100, and the P(C|A) and reason relation were centered on their means. Furthermore, the Contingency factor (a, b, c, d) outlined in Table 8 was encoded in three indicator variables representing the following contrasts: (x1) a - b, (x2) c - b, and (x3) d - b. The model was fitted using the R-package `lavaan` (Rosseel, 2012).

P(counterfactual) P(singular) Power P(indicative)



P(C|A) ~ x1 + x2 + x3 + machine, $R^2 = 33.3$	$R^2 = 29.9$	$R^2 = 35.7$	$R^2 = 33.7$
Reason ~ x1 + x2 + x3 + machine + P(C|A), $R^2 = 30.4$	$R^2 = 30.9$	$R^2 = 27.3$	$R^2 = 31.6$
y ~ x1 + x3 + machine + Reason, $R^2 = 31.9$	*y* ~ x3 + machine + P(C|A)*Reason, $R^2 = 54.9$	*y* ~ x1 + x3 + machine + Reason + P(C|A), $R^2 = 63.6$	*y* ~ x1 + x3 + Reason* P(C|A), $R^2 = 82.2$

Figure 8. SEM model. 'P(C|A)*R' = two-way interaction of mean-centered P(C|A) and Reason. Contingency contrasts: 'x1' = a – b; 'x2' = c – b; 'x3' = d – b. Only statistically significant effects (*p* < .05) are shown. The regressions for the two mediators (P(C|A), Reason) are fixed to have the same regression coefficients across groups.

The model in Figure 8 was arrived at by trimming down a saturated model. We did this through a combination of domain knowledge, statistical tests, and by introducing equality constraints between coeffiecients of the predictors of the four main outcome variables. Only statistically significant paths are displayed and were retained. Figure 8 shows that, except for counterfactuals, the linear models of the main outcome variable, y_j, were in each case capable of accounting for more than 50% of the total variance. In the case of indicative conditionals, the model accounted for over 82% of the variance. Global fit statistics moreover indicated that the covariance matrix predicted by the model did not significantly misfit the data, $\chi^2(59) = 73.56$, $p = 0.096$, and that the model met widely used benchmarks for fit measures in SEM modelling (Finch & French, 2015; Kline, 2016): RMSEA = 0.026, 90% CI [0.00, 0.043], $p_{\varepsilon_0 \leq .05} > 0.99$, CFI = .996, SRMR = 0.037, AIC = 1306.85, BIC = 1926.30.

What enabled this model to do comparably well was by imposing differences between the four main DVs corresponding to the dashed edges in the conceptual diagram (Figure 7) and as illustrated in the diagram of the fitted model (Figure 8). In contrast, imposing the constraint that all four main DVs were identical resulted in a model that significantly misfit the data, $\chi^2(68) = 377.18$, $p < 0.001$, and which performed worse in terms of the fit vs. parsimony trade-off, AIC = 1592.47, BIC = 2164.27. Similarly, imposing the constraint that the evaluation of indicative conditionals was identical to causal power and singular causation led to an inferior model that significantly misfit the data, $\chi^2(66) = 175.09$, $p < 0.001$, AIC = 1394.38, BIC = 1976.77. Finally, imposing the constraint that only the evaluation of indicative conditionals and causal power were identical led to an inferior model that significantly misfit the data, $\chi^2(61) = 88.92$, $p = 0.011$, AIC = 1318.21, BIC = 1927.06. Of the latter three, the last was, however, the most competitive. But it still failed to capture the differences between indicative conditionals and causal power displayed in Figure 8.

Across the 16 Contingencies × DV conditions, the main outcome variables were consistently rated higher on the 0-100% scale in the Machine Group than in the Regularity group ($M_{difference}$=17.57, SD = 5.25). Figure 8 shows that this effect was in part mediated through the influence of the Machine factor (0 vs. 1) on the reason relation assessment and the Ramsey test assessment of P(C|A). In addition, Figure 8 shows that all four dependent variables were influenced by the contingency and machine manipulations.

A further finding in Figure 8 is that while the reason relation assessment affected all four main dependent variables to varying degrees, the Ramsey test assessment of P(C|A) did not affect the evaluation of the counterfactual “if A had not been the case, then C would not have been the case”. Finally, a moderation of P(C|A) by qualitative reason relation assessments was only found for singular causation judgments and indicative conditionals. We test this moderated mediation effect below.

Ramsey Test Conditional Probabilities and Causal Power

It was found that participants' Ramsey Test conditional probabilities, $P(C|A)_{DV}$, were sensitive to the Machine vs. Regularity manipulation. Participants in the Machine Condition tended to overestimate $P(C|A)_{DV}$ when $P(C|A)_{design}$ = low (conditions: *b*, *c*), $\bar{x}_b = .43$, $t(185) = 6.67$, $p < .0001$, $\bar{x}_c = .48$, $t(176) = 9.86$, $p < .0001$. Conversely, participants in the Regularity Condition tended to underestimate $P(C|A)_{DV}$ when $P(C|A)_{design}$ = high (conditions: *a*, *d*), $\bar{x}_a = .65$, $t(166) = -5.82$, $p < .0001$, $\bar{x}_d = .66$, $t(201) = -5.55$, $p < .0001$. These divergences from the manipulated conditional probabilities are illustrated through the distances to the dashed lines in Figure 9.

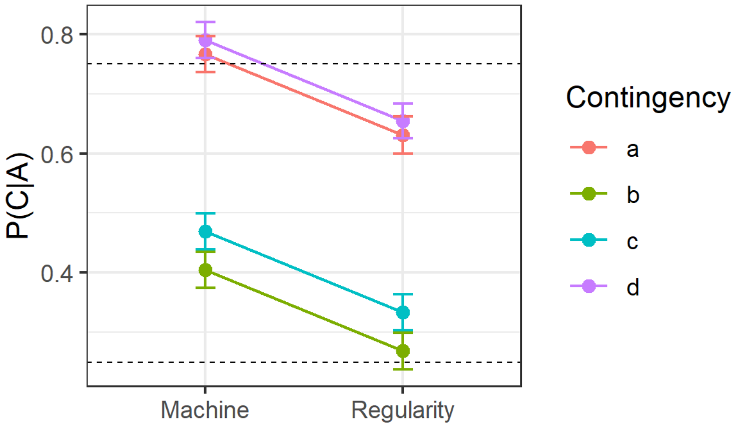


Figure 9. Ramsey Test conditional probabilities across conditions. The dashed lines indicate the manipulated conditional probabilities through the experimental design (see Table 8).

It was, moreover, found that $P(C|A)_{DV}$ was highly correlated with participants' estimates for causal power, $power_{DV}$: $r = 0.90$, $t(367) = 38.56$, $p < .0001$. Controlling for the other predictors shown in Figure 8, $P(C|A)_{DV}$ continued to be a significant predictor of causal power, $b = .60$, $z = 13.65$, $p < .0001$.

The high correlation between Ramsey Test conditional probability and $power_{DV}$ could be interpreted as follows. In the pilot study, it was found that many participants produced zero physically plausible, independent, alternative causes (>39%) in both the Machine and the Regularity conditions when prompted. This finding could in turn be interpreted as supporting van Rooij and Schulz's (2019) hypothesis that participants treat conditional probabilities as equal to causal power because they tend to ignore alternative causes. Such a tendency would count as a bias, insofar as participants also see trials in which the effect occurs in the absence of the target cause. In these trials the effect must be attributed to alternative causes.

However, a comparison with the manipulated conditional probability and causal power through the Contingency conditions invites a different interpretation. Based on Power$_{\text{design}}$, the pattern that would be expected for the indicator variables (x1, x2, x3) encoding the Contingency conditions is shown in Table 9:

Table 9. Comparison of causal power and P(C|A)

Indicator	Contingency	Power$_{\text{design}}$	Sign	P(C|A)$_{\text{design}}$	Sign
x1	a – b	.50 - .25 = .25	+	.75 - .25 = .50	+
x2	c – b	0 - .25 = -.25	-	.25 - .25 = 0	0
x3	d – b	0 - .25 = -.25	-	.75 - .25 = .50	+

Note. See Table 8 for the Contingency conditions.

As Table 9 shows, the predicted signs of the causal power estimates for the contingency contrasts are: +, -, -. In contrast, the predicted signs for the conditional probability estimates: +, 0, +. Figure 10 displays the signs of participants' causal power estimates in the data. More specifically, Figure 10 shows the total effects of x1, x2, and x3 on power$_{\text{DV}}$, along with the proportion that is mediated through P(C|A)$_{\text{DV}}$ alone:[81]

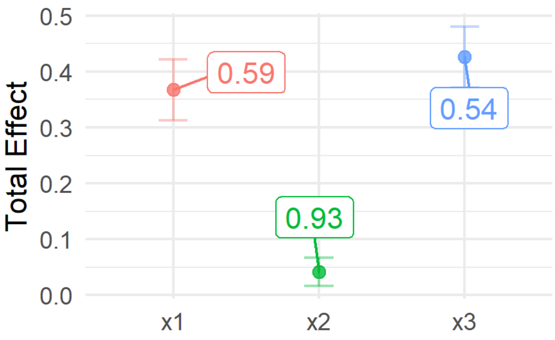


Figure 10. Total effect of x1, x2, and x3 on power$_{\text{DV}}$. The proportion of the total (positive) effect that is mediated through P(C|A)$_{\text{DV}}$ alone is labelled.

As is clear from a comparison between Figure 10 and Table 9, the magnitudes and signs of the total effects of x1, x2, and x3 on Power$_{\text{DV}}$ are more compatible with P(C|A)$_{\text{design}}$ than with Power$_{\text{design}}$. Had participants estimated Power$_{\text{DV}}$ based on the causal power calculated by the actual trials shown, the effects of x1, x2, and x3 would have had to follow the following order: +, -, -. Instead, the effects of x1, x2, and x3 followed the order of the manipulated conditional probabilities: +, 0, +.

[81] Note that there is a slight imprecision in these numbers due to the indirect effects of x1 (b = -.009, 95% CI [-.018, -.001]), x2 (b = -.006, 95% CI [-.011, .000]), and x3 (b = -.013, 95% CI [-.023, -.002]) through the mediator, Reason, with opposite signs. But these adjustments are so slight that they do not impact the interpretation substantially and they are thus ignored in the total (positive) effects displayed below.

The results therefore suggest that the high correlation between Ramsey Test conditional probability and power$_{DV}$ is to be accounted for by participants' causal power estimates as follows: Participants appear to be more sensitive to manipulations in conditional probabilities (P(C|A)$_{design}$) than to variations in the manipulated causal power (Power$_{design}$) in our experimental task. A simulation study in Appendix B shows what the expected relationship is between conditional probabilities and causal power for different types of statistical analyses. In the General Discussion, we will return to this issue and interpret van Rooij and Schulz's (2019) hypotheses in light of these results.

Moderated Mediation

To test for the influence of reason relation assessments on the indirect effects of Ramsey test conditional probabilities, a moderated mediation analysis was conducted (Hayes, 2018). In this analysis, the Reason factor is used as a moderator of the mediation by the Ramsey Test conditional probabilities. The Reason factor was recoded from its measurement on a five-point Likert-scale to values between 0 and 1: strong reason against (0.2), reason against (0.4), neutral (0.6), reason for (0.8), and a strong reason for (1.0). By trimming down a saturated model, the coefficients were constrainted to be zero for counterfactuals and causal power. Moreoever, as shown in Figure 8, the coeffeicients were set to be equal for singular causation and indicative conditionals. It was found that there was a significant interaction between P(C|A)$_{DV}$ and Reason$_{DV}$ for singular causation and indicative conditionals, $b = .34$, $z = 3.64$, $p < .0001$. In addition, it was found that there was a conditional effect of P(C|A)$_{DV}$ on causal power, $b = .34$, $z = 3.64$, $p < .0001$. Following Hayes (2018), the indirect effect of the experimental conditions through P(C|A)$_{DV}$ on singular causation and indicative conditionals can be viewed as moderated by Reason$_{DV}$, whenever a bootstrap interval of the index of partial moderated mediation does not cross zero (as found in Table 10).

Table 10. Indices of moderated mediation

X_i → P(C|A) → y		Moderator	Index: a_ib_j	95% Bootstrap CI
X_i =	x1	Reason	.12	[.056, .19]
	x2	Reason	.022	[.005, .040]
	x3	Reason	.13	[.060, .21]
	machine	Reason	.047	[.020, .073]

Note. The index 'a_ib_j' is a product out of the regression coefficients of X_i in the mediator regression model (a_i) and the regression coefficients of the moderator on the indirect path (b_j) in the outcome regression model. A bootstrap interval is used, because it has been shown in previous studies that the assumption of normality is violated for this index (Hayes, 2018).

The moderated mediation of P(C|A) by reason relation assessments for the outcome variable, P(if A, then C), replicates the influence of reason relations on P(if A, then C) from Experiments 1 and 2.

6.4.3 Summary

The main findings of Experiment 3 were as follows: First, support for the conceptual layer hypothesis (H_3) could be obtained, because models that treated indicative conditionals and explicit causal constructs (i.e. singular causation and causal power) equivalently were found to significantly misfit the data. Differences between the four main outcome variables thus emerged, which are illustrated in Figure 8. Most notably, it was found that there was no direct effect of Ramsey Test assessments of P(C|A) on counterfactual conditionals (“If A had not been the case, then C would not have been the case”) and that the interaction between Ramsey Test conditional probabilities and reason relation assessments could only be found for indicative conditionals and singular causation judgments. In contrast, no such interaction occurred for causal power judgments. Secondly, it was found that Ramsey Test conditional probabilities and measured causal power were highly correlated. It was considered whether this correlation should be interpreted considering the findings of a pilot study showing that many participants failed to produce physically plausible, independent, alternative causes in both the Machine and the Regularity conditions when prompted. Yet, this interpretation was rejected due to the finding that the causal power judgments deviated strongly from the manipulated causal powers. Instead it was found that participants were more sensitive to manipulations of conditional probability in their causal power judgments than to variations in the manipulated causal power in our experimental task.

The general finding of higher ratings in the Machine condition than in the Regularity condition suggests that participants rely on structural information that go beyond mere observed covariances for the four main outcome variables, when mechanistic knowledge is available. This is in agreement with previous findings (see, e.g., Johnson & Ahn, 2017) but is here also found for indicative and counterfactual conditionals. Since our experimental task had this knowledge component, participants had to integrate background information with the observed trials to form their subjective responses (as modelled by the SEM in Figures 7 and 8). But importantly, it was found that participants’ evaluations of indicative conditionals could not be equated with their subjective judgments of explicit causal constructs in a between-subjects comparison.

6.5 Experiment 4: Singular Causation

The results of Experiment 3 displayed in Figure 8 indicate that singular causation and indicative/counterfactual conditionals are influenced by similar factors and mediational processes. Still, it was found that the evaluation of indicative conditionals is not equivalent to the processing of explicit causal notions. According to the hypothesis of multiple conceptual layers of causal understanding (H_3), it is expected that indicative and counterfactual conditionals capture separate components of causal relations. To test this hypothesis, the goal of Experiment 4 was to probe whether participants' singular causation judgments could be predicted by their acceptance of indicative or counterfactual conditionals in a within subject design. This within-subject design was adopted to test whether participants' singular causation judgments could be predicted by their acceptance of indicative and counterfactual conditionals.

6.5.1 Method

Participants

A total of 594 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 3. The final sample after applying the *a priori* exclusion criteria consisted of 330 participants. Mean age was 40.02 years, ranging from 18 to 74. 46.1 % of the participants were male. 67.88 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Procedure

Participants were randomly assigned to the same 4 Contingency × 2 Group between-subjects conditions as in Experiment 3. The procedure was identical to the one in Experiment 3 with one exception: in Experiment 4, only the Singular Causation, Indicative Conditional, and Counterfactual Conditional dependent variables were included, and these were manipulated within subject.

6.5.2 Results and Discussion

To test whether singular causation judgments can be predicted by the probabilities assigned to indicatives and counterfactuals, a within-subject comparison was conducted across the 8 Group (Machine, Regularity) x Contingency (a, b, c, d) conditions. Three regression models were compared (see Table 11 below). First, a model that predicts singular causation judgments based on the Group factor (Machine vs. Regularity) and the probability assigned to indicative conditionals alone (M_1). Secondly, a model that is like (M_1) but additionally includes the probability assigned to counterfactual conditionals as a predictor (M_2). Thirdly, a model that is like (M_2) but additionally controls for the influence of the Contingency factor.

Table 11. Singular Causation Judgments

Model		*b*	SE	*p*	R^2	AIC	BIC
M1	Intercept	.38	.037	< .0001	.30	102.12	117.32
	Indicative	.47	.048	< .0001			
	GroupRegular	-.15	.032	< .0001			
M2	Intercept	.14	.041	< .001	.45	21.61	40.61
	Indicative	.44	.042	< .0001			
	Counterfactual	.38	.040	< .0001			
	GroupRegular	-.076	.029	.0089			
M3	Intercept	.18	.051	< .001	.46	24.67	55.06
	Indicative	.41	.052	< .0001			
	Counterfactual	.39	.043	< .0001			
	GroupRegular	-.077	.029	.008			
	Contingencyb	-.064	.045	ns			
	Contingencyc	-.026	.043	ns			
	Contingencyd	-.050	.038	ns			

Note. The lower AIC and BIC values indicate that M2 is superior to M1 and M3 in light of the parsimoni vs. fit trade-off.

The model comparison favors (M_2). It was thus found that a model that includes participants' evaluations of counterfactuals was a better fitting model than one that only included indicatives (M_1). This suggests that both the ratings of indicative and counterfactual conditionals were needed to predict singular causation judgments. It was also found that including ratings of counterfactuals accounted for unique variance when including a model that controls for the influence of the experimental conditions (M_3). Thus, even if we take differences in presented contingencies into account, the relationship between singular causation judgments and indicative and counterfactual conditionals holds.[82]

Pearl (2000) and Pearl and Mackenzie (2018) have argued that there are three types of queries that represent different layers of conceptual understanding of causal relations, which can be expressed via conditionals, as we have seen. Here we have not tested interventions. But the results of Experiment 4 indicate that participants' predictive judgments (expressed via indicatives)—and their counterfactual comparisons (expressed via counterfactuals)—are good

[82] To control for random effects due to variation across participants in a mixed regression analysis, trial replications would be needed of the DV factor. This would require presenting participants with multiple machines analog to the mouse-wheel machine in Figures 3 and 4. While such an analysis would be desirable, it goes beyond the limits of the present investigation. Aggregating the data and fitting models corresponding to M_1 and M_2 lead to similar results favoring M_2 over M_1 in both Experiment 3 and 4.

predictors of their singular causation judgments. This finding is in line with the hypothesis that there are different layers of conceptual understanding of causal relations that can be expressed by natural language conditionals (H_3).

More broadly, the finding that counterfactual judgments influence singular causation judgements is in line with causality theories from philosophy (Goodman, 1947; Lewis, 1973; Collins, Hall, & Paul 2004), computer science (Pearl, 2009), and statistics (Morgan & Winship, 2018; VanderWeele, 2015) emphasizing the close connection between counterfactuals and singular causal relations. Finally, in their accounts of singular causation, Pearl (2009, Ch. 10) and Halpern (2019) both build in counterfactual conditions in agreement with our results.

6.6 Experiment 5: Common Cause

The results of Experiment 4 suggest that there is more to the acceptance of a causal relation than the endorsement of indicative conditionals. In Experiment 5, this point was further corroborated through the investigation of a common-cause structure with two correlated effects. The use of such common-cause models permitted us to contrast probabilistic dependencies based on spurious correlations with probabilistic dependencies based on direct causal influence.

To further test the hypothesis (H_3) that reasoners grasp multiple conceptual layers of causal relations, Experiment 5 made a direct comparison of the acceptance of indicatives and non-backtracking,[83] interventionalist counterfactuals. Through the common-cause structure, we investigated the contrast between these two types of conditionals in the presence and absence of direct causal relations relating their antecedents and consequents. In our experimental task, the following common-cause version of the mouse wheel machine was implemented. First, a purple bowling ball drops on a mouse wheel, which sets off two sequences of events. One terminating with a yellow basketball following down. Another terminating with a red basketball falling, as shown in Figure 11 below:

[83] In back-tracking counterfactuals, one engages in abductive reasoning and starts reasoning backwards from, e.g., the non-occurrence of an event to the non-occurrence of its typical cause. When modelling interventions in a causal system, this type of reasoning is blocked in Pearl (2009). Pearl achieves this by the stipulation that the intervention sets a variable to a given value while removing the causal influence of variables that would normally have affected it.

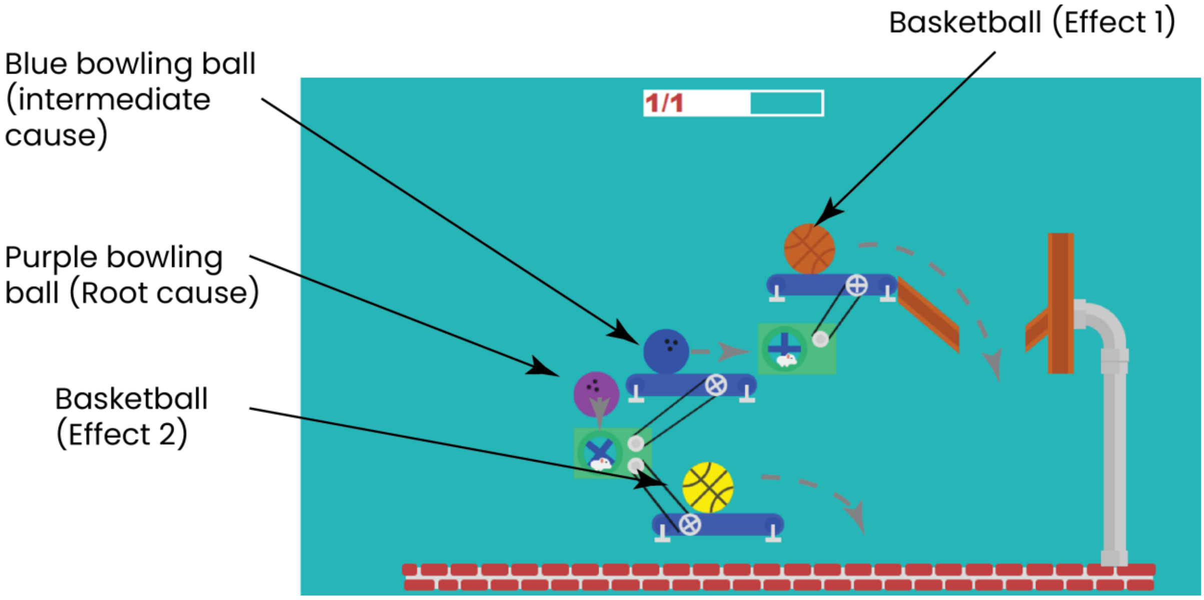

Blue bowling ball (intermediate cause)
Purple bowling ball (Root cause)
Basketball (Effect 2)
Basketball (Effect 1)
1/1

Figure 11. Annotated illustration of a common-cause trial in which the whole mechanism is visible. Instead of the annotation, participants saw animated trials. See https://osf.io/fa9rj/ for a video illustration.

In this common-cause scenario, causal relevance and probabilistic relevance come apart. The reason is that there was a probabilistic dependence between the yellow and the red basketballs, which was not grounded in a direct causal relation. So, although events with the yellow basketball is *relevant* to the probability of the red basketball falling down, the yellow basketball is not a cause for this and is thereby *not causally relevant*.

According to Lassiter (2017), the causal irrelevance of the yellow basketball for the red basketball is decisive for probabilistic counterfactuals. Yet, Lassiter holds that it should play no role for probabilistic indicatives, in line with the following hypotheses:

> (H_4) indicatives that *support* and indicatives that *do not support* counterfactuals can be empirically distinguished,
>
> (H_5) the use of indicatives and the acceptance of causal relations can be dissociated even in causal scenarios.

To examine these hypotheses, Experiment 5 contrasts indicatives and counterfactuals in predictive, diagnostic, and common-cause conditions. We moreover compare the assessment of these conditionals with singular causation judgments in situations where the causal relation is either present or absent. Our goal was to test for possible dissociations between the acceptance of indicative and counterfactual conditionals.

6.6.1 Method

Participants

A total of 949 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 4. The final sample after applying the *a priori* exclusion criteria consisted of 542 participants. Mean age was 40.16 years, ranging from 18 to 91. 39.48 % of the participants were male. 73.43 % indicated that the highest level of education they had was an undergraduate degree.

Design

The experiment had a mixed design. It contained one within-subject factor, DV (with three levels: indicative conditional vs. singular causation vs. counterfactual conditional). In addition, there were two between-subjects factors: Contingency (with four levels outlined in Table 12 below: a vs. b vs. c vs. d), and Condition (with three levels: predictive vs. diagnostic vs. common-cause). In total, 12 conditions were manipulated between subjects.

Table 12. Experimental design, contingency conditions

	P(E1|C)	$\Delta P_{E1,C}$	P(C|E1)	$\Delta P_{C,E1}$	P(E2|E1)	$\Delta P_{E2,E1}$
a	0.80	0.47	0.80	0.47	0.80	0.47
b	0.50	0.33	0.83	0.33	0.83	0.53
c	0.83	0.33	0.50	0.33	0.80	0.47
d	0.80	0.47	0.80	0.47	0.50	0.33

Note. The contingency conditions were introduced through the first initial trial and consecutive 15 randomly ordered blackbox trials. These trials were subject to the constraint that the last trial displayed was a <bowling ball, yellow basketball, red basketball> trial. This was done to enable participants to make singular causation judgments about whether the bowling ball caused the basketball to fall into the basket.

Materials and Procedure

Participants were randomly assigned to one of the 12 between-subjects conditions. The experimental procedure was similar to the one of Experiment 4. One difference was that Experiment 4 featured a group comparison between the machine vs. blackbox conditions. In contrast, in Experiment 5 all participants saw an initial common-cause machine trial (see Figure 11) and subsequently 15 blackbox trials, implementing the Contingency conditions outlined in Table 12. Because the common-cause version featured three events, there were eight possible combinations of events. To make the task less complex, participants were instructed in advance which of the three balls they should pay special attention to for answering the questions after the 15 blackbox trials.

Following these trials, participants were shown a block with two types of test questions (the dependent variables) in random order. One of these asked for the probability of an indicative conditional in one of the following three Conditions (predictive vs. diagnostic vs. common-cause), on a slider permitting continuous values between 0-100%.

Predictive Condition:
IF the purple bowling ball falls down, THEN the yellow basketball falls down.

Diagnostic Condition:
IF the yellow basketball falls down, THEN the purple bowling ball fell down.

Two Spuriously Related Effects of a Common-Cause:
IF the yellow basketball falls down, THEN the red basketball drops into the basket.

The second question asked for the probability of a counterfactual conditional. For the counterfactuals, participants were encouraged to imagine an intervention that would have prevented the antecedent from occurring:

> Imagine that we had prevented the purple bowling ball [/yellow basketball] from falling down (e.g., by constructing a safety net under it) [/e.g., by gluing it to the surface].

As a reminder, the statement describing the hypothetical intervention was displayed in grey on the following page. Participants were then asked to rate the probability of one of the following counterfactuals on a scale from 0 to 100% under this assumption:

Predictive Condition:
IF the purple bowling ball had NOT fallen down, THEN the yellow basketball would NOT have fallen down.

Diagnostic Condition:
IF the yellow basketball had NOT fallen down, THEN the purple bowling ball would NOT have fallen down.

Two Spuriously Related Effects of a Common-cause:
IF the yellow basketball had NOT fallen down, THEN the red basketball would NOT have dropped into the basket.

Following this block, participants were asked for singular causation judgments by assigning probabilities to the following statements:

Predictive Condition:

The purple bowling ball falling down caused the yellow basketball to fall down.

Diagnostic Condition:

The yellow basketball falling down caused the purple bowling ball to fall down.

Two Spuriously Effects of a Common-Cause:

The yellow basketball falling down caused the red basketball to drop into the basket.

6.6.2 Results and Discussion

To test whether participants' ratings for the indicative and counterfactual conditional statements were influenced by the Condition and Contingency factors, a mixed ANOVA was fitted to the data. The R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020) were used to this end. Condition (common-cause vs. diagnostic vs. predictive) and Contingency (a vs. b vs. c vs. d) were specified as between-subjects factors. DV (indicative vs. counterfactual vs. singular causation) was specified as a within-subject factor. The goal was to investigate possible dissociations between the probability of indicatives and counterfactuals within the levels of the Condition factor, in line with (H_4) and (H_5).

We found a significant three-way interaction between the Condition, Contingency, and DV factors, $F(11.30, 998.59) = 5.15$, $p < .0001$, $\eta^2_{\mathrm{G}} = .03$. In addition, a significant two-way interaction between the Condition and DV factors was found, $F(3.77, 998.59) = 44.66$, $p < .0001$, $\eta^2_{\mathrm{G}} = .07$. There were also significant simple effects of the DV factor, $F(1.88, 998.59) = 76.34$, $p < .0001$, $\eta^2_{\mathrm{G}} = .06$, and the Condition factor, $F(2, 530) = 123.01$, $p < .0001$, $\eta^2_{\mathrm{G}} = .20$.

The results are displayed in Figure 12 below:

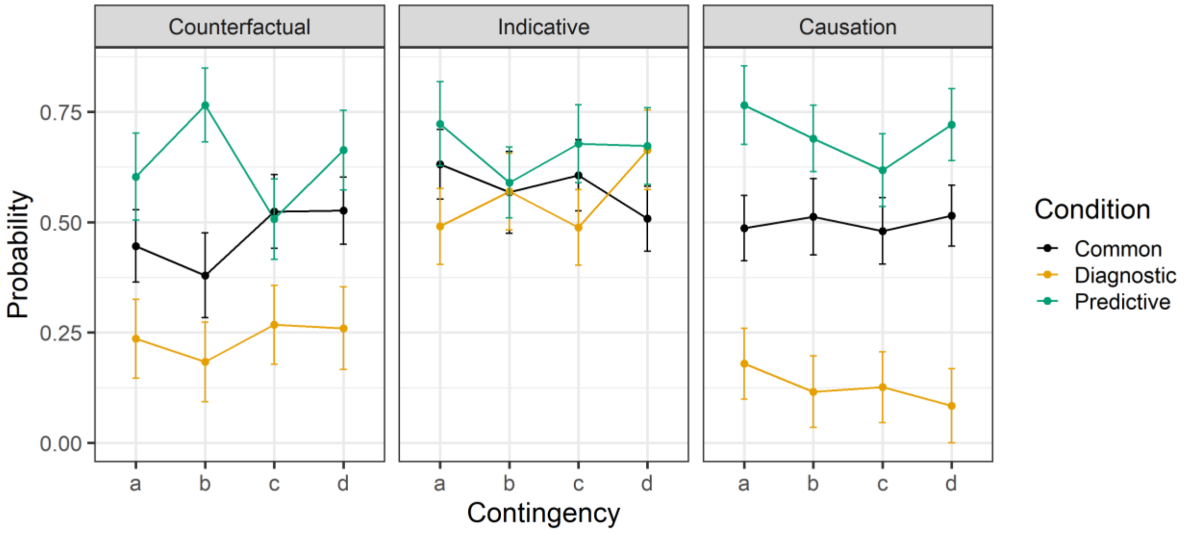


Figure 12. The three DVs are displayed across the 12 levels of the Contingency (a vs. b vs. c vs. d) × Condition (predictive vs. diagnostic vs. common cause) factors. ‘causation’ = singular causation judgment; 'indicative' = indicative conditional; 'counterfactual' = counterfactual conditional. The error-bars represent 95% CI intervals.

Most participants gave high ratings for the singular causation question in the predictive condition (*M* = 0.70, *SD* = 0.24) and low ratings in the diagnostic condition (*M* = 0.13, *SD* = 0.23). In contrast, they displayed more uncertainty about whether the two target variables were causally linked in the common-cause condition (*M* = 0.50, *SD* = 0.32), with 53 participants in the 3th quantile (≥ .75) and 54 participants in the 1th quantile (≤ .20).

Bonferroni-Holm corrected pairwise contrasts revealed dissociations between counterfactuals and indicatives. In the common-cause condition, counterfactuals were rated lower than the corresponding indicatives for Contingency a (*b* = -0.19, 95% CI [-0.31, -0.061]), *t*(530) = -3.58, *p* < .01) and Contingency b (*b* = -0.19, 95% CI [-0.33, -0.043]), *t*(530) = -3.10, *p* < .01). For the diagnostic condition, the same relationship was found for Contingency a (*b* = -0.26, 95% CI [-0.39, -0.12]), *t*(530) = -4.52, *p* < .0001), Contingency b (*b* = -0.39, 95% CI [-0.52, -0.25]), *t*(530) = -6.77, *p* < .0001), Contingency c (*b* = -0.22, 95% CI [-0.36, -0.086]), *t*(530) = -3.92, *p* < .001), and Contingency d (*b* = -0.40, 95% CI [-0.55, -0.26]), *t*(530) = -6.84, *p* < .0001). In the predictive condition, counterfactuals were rated higher than the corresponding indicatives for Contingency b (*b* = 0.18, 95% CI [0.049, 0.30]), *t*(530) = 3.32, *p* < .01) and lower than indicatives for Contingency c (*b* = -0.17, 95% CI [-0.31, -0.033]), *t*(530) = -2.97, *p* < .01).

Across Contingency conditions, it was found, on the one hand, that the three dependent variables were very similar in the predictive condition ($b_{counterfactual\ -\ indicative}$ = -0.03, 95% CI [-0.10, 0.038]), *t*(530) = -1.085, ns; $b_{indicative\ -\ causation}$ = -0.033, 95% CI [-0.09, 0.026]), *t*(530) = -1.34, ns; $b_{counterfactual\ -\ causation}$ = -0.064, 95% CI [-0.12, -0.0064]), *t*(530) = -2.67, *p* =

0.023). On the other, it was found that the three dependent variables differed increasingly in the common-cause condition ($b_{counterfactual\ -\ indicative}$ = -0.11, 95% CI [-0.17, -0.045]), t(530) = -4.09, p < .001; $b_{indicative\ -\ causation}$ = 0.080, 95% CI [0.025, 0.13]), t(530) = 3.52, p < .001; $b_{counterfactual\ -\ causation}$ = -0.030, 95% CI [-0.08, 0.024]), t(530) = -1.33, ns), and completely in the diagnostic condition ($b_{counterfactual\ -\ indicative}$ = -0.32, 95% CI [-0.39, -0.25]), t(530) = -11.06, p < .0001; $b_{indicative\ -\ causation}$ = 0.43, 95% CI [0.37, 0.49]), t(530) = 17.64, p < .0001; $b_{counterfactual\ -\ causation}$ = 0.11, 95% CI [0.053, 0.17]), t(530) = 4.65, p < .0001).

The results warrant the following conclusions. First, the acceptance of indicatives can clearly become dissociated from the acceptance of the corresponding counterfactuals and singular causation judgments corroborating (H_4) and (H_5). Secondly, it was found that counterfactual judgments tend to align with singular causation judgments. This in turn supports the hypothesis of a hierarchy of causal queries. On this view, singular causation judgments require affirmative answers to counterfactual queries (“does the consequent counterfactually depend on the antecedent?”), in addition to affirmative answers to the predictive queries (“is the antecedent a good predictor of the consequent?”) expressed by indicative conditionals.

The finding of a dissociation was most striking in the comparison between the predictive and diagnostic conditions. A factor contributing to this was the individual variation in whether participants accepted the existence of a direct causal relation in the common-cause condition. The use of blackbox trials may have made it more difficult for the minority who accepted a causal relation in the common-cause condition to distinguish between common-cause conditional and predictive conditionals.

Indicative conditionals can be acceptable both in the direction “if A, then C” and in the direction “if C, then A”. This is an indicator that indicatives do not themselves encode causal relations, but rather the inferential potential based on causal (and non-causal) probabilistic dependencies. Whereas causal relations are asymmetrical, our results are consistent with the probabilistic dependency between antecedent and consequents of indicative conditionals being symmetrical (Spohn, 2012a, Ch. 6; Skovgaard-Olsen, 2015).

In Ali et al. (2011), the alternative view is put forward that participants spontaneously recode causal relationships. Accordingly, the consequent can serve as the cause of the antecedent although the reverse direction would normally be expected. This recoding strategy is, however, challenged in cases like the one investigated in Experiment 5, where both the antecedent and the consequent are two effects of a common cause. It is worth noting, moreover, that Ali et al.’s (2011) case for the recoding hypothesis relies on indirect evidence from inference patterns, which showed deviations from participants' responses. We therefore regard this distinction between the spontaneous recoding hypothesis and the hypothesis of symmetry between antecedents and consequents of indicative conditionals (as introduced by the symmetry of probabilistic dependence) as a fruitful area for further inquiry.

6.7 Experiment 6: Control Study

The comparison between predictive and diagnostic conditionals in Experiment 5 involved a tacit comparison between a forward and backward temporal order of the antecedents and consequents. Yet, Experiment 5 only investigated common-cause conditionals in the forward direction, where the event mentioned in the antecedent occurred *before* the event mentioned in the consequent. To exclude possible confounds, Experiment 6 sought to contrast forward and backward common-cause conditionals within a single contingency condition. It was expected that a similar dissociation between indicative and counterfactual conditionals would be found in Experiment 6, and that this dissociation would be moderated by the temporal order (the antecedent occurring *before* vs. *after* the consequent).

6.7.1 Method

Experiment 6 followed the same method as Experiment 5 unless otherwise stated.

Participants

A total of 323 participants completed the experiment. The same sampling procedures and exclusion criteria were used as above. The final sample after applying the *a priori* exclusion criteria consisted of 166 participants. Mean age was 42.86 years, ranging from 19 to 74. 38.55 % of the participants were male. 77.71 % indicated that the highest level of education they had was an undergraduate degree.

Design

The experiment had a mixed design. The DV factor (with three levels: indicative conditional vs. singular causation vs. counterfactual conditional) was varied within subject. The Condition factor was varied between subjects (with four levels: predictive vs. diagnostic vs. common cause forward vs. common cause backward).

In contrast to Experiment 5, only Contingency a of Table 12 was used in Experiment 6. This contingency fixes the conditional probabilities and ΔP values of the examined conditionals to the same values: P(E1|C) = P(C|E1) = P(E2|E1) = P(E1|E2) = 0.80; $\Delta P_{E1,C}$ = $\Delta P_{C,E1}$ = $\Delta P_{E2,E1}$ = $\Delta P_{E1,E2}$ = 0.47. In total, 4 conditions were manipulated between subjects.

Materials and Procedure

In Experiment 6, the backward common-cause conditional was introduced:

> **Two Spuriously Correlated Effects of a Common-Cause, Backward:**
> IF the red basketball dropped into the basket, THEN the yellow basketball fell down.

In addition, Experiment 6 held the tense of all conditionals constant. Both the antecedent and consequents of all examined indicative conditionals were thus manipulated to be in past tense. To create a context of epistemic uncertainty suitable for indicative conditionals, participants were instructed for indicative conditionals that they had to evaluate these sentences with respect to a further unknown run of the animation. For counterfactuals and singular causation judgments, participants were instructed to evaluate the sentences while thinking back on the last trial that they had seen. Like in Experiment 5, this last trial was fixed to be a <bowling ball, yellow basketball, red basketball> trial.

6.7.2 Results and Discussion

An ANOVA with Condition (common-cause backward vs. common-cause forward vs. diagnostic vs. predictive) as a between-subjects factor and DV (indicative vs. counterfactual vs. singular causation) as a within-subject factor was fitted to the data. The R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020) were used. Like in Experiment 5, the goal was to investigate possible dissociations between the probability of indicatives and counterfactuals within the levels of the Condition factor, as a test of (H_4) and (H_5).

It was found that there was a significant two-way interaction between the Condition and DV factors, $F(5.79, 312.79) = 3.33$, $p < .01$, $\eta^2_G = .03$. In addition, significant simple effects of the Condition factor, $F(3, 162) = 23.58$, $p < .0001$, $\eta^2_G = .19$, and the DV factor, $F(1.93, 312.79) = 9.33$, $p = .0001$, $\eta^2_G = .03$, were found. The results are displayed in Figure 13 below:

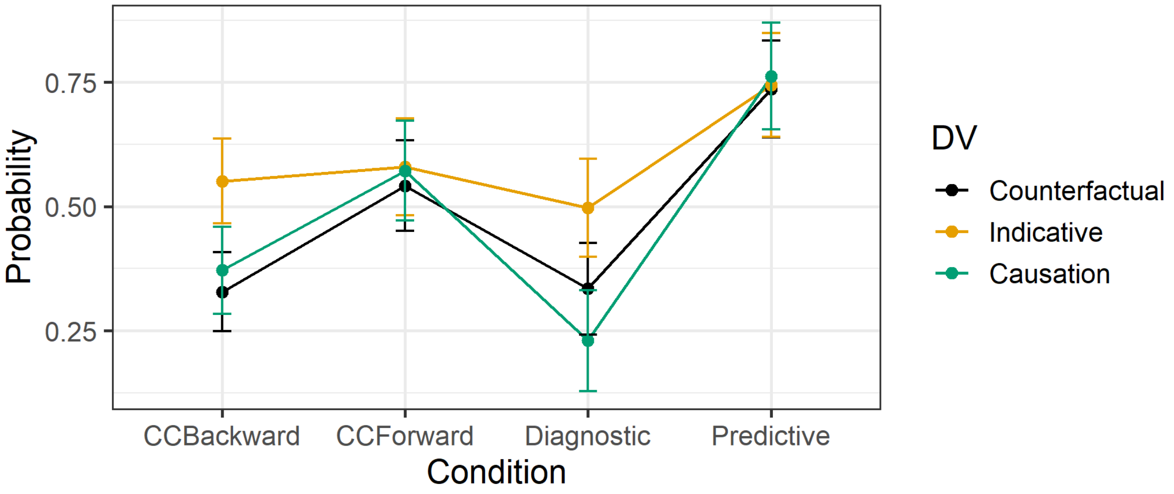


Figure 13. The three DVs are displayed across the 4 levels of the Condition factor. ‘CCBackward’ = common-cause backward; ‘CCForward’ = common-cause forward; ‘causation’ = singular causation judgment; 'indicative' = indicative conditional; 'counterfactual' = counterfactual conditional. The error-bars represent 95% CI intervals.

Bonferroni-Holm corrected pairwise contrasts revealed dissociations between counterfactuals and indicatives. Counterfactuals were rated lower than the corresponding indicatives in the common-cause backward condition (*b* = -0.22, 95% CI [-0.35, -0.098]), *t*(162) = -4.30, *p* = .0001) and in the diagnostic condition (*b* = -0.16, 95% CI [-0.31, -0.018]), *t*(162) = -2.71, *p* = .015).

Like in Experiment 5, the results warrant the following conclusions. First, the acceptance of indicative conditionals can become dissociated from the acceptance of the corresponding counterfactuals and singular causation judgments, in accordance with (H_4) and (H_5). Secondly, counterfactual judgments tend to align with singular causation judgments, in line with the hypothesis of a hierarchy of causal queries (H_3). But in contrast to Experiment 5, the dissociation of indicatives and counterfactuals was not found for forward common-cause conditionals. We attribute this difference of results to the procedural changes in Experiment 6, whereby past tense was adopted uniformly for the antecedents and consequents of all indicative conditionals.

One thing is striking about the results shown in Figure 13. Although the conditional probabilities and contingencies were identical for every condition, the indicative conditionals in the predictive condition were systematically higher than in all the other conditions. This finding may have resulted from the participants’ need to integrate their background knowledge, and knowledge of the mechanism from the first trial, with their learning experiences in the blackbox trials. Participants may thus have used assumptions about the underlying mechanism to provide clues about how stable the observed covariances were.

Alternatively, the finding could indicate that diagnostic and common-cause conditionals have different acceptability conditions than predictive conditionals. Accordingly, indicative conditionals in the diagnostic and common-cause conditions would have acceptability conditions that are systematically below the corresponding conditional probabilities even under positive contingency. Such a finding would be noteworthy, because it is not part of any of the main theories of indicative conditionals in the psychology of reasoning (see, e.g., Bennett, 2003; Evans & Over, 2004; Oaksford & Chater, 2007, 2010a; Rescher, 2007; Douven, 2015; Nickerson, 2015; Goodwin & Johnson-Laird, 2018).

Accordingly, Evans et al. (2007) state that "the Ramsey test predicts that belief in the conditional will be based on the probability of (q|p), regardless of the causal roles instantiated by p and q" (p. 639). To back this up, Evans et al. report evidence concerning predictive and diagnostic conditionals. Worth noticing in their results, however, is that the beta weight does change from .69 in the predictive conditional to .52 in the diagnostic conditional, when the probability of these conditionals is regressed on the corresponding Ramsey test conditional probabilities (see Evans et al. 2007, Table 2). This in turn would be consistent with the hypothesis of different acceptability conditions and the results reported here. Future research will have to determine whether the hypothesis of different acceptability conditions for various types of indicatives is correct in our trial-by-trial learning paradigm and in their paradigm.

6.8 General Discussion

The linguistic encoding of knowledge about causal relations plays a vital role for determining the basis for the cultural transfer of causal knowledge across generations. Causative verbs indicating the central contributing factor play a role in this transfer. An example is the verb “to break” in the example “the hammer broke the window” (Neeleman & van de Koot, 2012). Central among the linguistic constructions that facilitate the acquisition of causal knowledge are, moreover, natural language conditionals (Sloman, 2005, Ch. 11; Spohn, 2013). Conditionals play this role as a primary vehicle for expressing dependencies between variables (e.g., ‘if you hit it with a hammer, then it will break’). However, exactly which aspects of causal relations are linguistically encoded in indicative conditionals is still very much in dispute; with some authors interpreting recent findings of the role of probabilistic dependency as evidence for a causal interpretation, as we have seen. We will start by discussing what bearing our results have on that debate below and then turn to outlining a more general framework based on Pearl’s hierarchical theory of causation in which our various experimental findings can be interpreted in the remainder of the General Discussion.

Indicative Conditional, Causal Power, and the Relevance Effect

Experiment 1 followed previous studies (e.g., Skovgaard-Olsen, Singmann, et al., 2016, Skovgaard-Olsen, Kellen, et al. 2019) in replicating the Relevance Effect with verbal scenarios. In Experiment 2, it was found that the Relevance Effect could also be found in a trial-by-trial learning paradigm involving mechanistic knowledge.

Possible interpretations of the Relevance Effect reported by Skovgaard-Olsen, Singmann, et al. (2016) have played a role in recent work in the psychology of reasoning (see, e.g., van Rooij & Schulz, 2019; Oaksford & Chater 2020a, 2020b; Over & Cruz, 2018; Over, 2020). There is a strong temptation to interpret the Relevance Effect as indicating that indicative conditionals are often read causally as that the antecedent is a *cause* of the consequent (Oaksford & Chater, 2020a, 2020b; van Rooij & Schulz, 2019). The latter view connects with another broad theme; namely, the assumption that causal models underlie most of our subjective judgments of probability (Fernbach et al., 2011). On this view, causal models can thus provide a basic building block for the new paradigm in the psychology of reasoning by, *inter alia*, solving the puzzle of how the Ramsey Test is psychologically implemented. Accordingly, Evans et al. (2007) and Over (2020) both suggest that the Ramsey test is implemented via causal models.

In van Rooij and Schulz (2019), a further step was taken in connecting recent work on the probability of conditionals with theories of causal judgement. van Rooij and Schulz suggest that causal power can be used to account for the acceptability conditions of indicative conditionals (H_1). Since causal power in turn has been used to parameterize causal Bayes nets (Glymour, 2001; Fernbach et al. 2011; Oaksford & Chater, 2017; Aßfalg & Klauer, 2019), this hypothesis would directly show how the subjective probabilities of indicative conditionals could be based on causal models. In addition, van Rooij and Schulz (2019) also suggest as an auxillary hypothesis that participants' tendency to ignore alternative causes could explain why previous research has found evidence in support of [Eq1.] under some conditions (H_2). The reason being that causal power coincides with conditional probability whenever there are no alternative causes.

In line with this conjecture, it was found in the pilot study to Experiment 3 that 39.54% and 47.5% of the participants produced zero (plausible) alternative causes in the Machine condition and the Blackbox conditions, respectively. This finding might in turn explain why causal power and Ramsey test conditional probabilities were found to be highly correlated in Experiment 3 in the trial-by-trial learning paradigm.

To test van Rooij and Schulz's (2019) conjecture (H_2) directly, Experiment 1 made a between-subjects comparison of participants' judgments employing the verbal scenarios originally used to discover the Relevance Effect. In a pilot study preparing such a comparison, it was found, however, that participants had no trouble generating alternative causes for these stimulus materials both in the Positive Relevance condition and in the Irrelevance condition. In fact, participants tended to generate more alternative causes in the former condition than in the latter. They did this in spite of the fact that the irrelevance items presented participants with a candidate cause (e.g., Paul is wearing a shirt), which was patently useless for producing the effect (e.g., Paul’s car suddenly breaking down). To get a more direct critical test, we presented participants with the alternative causes that their peers had generated in a between-subjects comparison in Experiment 1. It made no difference for all the investigated effects whether participants were presented with alternative causes explicitly while making their judgments. These findings suggest that it is not the presence or absence of an accessible alternative cause that accounts for the Relevance Effect.

In a second direct test of van Rooij and Schulz's (2019) conjecture that causal power accounts for the acceptability of indicative conditionals (H_1), it was found in a model comparison in Experiment 1 that neither causal power nor Ramsey Test conditional probabilities alone could account for participants' ratings of P(if A, then C) across conditions. Instead, the analysis replicated Skovgaard-Olsen et al.'s (2016) finding that a model permitting P(C|A) to interact with the Relevance factor best accounted for participants' ratings. In Skovgaard-Olsen, Kellen, et al. (2019) patterns of individual variation in these results were investigated.

Given these negative findings, it is useful to return to the high correlation between causal power and Ramsey test conditional probability in Experiment 3. On closer inspection, it was found that participants' causal power ratings were more sensitive to the manipulated conditional probabilities than the manipulated causal power (see Table 9 and Figure 10). This could suggest that participants were biased in the other direction; by estimating conditional probabilities in a task designed to elicit their causal power judgments. Over et al. (2007, Experiment 2) also found that conditional probabilities calculated based on participants' responses were highly correlated with their ratings of causal strength (r = .87), and that the latter even correlated with probabilities of conjunctions to the same degree (r = .86). This finding, together with the much weaker associations of causal strength estimates with P(effect|¬cause), could also be interpreted as failures to give proper causal strength estimates in the investigated paradigms.

As a final option, one could adopt a causal power account but drop van Rooij and Schulz's (2019) auxillary assumption that participants' tendency to ignore alternative causes make them evaluate P(if A, then C) as P(C|A). In Appendix B, we investigate this possibility via a simulation analysis. Again, it is found that the simulation analysis did not turn out favorably for a causal power account of P(if A, then C).

Additionally, it was found in Experiment 3 that equating the evaluation of indicative conditionals with judgments of singular causation and causal power would result in a model that significantly misfit the data. In light of these various negative results (as well as further results discussed below), one must be careful not to make the slip from stating that the acceptability of indicative conditionals requires probabilistic dependency to the thesis that indicative conditionals are acceptable just in case there is causal relation between the antecedent and consequent. Instead, our results are consistent with the hypothesis (H_3) that causal relations involve a hierachy of causal queries, which goes beyond what is expressed by indicative conditionals alone.

Having dealt with causal power interpretations of indicative conditionals in relation to debates in the psychology of reasoning, we now turn to our remaining results and broaden our view by outlinning a general framework based on Pearl’s hierachical theory of causation in which our various experimental findings can be interpreted.

Learning Causal Relations Through Descriptions

According to Danks (2014): “A full account of causal learning from description remains an open research problem, particularly the question of when learners infer the absence of a causal relation (C does not cause E) from absence of information” (p. 68).

Several of our experiments can be interpreted as providing hints for constructing such an account. In Experiment 4 it was found that singular causation judgments could not be predicted by the probability of indicative conditionals alone. Instead it was found that the probability assigned to both indicatives and counterfactuals was needed to predict singular causation judgments. This finding already suggests that causal relations have multiple conceptual dimensions which are differentially encoded in indicatives and counterfactuals.

In our task involving counterfactuals, participants were asked to evaluate the probability of that the red basketball *would not* have fallen if the blue bowling ball *had not* fallen into the basket, after being shown a trial where both balls fell down. This type of task requires participants to evaluate the following counterfactual probability: P($Y_{X'}$ = false | X = true, Y = true). In words: under the assumption that both events actually occurred, what is the probability that Y would not have occurred had X not occurred. According to Pearl (2009), evaluating counterfactual expressions of this type is not possible based on causal Bayes nets, as illustrated in Appendix A. Instead the evaluation of counterfactual expressions requires a causal model with equations that represent in autonomous mechanisms of the data generating processes underlying directed edges, like in structural equation models (SEM), as we explain in Appendix A.

The counterfactual probability evaluates the causal *necessity* of the first event for the second event (*counterfactual query*). In contrast, predictive queries evaluate whether the occurrence of the antecedent is *sufficient for predicting* the consequent. By showing that singular causation judgments cannot be predicted by the acceptance of indicative conditionals alone, our results indicate that participants are sensitive to the counterfactual dimension of causal judgments.

For example, when a colleague says “Germany got the first wave of Covid-19 under control because of masks and social distancing”, and intends a causal interpretation, then this involves accepting the counterfactual, “If Germany had not introduced masks and social distancing, the first wave of Covid-19 would not have gotten under control”. In Spohn (2013, p. 1100), these sentences are taken as equivalent. Our results suggest that the colleague would also have to accept indicative conditionals like “if masks and social distancing are introduced, then Covid-19 will get under control” to make this type of causal attribution. The debate with the colleague over the causal attribution can then be focused on arguments concerning the acceptance/rejection of these indicative and counterfactual conditionals.

Corroborating the hypothesis of differential encoding of multiple conceptual dimensions, it was found in Experiments 5 and 6 that the probability of indicatives and counterfactuals could become dissociated in causal scenarios (H_4). This result was obtained by investigating diagnostic and common-cause conditionals in addition to predictive conditionals. Usually,[84] the focus in the psychology of reasoning has been on the acceptance of predictive, indicative conditionals. Theories have thus been formulated for the probability of indicative conditionals, which do not consider possible asymmetries between the probability of predictive, diagnostic, and common-cause indicative conditionals. Yet such asymmetries were found when holding P(consequent|antecedent) constant in Experiment 6.

Taken together, our finding of the need to predict singular causation judgments based on both indicatives and counterfactuals (Experiment 4) and the dissociations between the latter (Experiments 5 and 6) point in the same direction. They both suggest that one part of an account of causal learning from description may consist in subtle patterns of acceptance and

[84] One notable exception is Ali et al. (2010, 2011), which complement our results.

rejection of indicatives and counterfactuals. For instance, the speaker's unwillingness to assert “bad weather would not be coming, if the barometer had been prevented from falling” after having stated “if the barometer falls, bad weather is coming” would suggest that the speaker does not take his/her answer to a predictive query as supporting a causal relation.

Accordingly, the acceptance of an indicative conditional suggests that there is a symmetric, evidential relevance relation between two variables or propositions. But this does not yet imply that the evidential relationship is based on direct causation. As Edgington (2008, p. 18) observes, it is never contradictory to assert ‘If A happens, B will happen, but A won’t cause B to happen’. In contrast, the acceptance of interventionalist, non-backtracking counterfactuals suggests that there is an asymmetric, direct causal relation. This means that learners should be able to infer the absence of a causal relation from a verbal description indicating either that there is no probabilistic dependency (because the indicative rejected) or that it is a *mere* probabilistic dependency (because the counterfactual is rejected).

Oaksford and Chater (2010b, 2020) have suggested that conditionals describing inferential dependencies can be viewed as structure building operators in causal Bayes nets. The account we have unfolded above is in accordance with this general idea. But the hypothesis of differential linguistic encoding of causal relations through conditionals advanced in this paper opens up for more detailed investigations of the construction of causal models based on linguistic testimony. To illustrate, blackbox observations of three events may either correspond to a causal chain, a common-cause structure, or causal structures with hidden variables. Through indicative conditionals, the edges of the graph can be conveyed. Through the tense of the antecedents and consequents, temporal cues about the ordering of events can be given (e.g., “If it rains, then the streets will be wet” vs. “If the streets are wet, then it rained”). Such temporal cues can be used to infer the direction of edges. Moreover, the acceptance and rejection patterns of interventionalist, non-backtracking counterfactuals can be used to read off the direction of edges. For instance, in a situation where it rains and the streets are wet, “If we had built a pavilion, then the street would not have been wet” sounds acceptable, but “If we had built a pavilion, then it would not have rained” sounds off.

A further component of the ability to infer a qualitative causal structure is the ability to imagine a mechanism whereby cause and effects are related (Lagnado et al., 2007; Johnson & Ahn, 2017). Our use of the contrast between a Blackbox and a Machine condition led to the finding in Experiment 3 of higher ratings of the four examined outcome variables when mechanistic knowledge was available. This finding suggests that participants rely on structural information that go beyond mere observed covariances when evaluating both conditionals and explicit causal constructs like singular causation and causal power. Assumptions about the underlying mechanism provides clues about how stable observed covariances are and permit participants to make distinctions between predictive/diagnostic relationships and effects of a common cause as in Experiments 5 and 6.

Causal vs. Evidential or Informational Relevance

In Spohn (2010, 2012a, Ch. 14), the distinction between evidential and causal relevance is expressed through the attempt of explicating causal relations as a specific case of a generic reason relation. Pearl (2009) draws a parallel distinction as follows:

> *Informational relevance* is concerned with questions of the form: "Given that we know Z, would gaining information about X give us new information about Y?" *Causal relevance* is concerned with questions of the form: "Given that Z is fixed, would changing X alter Y?" (pp. 234-235, italics added)

The distinction between the evidential and causal relevance of factors also plays a role in distinguishing between purely predictive uses of regression approaches from causally interpreted models in statistics (Gelman & Hill, 2007; Kline, 2016; Pearl, Glymour, & Jewell, 2016; Shipley, 2016; Morgan & Winship, 2018). The distinction is moreover central in discussions over the opposition between evidential and causal decision theory (Hitchcock, 1993; 1996; Meek & Glymour, 1994; Pearl, 2009; Spohn, 2012b).

According to Danks (2014), the graphical models in Bayes nets "can be understood as compact representations of relevance relations, where different types of graphical models present different types of relvance (e.g., informational, causal, probabilistic, communicative)" (p. 39). In causal Bayes nets parameterized via base rates and causal power (see Figure 2), the directed edges represent relations of *causal relevance*. In contrast, in undirected graphical models, the edges represent symmetric, *evidential relevance* relations (ibid, Ch. 3). In cases of confounding arising via common-cause scenarios, and other cases of spurious correlations,[85] causal relevance and probabilistic relevance come apart. For a psychological theory of probabilistic reasoning, ΔP is often used to represent evidential relevance[86] and causal power can be used to represent causal relevance.

[85] In addition to spurious correlations created by a common cause, spurious correlations are introduced by conditioning on either a collider or the descendant of a collider in common-effects structures (Pearl et al., 2016).

[86] It should be noted, though, that the factorization of undirected graphical models permits the use of any non-negative function defined over the variables in a clique (Højsgaard et al., 2012), yet ΔP can take negative values. However, ΔP is only one of a larger class of

For a causal Bayes net like Figure 2, the parents of a variable represent all the variables that are directly causally relavant to the given variable (Spohn, 2010). Bayes nets are normally only used to encode variables that are at least unconditionally relevant to one another. Answering predictive queries in a Bayes net via conditionalization is therefore unlike the cases of missing-link conditionals, where conditionalization is applied to variables that are categorized as being completly unrelated. Hence, answering predictive queries based on Bayes nets is akin to making predictions based on reason relations.

To acknowledge the counterfactual dimension of causal relations, causal power can also be replaced with the following counterfactual notion of sufficiency: $P(Y_X = \text{true}|Y = \text{false}, X = \text{false})$. In words: under the assumption that both events did not occur, what is the probability that Y *would* have occurred *had X occurred*. This counterfactual concept of sufficiency is identifiable based on Cheng's (1997) account of causal power, provided that no confounding is present and that the cause is generative (Pearl, 2009, ch. 7). The counterfactual notion is, however, stronger than the evidential relationship that we take indicative conditionals to express. The reason is that evidential relevance does not require that the antecedent and the consequent are actually false, but only that the antecedent can be used to *predict* the occurrence of the consequent (as a sufficient reason for believing in the consequent).

Coming from linguistics, Lassiter (2017) puts forward the view that the causal irrelevance of a factor is decisive for probabilistic counterfactuals. At the same time, Lassiter argues that such causal irrelevance plays no role for probabilistic, indicative conditionals. Lassiter argues this point by considering the reversal of truth values of the counterfactual "If Fran had made her flight, it is likely that she would have died". Although this counterfactual would normally be considered true after a plane crash, Lassiter argues that its truth value reverses, when considering the manipulation of the causally relevant factor that Fran is a highly skilled pilot. In contrast, when evaluating the indicative conditional, "If Fran made the flight it is likely that she died", the fact that the plane crashed is held fixed. Varying information about Fran's skills as a pilot should therefore make no difference.[87]

Lassiter's (2017) formal linguistic analyses are in line with Pearl's (2009) idea of a hierarchy of causal queries. They are also congenial to the possibility of mapping natural language expressions of indicatives onto the processing of generic predictive queries and

confirmation measures (Crupi et al., 2007) and measures of covariation (Hattori & Oaksford, 2007), which all merit further empirical investigation.

[87] For a dissenting perspective see Over & Cruz (2019) and Over (2020), who hold that counterfactuals can "collapse" to indicative conditionals in examples of this kind.

counterfactuals onto the processing of distinctively causal, counterfactual queries, as we have done in the present study. The dissociations of the probability of indicatives and counterfactuals in Experiments 5 and 6 in situations where ratings of singular causation are low corroborate this hypothesis. These dissociations corroborate a conceptual distinction between indicatives that *support* counterfactuals and indicatives that *do not support* counterfactuals (H_4) due to the absence of direct causal relations.

Viewed from this perspective, it is worth highlighting that Kirk (2013) notes in his book on experimental design that scientific hypotheses share the characteristic that they "can be reduced to the form of an *if-then* statement. For example, "*If* John smokes, *then* he will show signs of high blood pressure" (p. 49). Kirk proceeds to explain how such if-then statements are to be evaluated through statistical hypothesis testing and confidence interval estimation. But it would have been highly controversial, if he had then gone on to state that these methods of classical statistics were themselves sufficient for establishing causal claims. For this, statistical methods for causal inference make use of procedures for evaluating counterfactuals (Morgan & Winship, 2018; VanderWeele, 2015; Pearl, Glymour, & Jewell, 2016). In addition, the experimental method investigates the scope for intervention, which can now also be emulated through Pearl's (2009) do-calculus based on observational studies.

In other words, validating a scientific hypothesis expressed as an indicative conditional is only the first step towards establishing a causal relation. In addition, it must also be established whether the probabilistic dependency that the conditional expresses can form the basis for intervention, whenever feasible. Secondly, it must be established whether it supports counterfactual conditionals, which are used for causal explanation (e.g., a colleague claiming that "Germany has gotten the first wave of Covid-19 under control *because of* masks and social-distancing"). In short, the assessment of causal relations requires probabilistic *prediction*, investigation of *intervention*, and counterfactually based *explanations*.

The low singular causation ratings in Experiment 6 to the backward common-cause and the backward diagnostic cases further suggest that participants recognize temporal precedence as a requirement of (direct) causal relevance. In both cases, where the antecedent occurred *later* than the consequent, very low singular causation judgments were obtained. Yet, the dissociation of these singular causation judgments with the probability of indicative conditionals suggests that participants accept that the antecedent may nevertheless be evidentially relevant for the consequent, in spite of its low (direct) causal relevance.

Finally, we contrast mental model theory with the account above and make some further comparisons.

Alternative Frameworks

On the newest version of mental model theory (Khemlani et al., 2018), indicatives are viewed as conjunctive assertions about possibilities as shown in table 13:

Table 13. Mapping between indicative and counterfactuals, MMT

Row	Partition		Factual: *If A then C*	Counterfactual: *If A had happened, then C would have happened*	Counterfactual with Neg.: *If A had not occurred, then C would not have occurred*
1	A	C	Possibility	Counterfactual possibility	Fact
2	A	Not-C	Impossibility	Impossibility	Counterfactual possibility
3	Not-A	C	Possibility	Counterfactual possibility	Impossibility
4	Not-A	Not-C	Possibility	Fact	Counterfactual possibility

Note. Quelhas et al. (2018) call indicative conditionals "factual conditionals". The last "Counterfactual with Negations" column was added here.

On this view, 'if the sun is setting, then the sky is red' makes a categorical assertion that it is *impossible* that the sun is setting and the sky is not red, and that it is *possible* that:

> the sun is setting and the sky is red,
> the sun is not setting and the sky red,
> the sun is not setting and the sky is not red.

In Johnson-Laird and Khemlani (2017), various causal relations are also explicated in terms of mental model theory. Interestingly, Johnson-Laird and Khemlani distinguish between a weak and a strong notion of causation. On the weak notion, 'the sun is setting causes the sky to be red' asserts the same three possibilities as 'if the sun is setting, then the sky is red'. The only difference is that the weak notion of causation imposes the temporal constraint that 'the sun is setting' occurs *before* 'the sky is red', whereas indicative conditionals would be compatible with either temporal direction. Hence, on mental model theory, the weak notion of causation is almost identical in meaining to indicative conditionals, but indicative conditionals need not express a causal relation.

In our account, we have emphasized that in addition to accepting indicative conditionals, and respecting a temporal order, counterfactual conditionals of the type 'if the sun had not set, then the sky would not have been red' should be accepted in causal attributions as well. The results from Experiments 4-6 have corroborated this view.

Inspecting Table 13, a special problem emerges for mental model theory in taking this finding on board. The problem is that while the indicative conditional asserts that it is *impossible* that the sun is setting and the sky is not red, the counterfactual with negated antecedent and consequent asserts that this is a *counterfactual possibility*. However, on the notion of impossibility that Johnson-Laird and Khemlani (2017, p. 170) adopt, there exist no possibilities in which an impossible proposition holds. But this means that in accepting an indicative conditional, 'if A, then C', and the counterfactual with negated clauses, 'if A had not occurred, then C would not have occurred', as part of causal attributions, one is depicted as inconsistently claiming *both* that A and not-C is a counterfactual possibility *and* that there are no possibilities in which A and not-C holds. We can therefore conclude that Pearl's hierarchy of causal queries does not sit well with the revised mental model theory.

In philosophy and linguistics, the possible worlds semantics of Stalnaker (1968) and Lewis (1973) remain popular alternatives. Pearl (2009, Ch. 7) showed that it was possible to use his account of interventions in causal models to explicate the elusive notion of similarity in Lewis (1973). In doing so, Pearl showed that it was possible to derive the same conditional logics based on his structural semantics for counterfactuals as on Lewis' account. On this logic, conditional sufficiency, or and-to-if inferences, are valid. For indicative conditionals, these types of inferences are, however, the focus of a recent controversy in the psychology of reasoning (Over & Cruz, 2018; Skovgaard-Olsen, Kellen, et al., 2019).

At the time of Nute (1980), they were already considered problematic for counterfactual conditionals. Accordingly, Nute (1980) discusses various ways of weakening possible worlds semantics into a logic, where they are invalid. Lewis (1973) earlier showed that he could apply his truth conditions for this logic as well if he allowed that other possible worlds could be *as similar* to the actual world as the actual world itself.

In a causal model, this would correspond to considering further possible values of the background variables characterizing the current situation than the ones actually instantiated, and calculating the effects of forcing the antecedent to be true under those circumstances as well. This could give rise to cases where the consequent is false leading to a failure of conjunctive sufficiency. It would be interesting to see if Pearl could follow the extensions of possible worlds semantics in Nute (1980) to evaluate the counterfactuals with respect to a set of sufficiently similar possible worlds in case the antecedent is true in the actual world.

Other conditional logics have been developed along these lines to avoid conjunctive sufficiency also for indicative conditionals. For instance, Vidal (2017) builds on Nute (1980) but introduces a two-stage impementation of the Ramsey Test that brackets the current beliefs and disbeliefs in the antecedent before evaluating the consequences of adding the antecedent to one's belief set. Simiarly, Rott (2019) has developed a logic for an expanded notion of the Ramsey Test to ensure that the antecedent is relevant for the consequent, which he suggests could either be part of the truth or acceptability conditions of indicative conditionals. See further Raidl (2020) for an overview of several such formal systems.

6.9 Conclusion

In sum, the evidence across the six experiments we reported is most consistent with the view that indicative conditionals encode inferential relations (as shown by the Relevance Effect, which was replicated in Experiments 1 and 2) and are used to answer predictive queries. Following Skovgaard-Olsen, Collins, et al. (2019), these inferential relations may be viewed as conventional implicatures. The results also suggest that there are multiple layers of conceptual understanding involved in causal relations that are differentially encoded in indicative and counterfactual conditionals, which has not been demonstrated before. Both the acceptance of indicatives and counterfactuals are required to predict singular causation judgments (Experiment 4). However, when the acceptance of indicative and counterfactual conditionals become dissociated (Experiments 5 and 6), the acceptance of counterfactuals track singular causation judgments and the (direct) causal relevance of the antecedent for the consequent. In contrast, indicative conditionals track evidential relevance.

Moreover, although causal power may be used to parameterize causal Bayes nets (Glymour, 2001), and its application to indicative conditionals can be theoretically motivated (van Rooij & Schulz, 2019), it turns out empirically that causal power does not fit our data for indicative conditionals (Experiments 1, 3, Appendix B). Instead, an account that assumes that participants make reason relation assessments using conditional probabilities while being sensitive to when the antecedent lowers or raises the probability of the consequent turns out to better account for our results. This is in line with the idea of indicative conditionals as answering predictive queries requiring evidential relevance without necessarily representing causal relevance.

References

Ali, N., Schlottmann, A., Shaw, A., Chater, N., and Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In: Oaksford, M. and Chater, N. (Ed.), *Cognition and Conditionals* (pp. 117-134). Oxford: Oxford University Press.

Ali, N., Chater, N., and Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403-18.

Arlo-Costa, Horacio (2007). The Logic of Conditionals. In E. N. Zalta (eds.), *The Stanford Encyclopedia of Philosophy* (spring 2016 Edition). Retrieved from: <http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/>.

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Andreas, H., & Günther, M. (2018). A Ramsey Test Analysis of Causation for Causal Models, *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy074

Aßfalg, A., & Klauer, K. C. (2019). Reasoners Consider Alternative Causes in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 1–61.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379-384.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bouton, M. E. (2016). *Learning and Behvaior: A Contemporary Synthesis.* Oxford: Oxford University Press.

Brandom, B. (1994). *Making it Explicit*. Cambridge, Mass.: Harvard University Press.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61-83.

Cheng, P. W., & Lu, H. (2017). Causal Invariance as an Essential Constraint for Creating a Causal Representation of the World: Generalizing the Invariance of Causal Power. In *Waldmann, M. R. (Eds.), The Oxford Handbook of Causal Reasoning* (pp. 65–84). Oxford: Oxford University Press.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, *104*(2), 367–405.

Collins, J., N. Hall, and L. A. Paul (Eds.) (2004). *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science*, *74*, 229-252.

Cruz, N., Over, D., Oaksford, M., & Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of the 38th annual conference of the cognitive science society (pp. 1104–1109). Austin, TX: Cognitive Science Society.

Cummins, D. D., Lubart, T., Alsknis, O., and Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition*, *19*, 274-282.

Cummins, D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646–658.

Cummins, D. D. (2014). The impact of disablers on predictive inference. *Journal of Experimental Psychology: Learn. Mem. Cogn., 40*, 1638–1655

Danks, D. (2014). *Unifying the Mind. Cognitive Representations as Graphical Models.* Cambridge, Massachusetts: The MIT Press.

Darwiche, A. (2009). *Modelling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press.

Douven, I. (2016). *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Edgington, D. (2008). I-Counterfactuals. *Proceedings of the Airtotelian Society, 108*(1), 1-21.

Elqayam, S. & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction. In S. Elqayam, J. Bonnefon, and D. E. Over (Eds.), *Thinking & Reasoning, 19* (3-4), 249-265.

Evans, J. St. B. T. (2020). The suppositional conditional is not (just) the probability conditional. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 57-70). London: Routledge.

Evans, J. St. B. T. and Over, D. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Handley, S., Hadjchristidis, C., Thompson, V., Over, D., & Bennett, S. (2007). On the basis of belief in causal and diagnostic conditionals. *The Quarterly Journal of Experimental Psychology*, *60*(5), 635–643.

Evans, J. S. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335.

Evans, J. S. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review*, *112*(4), 1040–1052.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning, *Psychological Science, 21*, 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185.

Fernbach, P.M., and Erb, C.D. (2013).A quantitative theory of conditional reasoning. *Journal* *of Experimental Psychology: Learn. Mem. Cogn., 39*, 1327–1343

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument and Computation, 4*(1), 64–88.

Finch, W. H., French, B. F. (2015). *Latent variable modeling with R*. New York: Routledge.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Glymour, C. (2001). *The Mind's Arrows, Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: the MIT Press.

Goodman, N. (1947). The Problem of Counterfactual Conditionals. *The Journal of Philosophy*, *44*(5), 113.

Goodwin, G. P., and Johnson-Laird, P. N. (2018). The Truth of Conditional Assertions. *Cognitive Science, 42*, 2502-2533.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review, 111*(1), 3–32.

Halpern, J. (2019). *Actual Causality*. Cambridge, MA: the MIT Press.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2. Edition). New York: The Guilford Press.

Hattori, M. and Oaksford, M. (2007). Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis. *Cognitive Science*, *31*, 765-814.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Hitchcock, C. (1993). A Generalized Probabilistic Theory of Causal Relevance. *Synthese, 97*, 335-364.

Hitchcock, C. R., (1996). Causal Decision Theory and Decision-Theoretic Causation. *Noûs*, *30*(4), 508–526.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). Graphical Models with R. New York: Springer.

Johnson, S. G. B., & Ahn, W. (2017). Causal Mechanisms. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 127–146). Oxford: Oxford University Press.

Johnson-Laird, P. N. and Khemlani, S. S. (2017). Mental Models and Causation. In M. R. Waldmann (Ed.)*, Oxford library of psychology*. *The Oxford handbook of causal reasoning* (p. 169–187). Oxford University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 169–187). Oxford: Oxford University Press.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science*, *42*(6), 1887–1924.

Kern-Isberner, G. (2001). *Conditionals in nonmonotonic reasoning and belief revision : considering conditionals as agents*. Berlin: Springer.

Kirk, R. E. (2013). *Experimental Design. Procedures for the Behavioral Sciences.* London: Sage Publications. (*4th Edition*)

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Edition). New York: The Guildford Press.

Kratzer, A. (2012). *Modals and Conditionals.* Oxford: Oxford University Press.

Krzyżanowska, K., Wenmackers, S., & Douven, I. (2013). Inferential Conditionals and Evidentiality. *Journal of Logic, Language and Information*, *22*(3), 315-334.

Krzyżanowska, K., Collins, P. J., & Hahn, U. (2017). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199–205.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. (2014). Causal Responsibility and Counterfactuals. *Cognitive Science, 37*, 1036-1073.

Lagnado, D. A., Waldmann, W. R., Hagmayer, Y., and Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (p. 154–172). Oxford: Oxford University Press.

Lassiter, D. (2017). Probabilistic language in indicative and counterfactual conditionals. *Proceedings of SALT, 27*, 525-546.

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science, 18*(11), 1014–1021.

Liu, M. (2019). Current issues in conditionals. *Linguistic Vanguard, 5*(3), 1-8.

Luhmann, C. C., & Ahn, W. (2005). The meaning and computation of causal power: A critique of Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112,* 685–692.

Manktelow, K. (2012). *Thinking and Reasoning: an Introduction to the Psychology of Reason, Judgment and Decision Making.* Hove: Psychology Press.

Mayrhofer. R. & Waldmann, M. (2015). Agents and Causes: Dispositional Intuitions As as a Guide to Causal Structure. *Cognitive Science*, *39*, 65-95.

Meder, B., Mayrhofer, R., and Waldmann, M. R. (2014). Structure Induction in Diagnostic Causal Reasoning. *Psychological Review, 121*(3), 277-301.

Meek, C. and Glymour, C. (1994). Conditioning and Intervening. *The British Journal for the Philosophy of Science*, *45*, 1001-1021.

Morgan, S. L. and Winship, C. (2018). *Counterfactuals and Causal Inference* (2th Edition). Cambridge: Cambridge University Press.

Neeleman, A., & van de Koot, J. (2012). The Linguistic Expression of Causation. In M. Everaert, T. Siloni, M. Marelj (Eds.), *The Theta System: Argument Structure at the Interface* (pp. 20-51). Oxford: Oxford University Press.

Nickerson, R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175-220.

Nickerson, R. (2015). *Conditional Reasoning*. *The Unruly Syntactics, Semantics, Thematics, and Pragmatics of "If"*. Oxford: Oxford University Press.

Nute, D. (1980). *Topics in Conditional Logic*. Dordrecht: Reidel.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (Eds.). (2010a). *Cognition and conditionals: probability and logic in human thinking*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2010b). Causation and conditionals in the cognitive science of human reasoning. *Open Psychology Journal, 3*, 105-118.

Oaksford, M., & Chater, N. (2017). Causal Models and Conditional Reasoning. In M. R. Waldmann (Eds.), *The Oxford Handbook of Causal Reasoning* (pp. 327–346). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2020a). New Paradigms in the Psychology of Reasoning. *Annual Review of Psychology*, *71*(1), 1–26.

Oaksford, M., & Chater, N. (2020b). Integrating Causal Bayes Nets and Inferentialism in Conditional Inference. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 116-132). London: Routledge.

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 680–693.

Over, D. E. (2020). The Development of the New Paradigm in the Psychology of Reasoning. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 243-263). London: Routledge.

Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In L. J. Ball and V. A. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 434– 450). Hove, UK: Psychology Press.

Over, D. E., & Cruz, N. (2019). Philosophy and the psychology of conditional reasoning. In A. Aberdein & M. Inglis (Eds.), *Advances in experimental philosophy of logic and mathematics* (pp. 225-249). London: Bloomsbury Academic.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2th Ed.). Cambridge: Cambridge University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plasuible Inference*. San Francisco: Morgan Kaufman Publishers.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics*. Sussex: Willey.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Baisc Books.

Pelley, M. E. L., Griffiths, O., and Beesley, T. (2017). Associative Accounts of Causal Cognition. In Waldmann, M. R. (Eds.), *The Oxford Handbook of Causal Reasoning* (pp. 13–28). Oxford: Oxford University Press.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*(2), 206–217.

Politzer, G., & Bonnefon, J. F. (2006). Two varieties of conditionals and two kinds of defeaters help reveal two fundamental types of reasoning. *Mind and Language*, *21*(4), 484–503.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2018). The Relation Between Factual and

Counterfactual Conditionals. *Cognitive Science*, *42*(7), 2205–2228.

Raidl, E. Definable Conditionals. *Topoi* (2020). https://doi.org/10.1007/s11245-020-09704-3

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107.

Rehder, B. and Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, *45*, 245-260.

Reips, U. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology, 49*(4), 243–256.

Rescher, N. (2007). *Conditionals*. Cambridge, MA.: The MIT Press.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36.

Rott, H. (1986). Ifs, though, and because. *Erkenntnis*, 25, 345–70.

Rott, H. (2019). Difference-Making Conditionals and the Relevant Ramsey Test. *The Review of Symbolic Logic*. doi: https://doi.org/10.1017/S1755020319000674

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139.

Sikorski, M., van Dongen, N. & Sprenger, J. (2019). *Causal Conditionals, Tendency Causal Claims and Statistical Relevance.* [Preprint] Retrieved from: http://philsci-archive.pitt.edu/16732/

Shipley, B. (2016). *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.

Singmann, H., Bolker, B., Westfall, J. & Aust, F. (2020). *afex: Analysis of Factorial Experiments.* R package version 0.28-0. https://CRAN.R-project.org/package=afex

Skovgaard-Olsen, N. (2015). Ranking Theory and Conditional Reasoning. *Cognitive Science*. *40*(4), 848-880.

Skovgaard-Olsen, N. (2016). Motivating the Relevance Approach to Conditionals. *Mind and Language*, *31*(5), 555–579.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, *150*, 26–36.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2017). Relevance and Reason Relations. *Cognitive Science*, *41*(S5), 1202–1215.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., & Klauer, K. C. (2019). Cancellation, negation, and rejection. *Cognitive Psychology*, *108*, 42–71.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, *126*(5), 611–633.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., & Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of “and”, “but”, “therefore”, and “if–then.” *Thinking and Reasoning*, *23*(4), 449–482.

Sloman, S. A. (2005). *Causal Models. How People Think about the World and its Alternatives*. Oxford: Oxford University Press.

Sloman, S. A. and Lagnado, D. A. (2005). Do We "do"? *Cognitve Science*, *29*, 5-39.

Spohn, W. (2010). The Structural Model and the Ranking Theoretic Approach to Causation: A Comparison. In Dechter, R., Geffner, H., Halpern, J. Y. (Eds.), *Heuristics, Probability and Causality. A Tribute to Judea Pearl* (pp. 493-508). San Mateo, CA: Kauffmann.

Spohn, W. (2012a). *The Laws of Beliefs*. Oxford: Oxford University Press.

Spohn, W. (2012b). Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box. *Synthese*, *187*(1), 95–122.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, *37*(6), 1074–1106.

Stalnaker, R. C. (1968). A Theory of Conditionals. In: Rescher, N. (Eds.), *Studies in Logical Theory (pp. 98-112)*. Oxford: Basil Blackwell.

Stephan, S. & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*, 242-257.

Vance, J., & Oaksford, M. (2020). Explaining the implicit negations effect in conditional inference: Experience, probabilities, and contrast sets. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000954

Vandenburgh, J. (2020). Conditional learning through causal models. *Synthese*. https://doi.org/10.1007/s11229-020-02891-x

VanderWeele, T. J. (2015). *Explantion in Causal Inference.* Oxford: Oxford University Press.

van Rooij, R., & Schulz, K. (2019). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

Vidal, M. (2017). A compositional semantics for 'even if' conditionals. *Logic and Logical Philosophy*, *26*, 237-276.

Vidal, M., & Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, *29*(6), 769–781.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol 34. Causal learning* (pp. 47–88). San Diego, CA: Academic Press.

Waldmann, M. R. (Eds.). (2017). *The Oxford Handbook of Causal Reasoning*. Oxford: Oxford University Press.

Waldmann, M. R., and Hagmayer, Y. (2005). Seeing versus Doing: Two Modes of Accessing Causal Knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216-227.

Walton, D. (2004). *Relevance in Argumentation*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Wilson, D. and Sperber, D. (2004). Relevance Theory. In Horn, L.R. & Ward, G. (eds.), The Handbook of Pragmatics. Oxford: Blackwell, 607-632.

Appendix A: Bayes Nets and SEM

We here illustrate the difference between causal Bayes nets and structural equation modelling (SEM) in Pearl's (2009) theory of causal inference. While Pearl (1988) earlier argued that one could explain causal inferences solely in terms of causal Bayes nets, he later revised this account due to the need for structural equation models for counterfactual reasoning (Pearl, 2009; Pearl & Mackenzie, 2018).

On Pearl's (2009) current account, there are three irreducible layers of conceptual understanding of causal relations: 1) statistical associations for predictive inference (which can be computed by conditionalization, e.g., via Bayes nets), 2) predictions based on interventions (which are observed through manipulations in randomized, experimental studies),[88] and 3) counterfactual inferences (which can only be computed based on structural equation models of the data generating processes, as we show below).

Bayes nets encode a set of conditional independence statements to simplify the specification of a joint probability distribution over a set of causally relevant variables (Darwiche, 2009), such as the following:

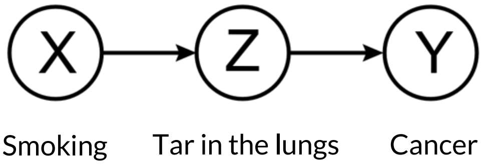
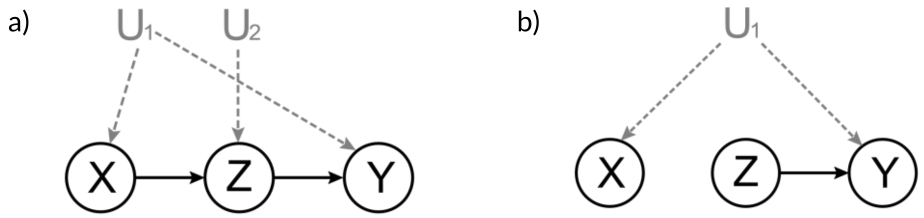


Figure A1. Bayes net representing a causal chain.

[88] In addition, these interventions can now also be computed by applying Pearl's (2009) do-calculus to observational studies (see also Morgan & Winship, 2018).

To illustrate their use in answering the queries above, the occurrence of the effect (e.g., cancer) can be predicted by conditionalizing on information about its possible causes (e.g., smoking), P(cancer|smoking). To evaluate the effect of an intervention (e.g., a hypothetical treatment designed to remove tar in the lungs), graph-surgery can be performed on the Bayes net. Graph surgery works by removing all incoming edges to the node intervened on, setting it to a given value (e.g., Z=0), and calculating the effects of the intervention on the descending nodes, P(cancer|*do*(tar=0)).

Finally, we can evaluate the counterfactual scenario in which we consider whether the the patient *would have been* cured, if the tar *had been* removed. But we need to make this evaluation while taking into account that the patient is in fact in a condition in which he has cancer and tar in his lungs. As a result, we need to be able to do both: 1) conditionalize on the factual information (cancer=1, tar=1) to update our distribution of the boundary conditions (U) representing the actual circumstances, and 2) perform graph surgery to calculate the effects of our counterfactual intervention. However, this latter step is not possible without structural equations representing the causal mechanisms underlying the causal diagram, which are shown below in Figure A2.



$$X = f_1(U_1) \quad Z = f_2(X, U_2) \quad Y = f_3(Z, U_1)$$

$$X = f_1(U_1) \quad Z = 0 \quad Y = f_3(Z, U_1)$$

Figure A2. Left-side: structural equation model of the causal chain in Figure A1 with structural equations determining the values of endogenous variables X, Y, Z as a function of their parents and the exogenous variables, U_1 and U_2, representing the boundary conditions. Right-side: sub-model obtained by performing graph surgery on a) by replacing the equation for Z with Z = 0 and removing all edges to Z.

In this case, the boundary conditions might be unknown factors influencing both the amount of tar in the patient's lungs (U_2) and whether the patient smokes and has cancer (U_1). The structural equations in Figure A2 are used to update the distribution of the boundary conditions based on the available evidence, P(U | smoke=1, cancer=1, tar=1). This updated distribution *remains invariant* when considering the counterfactual scenario in which an intervention is introduced to set tar=0, through graph surgery to generate the submodel displayed as b) above. Finally, the counterfactual probability, P(cancer=$0_{\text{tar=0}}$ | cancer=1, tar=1), is calculated based on both the updated distribution of the boundary conditions and the submodel, where the graph surgery has been applied (Pearl, 2009, Ch. 7). Since Bayes nets lack structural equations representing the influence of the boundary conditions, Bayes nets cannot handle cases, where we both update based on the evidence (cancer=1, tar=1) and consider *what would have* happened if tar *had been* 0 under the actual circumstances.

Formally, this double evaluation of 1) an update by factual information (cancer=1, tar=1) concerning the actual world and 2) computation of probabilities in counterfactual scenarios (cancer=0, tar=0) would give rise to inconsistency, if represented by standard probability theory via conditionalization alone. The use of structural models illustrated in Figure 2A prevents this by separating the update that is kept invariant between the models a) and b), and computating the counterfactual update in the submodel b) only. To represent this type of computational query, Pearl introduces the notation $P(Y_{X'} = \text{false} \mid X = \text{true}, Y = \text{true})$. In words: under the assumption that both events actually occurred ($P(\cdot \mid X = \text{true}, Y = \text{true})$),[89] what is the probability that Y would not have occurred had X not occurred ($P(Y_{X'} = \text{false} \mid \cdot)$).

Sometimes the term ‘Structural Causal Model’ (SCM) is used by Pearl to emphasize the integration of SEM as a statistical tool with causal graphs, a counterfactual semantics, and an explicit causal interpretation of strutural equations. Recent books on SEM have integrated many of these developments (see, e.g., Kline, 2016; Shipley, 2016). We provide further details on structural equation models, when we apply them as a statistical tool in Experiment 3.

Appendix B: Simulation Analysis, Causal Power

In Appendix B, we consider the option of adopting a causal power account while dropping van Rooij and Schulz's (2019) auxillary assumption that participants' tendency to ignore alternative causes make them evaluate P(if A, then C) as P(C|A). Instead, the causal power account of the acceptability of indicative conditionals could be strengthened by the observation that the equation in Cheng (1997) requires causal power and P(C|A) to be highly correlated for generative causes. This observation might in turn account for the positive association between P(if A, then C) and P(C|A). To examine exactly how strongly P(C|A) and P(if A, then C) would be associated on a pure causal power account, a simulation analysis was carried out with 488422 probability distributions generated through gridsearch (see Table B1, upper part):

[89] The dot, · , is here used as a placeholder for an event, proposition, or random variable.

Table B1. Simulation Analysis, Pure Causal Power Account

	r(Y, P(C|A))	r(Y, P(C|¬A))	m1: (Y ~ P(C|A))	m2: (Y ~ P(C|A) + P(C|¬A))
		Simulation		
Y = power	$r_{Y,P(C\mid A)} = .82$	$r_{Y,P(C\mid -A)} = 0.03$	β1 = .82,	β1 = 1.08, β2 = -.52
	$r_{Y,P(C\mid A).P(C\mid -A)} = .94$	$r_{Y,P(C\mid -A).P(C\mid A)} = -.79$	$R^2 = .68$	$R^2 = .88$
Y = ΔP	$r_{Y,P(C\mid A)} = .5$	$r_{Y,P(C\mid -A)} = -.5$	β1 = .5,	β1 = 1.0, β2 = -1.0
	$r_{Y,P(C\mid A).P(C\mid -A)} = 1.0$	$r_{Y,P(C\mid -A).P(C\mid A)} = -1.0$	$R^2 = .25$	$R^2 = 1$
		Experiment 1		
Y = If	$r_{Y,P(C\mid A).P(C\mid -A)} = .72$	$r_{Y,P(C\mid -A).P(C\mid A)} = .04$	β1 = .75	β1 = .74, β2 = .03
Y = power	$r_{Y,P(C\mid A).P(C\mid -A)} = .69$	$r_{Y,P(C\mid -A).P(C\mid A)} = -.42$	β1 = .61	β1 = .74, β2 = -.37

Note. The comparison is based on Positive Relevance conditions only. Upper half: correlation and least square regression analysis of simulated data based on 488422 probability distributions, which were generated meeting the criterion of Positive Relevance. Lower half: reanalysis of the Positive Relevance condition of Experiment 1 based on mixed regression models.

As the simulation shows, it is required that a causal power construct not only is strongly positively associated with P(C|A), β1 = 1.08, in a regression analysis, but also negatively associated with P(C|¬A), β2 = -.52. In Over et al. (2007), it was assumed that on a causal analysis, it would be required that the negative assocaition of P(C|¬A) with P(if A, then C) would be of the same magnitude as the positive association of P(C|A). However, as the simulation analysis shows, this constraint only holds for ΔP. In contrast, on a causal power account, the absolute magnitude of the positive association of P(C|A) is twice that of the negative association with P(C|¬A). Nevertheless, in previous studies—like Evans et al. (2007) and Over et al. (2007)—it was found that although weak, negative associations between P(C|¬A) and P(if A, then C) did occur, they were of a much smaller magnitude than the ones shown above.

In the lower part of Table B1, a reanalysis of parts of the data from Experiment 1 was carried out with the type of mixed regression model reported in Table 5. This type of model also contains a random intercept controlling for differences between scenarios, while estimating fixed, mean effects. The required negative association of P(C|¬A) with P(if A, then C) was not obtained for this subset of the data. P(C|¬A) was, however, negatively associated with causal power. Thus, like the model comparison in Table 5, this reanalysis did not turn out favorably for a causal power account of P(if A, then C).

Part III
Norm Conflicts and Individual Differences

Chapter 7:
Norm Conflicts and Conditionals[90]

Niels Skovgaard-Olsen,
David Kellen,
Ulrike Hahn,
Karl Christoph Klauer

Suppose that two competing norms, N_1 and N_2, can be identified such that a given person's response can be interpreted as correct according to N_1 but incorrect according to $N_{2.}$ Which of these two norms, if any, should one use to interpret such a response? In this paper we seek to address this fundamental problem by studying individual variation in the interpretation of conditionals by establishing individual profiles of the participants based on their case judgments and reflective attitudes. To investigate the participants' reflective attitudes we introduce a new experimental paradigm called the Scorekeeping Task. As a case study, we identify the participants who follow the Suppositional Theory of conditionals (N_1) versus Inferentialism (N_2) and investigate to what extent internally consistent competence models can be reconstructed for the participants on this basis. After extensive empirical investigations, an apparent reasoning error with and-to-if inferences was found in one of these two groups. The implications of this case study for debates on the proper role of normative considerations in psychology are discussed.

[90] This chapter has been published as:

Skovgaard-Olsen, N., Kellen, D., Hahn, U., and Klauer, K. C. (2019). Norm Conflicts and Conditionals. *Psychological Review*, *126*(5), 611-633.

Authors' Note: This work was supported by grants to Wolfgang Spohn and Karl Christoph Klauer from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). David Kellen received support from Swiss National Science Foundation Grant 100014_165591. Ulrike Hahn received support from the Humboldt Foundation.
Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).
Supplementary Materials: https://osf.io/9fm45/

7.1 Introduction[91]

In this paper we put forward an experimental framework for dealing with cases of conflicting norm in psychological research. This problem arises when multiple norms can be applied to reasoning tasks, which yield conflicting verdicts on what counts as correct reasoning. A good example is Wason's selection task (Wason, 1968), in which participants are asked to select which of four cards to turn over in order to find out whether a certain conditional rule (that is a rule with the structure ''if A, then C') is true or false. In its original version, Wason's task was only solved as intended by a small minority of the most cognitively able participants (ca. 10%). Many variations of this classical task have been explored in more than 300 published articles (Ragni, Kola, and Johnson-Laird, 2017). Most importantly, however, the exceedingly poor performance of participants observed by Wason prompted the development of alternative theoretical accounts that, based on information theory (Oaksford and Chater, 1994; Klauer, 1999) or a different semantics of the conditional (Baratgin, Over, and Politzer, 2013), recast the majority of the responses as rational. Recently, Elqayam & Evans (2011) criticized such developments by arguing that they involve a fallacious "is-to-ought' inference: one cannot infer from the fact that something is the case that it *should* be the case (e.g., the fact that cash payments to avoid taxes are common does not imply that tax avoidance is legitimate). In other words, descriptive facts about what *is* or is not the case do not license normative conclusions about what *ought* to be the case. This characterization of what have been extremely influential developments in the study of reasoning is a central plank in Elqayam and Evans' (2011) argument against a central role for normative considerations in the study of higher level cognition more generally. Elqayam and Evans argue that theories of higher mental processing would be better off if freed from normative considerations, not just in the area of reasoning, but also in judgement and decision-making.

This recommendation is not only at odds with long research traditions in those areas, but also comes after two decades of expansion of normatively oriented approaches and

[91] **Acknowledgement**: We would like to thank the audiences at the following conferences for their valuable feedback: European Society for Philosophy and Psychology, Rijeka 2018; Annual Meeting of New Frameworks of Rationality, 2017; Meeting of the Cognitive Science Society, 2017, London; Reasoning Club, 2017, Turin; International Conference of Thinking, 2016, Providence. We furthermore thank Shira Elqayam, Keith Stenning, David Over, Vincenzo Crupi, Katya Tentori, Wolfgang Spohn, Eric Raidl, Igor Douven, and Mike Oaksford for important discussions and students (esp. Hannes Krahl, Mareike Makosch, Johanna Weymann, Markus Steiner, and Lucy Ungerathen) for their help.

explanations within domains such as categorization, language processing, language learning, memory processes, and perception, in the form of ideal observer models (e.g., Geisler, 2011), Bayesian models of cognition (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010) or “rational analysis” (Anderson, 1991; Chater & Oaksford, 1998). It is thus unsurprising that Elqayam and Evans’ suggestions prompted vigorous debate (see, e.g,. the open peer commentary to Elqayam & Evans, 2011; or the papers in Elqayam & Over, 2016). This debate is itself part of a wider foundational discussion not just about psychological methods, but also about the quality and nature of psychological theorizing and explanation (see, e.g., Gigerenzer, 1998; Jones & Love, 2011; Bowers & Davis, 2012; Chater et al., 2006; Chater, 2009; Hahn, 2014; Chater et al., 2018).

In this paper, we seek to advance this debate by focusing on a central issue for normatively oriented theorizing across these areas, namely the issue of arbitration between competing norms with respect to participant performance. Specifically, we seek to provide both conceptual clarification vis a vis charges of fallacious is-to-ought inference and provide a novel methodological tool for use in these contexts. The tool is a new experimental task we have called the *Scorekeeping Task*, which is used in tandem with Bayesian mixture models to develop profiles of the participants at the individual level. We use this task in a case study: investigation of how individuals think about *indicative conditionals*, natural language statements such as ‘If I forget to pay the rent, then my landlord will complain’ that follow the general form ‘if A, then C’, as prompted by Wason’s original (1968) research. Through application of the Scorekeeping Task to a currently contentious issue in the study of conditional reasoning, we will show how this method defuses arguments about the inappropriate use of normative considerations, how it clarifies the respective roles of normative and descriptive considerations, and how it provides novel empirical and theoretical insights into a core question of how conditionals are represented and used by people.

The paper proceeds in three parts: In the first, we detail further the normative debate and conceptual issues. In the second part, we describe the empirical case study and its findings. In the third and final part, we discuss the wider implications not just to the study of reasoning but to examples of norm conflict in other areas of cognition.

The Normative Foundation

One common strategy in cognitive science consists of using normative theories as competence models describing the idealized knowledge possessed by an agent in a given domain (e.g., sentence parsing, deductive reasoning, or decision making) upon which processing is based. Since the competence models prove to be too efficient in solving the problems vis-á-vis psychologically realistic performance, they are augmented through independently testable assumptions about performance factors (e.g., working memory constraints) involved in applying the idealized knowledge, which may lead to performance errors (Cooper, 2002). A fruitful way to view the competence models of logic, probability theory, and decision theory is as providing *consistency conditions* on belief, degrees of belief, and choices, respectively (Chater and Oaksford, 2012). However, care needs to be taken since competing formal systems exist, for example, non-monotonic logic as an alternative to classical logic (Stenning and van Lambalgen, 2008), ranking theory as an alternative to probability theory (Spohn, 2012), and risk-weighted expected utility theory as an alternative to expected utility theory (Buchak, 2013). So what we can say is that each of these systems codifies one way of being consistent within their respective domains.

The normative foundation of our individual-profiling approach to the problem of arbitration has two legs to stand on. The first is the *Principle of Charity*, which says roughly that we should choose as a default interpretation the one that renders participants rational, when the data allow for a choice (Thagard and Nisbett, 1983; C. J. Lee, 2006). The second is a modification of Carnap's (1937) *Principle of Tolerance*. According to Carnap, only external, pragmatic reasons can be given for adopting a particular logical framework, but each logical system should be well-formed and come with its own framework-internal notion of what counts as correct reasoning (Steinberger, 2016). We have argued elsewhere that those ‘pragmatic reasons’ ideally need to be formally elucidated themselves (see, e.g., Corner & Hahn, 2013; Hahn, 2014), an issue we return to later in this paper. However, in this paper, we are not interested in making claims about the normative status of the formal theories *per se*. We note only that we believe, in general, that people may value different epistemic goods and so could rationally come to choose different rational norms. In keeping with this, our modified Principle of Tolerance permits different participants to adopt divergent norms when approaching a reasoning task.

Chater and Oaksford's (2012) focus on the consistency conditions imposed by normative theories is important since consistency makes up a minimal condition for any well-formed, formal system. So through the requirement that regardless of which reasoning system the participants adopt, it should at least be well-formed, we use internal consistency as a constraint on our competence models. One goal of the empirical investigations is then to probe how far we can succeed in reconstructing consistent competence models of the participants, when we charitably allow participants to adopt different norms. Our individual-profiling approach thereby assesses the participants only relative to a reasoning system that they have themselves committed to. In this, we follow Stenning and van Lambalgen (2004, 2008), who make the observation that competing logics (e.g., classical logic, intuitionistic logic, non-monotonic logic, deontic logic) can be represented as a choice of parameters like a) selection of formal language, b) its semantics, and c) a definition of valid arguments in the language. Their point is that before we can even begin to assess the performance of participants, we need to gain independent evidence of the participants' choices with respect to a), b), and c) in order to have a well-defined problem. Ultimately, their goal is to show that there is wide individual variation concerning these parameter settings, and that once we map out these sources of individual variation, much of what has been diagnostized as reasoning errors (e.g., in the Wason selection task) will diminish.

To take another much discussed example, in the literature on the *conjunction fallacy*, measures have been taken to ensure that participants have the right understanding of probability (Hertwig and Gigerenzer, 1999), and accept basic entailments ('A and B ⊨ A?') (Tentori, Bonini, and Osherson, 2004) that would commit them to the requirement that P(A and B) ≤ P(A). The present approach goes further by virtue of its focus on individual variation and in its recommendation that the attribution of reasoning errors should only be made based on independent evidence concerning the adherence of each individual to a given set of norms.

The moderate relativism underlying *relative* attributions of reasoning errors constitutes a radical departure from the tradition in psychology of designing experiments with one preconceived notion of correct reasoning. Such a moderate relativism is also found in the approaches of Elqayam (2012) or Stupple and Ball (2014). Our own approach differs from those in a number of ways, however. First, as this paper seeks to argue, we believe there is a unique role for normative theories in the study of cognition, whereas the grounded-rationality approach in Elqayam (2012) takes an essentially descriptive stance to psychology. Furthermore, whereas Elqayam (2012) holds that reasoning according to Bayes' rule is a normative requirement *only* for participants who adopt the epistemic goal of conforming to this rule, we maintain that this requirement may *follow* from other commitments that participants adopt.[92] As we will discuss below, one of the key arguments in the literature on the normative foundations of Bayesianism demonstrates how, for a particular measure of inaccuracy, minimizing the inaccuracy of one's beliefs requires "being Bayesian", that is, assigning subjective degrees of belief in line with the probability calculus and using Bayes'

[92] For instance, Costello and Watts (2014) argue that individuals will conform to the axioms of probability theory when generating probability estimates based on the count of retrieved instances as these conform to the basic principles of set theory that underlie probabilities.

rule for belief revision (Pettigrew, 2016). What is at issue here is a wider point: ‘norm endorsement’, as Elqayam envisions it, may indeed provide a basis for “ought”: “I ought to exercise, because I feel I ought to exercise” is one potential way of providing a descriptive basis for a normative claim in order to bridge the difficulty of *is* to *ought* inference (for more detailed discussion see Corner & Hahn, 2013). However, such endorsement or ‘norm adoption’ does not have to be bestowed in a piece-by-piece fashion, because putatively normative formal systems are exactly that, *systems*. This means that anyone who wishes to assign probabilities is, on some level or other, normatively committed to assigning coherent probabilities (i.e., in line with the axioms of probability theory, see, e.g., Jaynes, 2003), because that is what “probability” *means*. To illustrate with simple examples, someone who wishes to assign probabilities to events *must*, on some level accept the fact that the conjunction fallacy is an error, that is, a norm violation.[93] And this is true even for a resource limited cognitive agent who generates the conjunction fallacy only due to some internal noise (Costello & Watts, 2014), or because they are using a cheap and cheerful averaging strategy which suffices for their present needs given their aims and resource constraints (Juslin, Nilson & Winman, 2009). In other words, the reasoner might not *care* much about the error itself, or even be able to realistically do much about it, and such considerations should certainly be included in one’s evaluation of the system. But the conjunction error will still be an error, by virtue of the fact that the agent has agreed to assign probabilities in the first place.

These considerations reveal the fundamental role of consistency in evaluating not just reasoning, but also argumentation, judgment, or decision-making performance. Consequently, we constrain relativism on a theoretical level through the requirement that the competence models should be well-formed formal systems and should meet minimal consistency requirements, and that these systems ultimately have a well-founded pragmatic justification. And on a practical level, consistency is a cornerstone of our tests.

Eliciting Reflective Attitudes through the Scorekeeping Task

One way of guarding against attributing reasoning errors based on a mere case of miscommunication between the participant and the experimenter (Hilton, 1995) is to use the participants' considered judgments as a basis for the assessment. Tversky and Kahneman (1983) treated judgments as fallacies (as opposed to 'errors', or 'misunderstandings') only

[93] We are here using 'conjunction' as a technical term referring to a logical/probabilistic relationship rather than as referring to natural language "AND", which may be interpreted in different ways. For instance, ‘Kiss my dog and you’ll get fleas’ conveys the conditional meaning “If you kiss my dog you’ll get fleas” (Bhatt and Pancheva, 2006).

when participants were disposed to accept (after suitable explanation) that they had made a non-trivial, conceptual error; an error which the participants had the competence to avoid. In other words, Tversky and Kahneman (1983) consider it to be diagnostic of the presence of a fallacy that the participants could be brought to realize that they have made a mistake based on a conceptual misunderstanding.[94] Similar requirements concerning the need for the agents' considered judgments figure in the discussion of apparent violations of decision theory in Macnamara (1986), Spohn (1993), and Bermudez (2011, Chap. 2).

Implicit here is the assumption that it is the considered judgments/choices, or reflective attitudes, of a participant that reveals the normative principles that this person is committed to (Stein, 1996, Chap. 5). As part of a charitable assessment, it is therefore worth exploring new ways of designing experiments for eliciting participants' reflective attitudes.

One influential method of eliciting reflective attitudes is through *reflective equilibrium* (Goodman, 1965; Rawls, 1971). Reflective equilibrium is a method for arriving at considered judgments based on the coherence of case judgments and endorsed principles. The goal is to strike a balance between having to accept counterintuitive judgments of cases based on endorsed principles and judging contrasting cases in a way which can be consistently codified in a set of principles. In Spohn (1993), it is argued that normative principles are the outcome of a reflective equilibrium and that these normative principles enter into a wider reflective equilibrium with a charitable interpretation of the participants' responses. The method of reflective equilibrium is appropriate for eliciting considered judgments in academic disciplines but requires a level of cognitive resources that makes it less suited for naive participants (but see Stupple and Ball, 2014).

A different approach to eliciting the participants' reflective attitudes is adopted by Kneer and Machery (2019). In relation to moral judgments, they argue that isolated case judgments in between-subject designs are prone to the influences of performance errors like hindsight bias. As a solution, they propose a test of participants' moral competence based on the considered judgments they make when comparing multiple cases that differ in important conceptual dimensions (for related concerns, see Birnbaum, 1999). In addition, Kneer and

[94] But as pointed out by a reviewer, Tversky and Kahneman may not have implemented this requirement generally in their other work on cognitive illusions outside the conjunction fallacy. However, Slovic & Tversky (1974) adopted a related approach when studying paradoxes of decision theory, and more recently Keith Stanovich reviewed a body of research on participants' postexperimental endorsement of the rational principles they violated (Chater et al., 2018, pp. 811).

Machery also investigated the participants' endorsement of abstract principles and found it to be moderately correlated with their other measures.

Given well-known findings showing that participants often lack introspective access to the psychological processes that lead to their responses and tend to confabulate a rationalization if asked for the reasons behind their responses (for a review, see Evans, 2007, Chap. 7), we believe that participants' explicit avowals of normative principles is not by itself a reliable source. This also becomes vivid in the presence of moral dumbfounding when it is investigated whether people can provide reasons and articulate moral principles matching their judgments and endorsed principles (McHugh, McGann, Igou, & Kinsella, 2018).

Moreover, to avoid participants displaying one reflective attitude when presented with one pair of cases, and another when presented with a different pair with no attempt at integration, we seek to elicit commitments through the participants' own normative behavior. To do this, we introduce a novel scorekeeping task where we put participants in the position of judging how well their peers argued for their mutually incompatible responses and where we equip the participants with normative actions. The task of participants consists in applying sanctions and assigning burden of proofs to the one of their peers who has provided the weakest advocacy of his or her responses.

We take the commitments the participants adopt in this argumentative setting as binding, in the sense that they can be used as a basis for attributing reasoning errors to the participants. This is based on the simple principle that it is always appropriate to hold a person responsible to the norms that he/she uses to criticize her peers with–itself a kind of consistency requirement. For example, Brandom (1994) has argued that agents can be held responsible to comply with norms only insofar as they express some sort of recognition of being bound by these norms. In particular, Brandom has emphasized that one implicit way of recognizing boundedness to a norm, which does not rely on explicitly avowing normative principles, consists in criticizing and sanctioning others based on violations of this norm. This thought then opens up a new avenue of psychological research into which norms the participants hold their peers accountable to in argumentative settings (Skovgaard-Olsen, 2017). Moreover, it is very much in line with recent developments emphasizing that the evolutionary function of reasoning is argumentative: to devise and evaluate arguments intended for persuasion (Mercier & Sperber, 2011, 2017).

The experimental framework provided by the Scorekeeping Task is used as a means for probing into the *participants' own understanding of the task*, their goals in completing it, and their understanding of the logical concepts involved in it. Throughout the task, the participants' own reflective attitudes are elicited. This enables a comparison between the participants' reflective attitudes and their case judgments to investigate their agreement and to initiate a search for covariates that characterize the participants who are classified into different profiles of reflective attitudes and case judgments. Finally, reasoning errors can be defined and studied as cases in which the participants fail to comply with the logical consequences of the norms they hold their peers accountable to.

We next illustrate these various tools by putting them to use in a case study.

Case Study: Norms and the Interpretation of Indicative Conditionals

Research on conditionals appears in Elqayam and Evans's (2011) critique as one of the areas in the psychology of reasoning that is plagued by the existence of multiple normative accounts and seemingly fallacious 'is-to-ought' inferences. Therefore, it constitutes an ideal case study for our individual profiling approach.

Conditionals play a key role in reasoning and argumentation in general. For instance, when identifying the type of questions that are amenable to experimental research, Kirk (2013) notes in his book on experimental design that they "should be reducible to the form, *if A, then B*". But despite this prominence, the meaning of the natural language conditional is a matter of longstanding theoretical debate that is far from resolved, with many competing views (Nickerson, 2015). Our case study will contrast two of these views and seek to demonstrate tools for adjudicating between them. The non-specialist reader may simply take this fact at face value.

The first of the two normative perspectives on conditional reasoning examined here is based on the work of Adams (1965), Edgington (1995a), and Bennett (2004). According to this prominent view, the probability of an indicative conditional is evaluated by the *Ramsey Test*:

> RAMSEY TEST: to evaluate 'if A, then C' add the antecedent (i.e. A) to the background beliefs, make minimal adjustments to secure consistency, and evaluate the consequent (i.e. C) on the basis of this temporarily augmented background beliefs.

Quantitatively, this introduces the following equivalence prediction:

$$P(\text{if } A, \text{ then } C) = P(C|A),$$

which is referred to as the *conditional-probability hypothesis*.[95] This equivalence implies the inequality

$$P(\text{if A, then C}) \geq P(A,C),$$

as $P(C|A) \geq P(A,C)$ holds by probability theory.

Much of the recent work in psychology of reasoning has been strongly influenced by these views of the conditional (Evans & Over, 2004; Oaksford & Chater, 2007; Baratgin, Over, and Politzer, 2013; Pfeifer, 2013), which we will here refer to as the *Suppositional Theory of Conditionals* (henceforth ST). Inspired by the conditional probability hypothesis and the Ramsey test, Evans and Over (2004) express the view that 'if' is a linguistic device for triggering a process of hypothetical or suppositional reasoning. In addition, Evans and Over (2004) embed ST within a dual-process framework that seeks to distinguish heuristic and analytic processes. But here we just take ST as denoting the theses above, which share a wider appeal. Indeed, in a recent introduction to conditionals in cognitive science, the conditional probability hypothesis is presented as "fundamental" to a new probabilistic paradigm in cognitive psychology (Nickerson, 2015, p.199), and in Oaksford and Chater (2017) it is said to be "at the heart of the probabilistic *new paradigm* in reasoning" (p. 330).

The Ramsey Test was a direct source of inspiration for several further theories in belief revision and conditional logics (Arlo-Costa, 2007). For theories inspired by the Ramsey test, the TT cell of truth tables, where both the antecedent and the consequent take the value 'True', functions as a trivial instance in which the conditional is true. Testing whether the consequent is true under the supposition that the antecedent is true reduces to testing whether the consequent is true, whenever the antecedent is already known to be true. Accordingly, inferences from conjunctions ('A and C') to conditionals ('If A, then C'), the so-called *and-to-if inferences*, are valid for theories of conditionals based on the Ramsey test.

An example of an and-to-if inference is inferring '*if* Craig pays for the dinner, *then* Matthew will invite Craig out to the movies' from observing 'Craig paying for the dinner *and* Matthew inviting Craig out to the movies'. As Edgington (1995b) points out, we may not have much need to infer a conditional if we already know that the conjunction is true. But this does not mean that we are permitted to consider the conditional false, either. Indeed, Edgington argues that someone rejecting the conditional, '*if* Craig pays for the dinner, *then* Matthew will invite Craig out to the movies', would have to admit that they were wrong, if it

[95] Variants of this hypothesis have been discussed under different names such as 'Stalnaker Hypothesis', 'Adam's Thesis', and 'The Equation' in the literature (Oaksford and Chater, 2010; Douven, 2015).

turned out to be true that Craig pays for the dinner *and* Matthew invites Craig out to the movies. According to ST, the participants are predicted to conform to the following inequality in the so-called *uncertain and-to-if inference*, where they are presented with ‘A and C’ as a premise and ‘if A, then C’ as a conclusion and asked to assign probabilities to each:

$$P(\text{Conclusion}) \geq P(\text{Premise})$$

This prediction was directly tested by Cruz *et al.* (2015), who found that participants conformed to this inequality at above-chance levels.[96]

However, not all agree that P(if A, then C) = P(C|A) applies universally to all sentences with the syntactic form of a conditional. As pointed out by Edgington (1995a), one common objection is that the conditional probability hypothesis does not apply to conditionals containing sentences that are mutually irrelevant like ‘If Napoleon is dead, Oxford is in England’. These conditionals, which have come to be known as missing-link conditionals, represent an explanatory challenge for ST (Douven, 2017).

According to a rivaling approach known as *inferentialism*, the oddness of missing-link conditionals is interpreted as indicating that conditionals express *reason relations* or condensed arguments (Ryle, 1950; Rott, 1986; Strawson, 1986; Brandom, 1994; Read, 1995; Rescher, 2007; Spohn, 2013; Olsen, 2014; Douven, 2015; Krzyżanowska, 2015; Skovgaard-Olsen, 2016b). Proponents of inferentialism are also inclined to point out that inferences from and-to-if become a lot less plausible once missing-link conditionals are considered. Suppose we learn some irrelevant fact about Craig in the example above, which is unknown to Matthew. Say, Craig’s grandmother has a dog. And suppose further that it is still the case that Matthew invites Craig out to the movies. In that case, the conditional ‘If Craig’s grandmother has a dog, then Matthew will invite Craig out to the movies’ sounds bizarre to someone who tends to view the conditional as expressing a reason relation, although we know that the conjunction happens to be true. With the introduction of inferentialism to the psychology of reasoning, there is currently a considerable interest in and-to-if inferences. According to Over and Cruz (2018), these inferences represent "an important high-level dividing line between theories of conditionals". In Skovgaard-Olsen, Singmann, and Klauer (2016) a probabilistic

[96] In the Online Supplementary Materials, we discuss how prediction-performance levels from the different accounts can be compared to chance in the Bayesian mixture model used in our analyses. This chance correction is very similar to the one adopted by Cruz et al. (2015) and Evans et al. (2015).

implementation of inferentialism was given as a descriptive thesis, which employs the following explication of the reason relation, following Spohn (2012, Chap. 6):

A is *positively relevant* for C (and a reason *for* C) iff ΔP > 0
A is *negatively relevant* for C (and a reason *against* C) iff ΔP < 0
A is *irrelevant* for C iff ΔP = 0
For ΔP = P(C|A) - P(C|¬A)

The underlying intuition is that what we mean when we say that A is a reason *for* C is that A raises the probability of C. When we assume that A is the case, C becomes more likely as compared to when we assume that A is not the case. In the case of irrelevance, we can either assume A or ¬A, and the probability of C will stay the same, because A makes no difference for our degree of belief in C. The theory here follows Spohn's (1991, 2012) explication of the reason relation in terms of probability difference making, which treats causality as a special case of the generic reason relation. In Hahn and Oaksford (2007) similar ideas were applied to analysing informal arguments. Moreover, in the psychological literature on causation, ΔP > 0 has likewise been taken to be a necessary, but not sufficient, condition for judging causality. Or rather: the causal power, W_C, is a scaled version of ΔP (Cheng, 1997):

$$W_C = \frac{\Delta P}{1 - P(E|\neg C)} \quad \textit{for E = effect, C = cause}$$

Theories emphasizing causal interpretations of indicative conditionals, like Ali et al. (2010) and van Rooji & Schulz (2018), could be cast as special cases of an inferentialist approach to conditionals. The inferentialist approach is more general, however, because it applies equally well to diagnostic inferences from effects to causes, correlations in common cause scenarios, context-specific correlations in the absence of stable causal relations, and non-causal deductive inferences. Skovgaard-Olsen (2016a) moreover established a connection between the inferentialist view and Rescorla and Wagner's work on classical conditioning. Skovgaard-Olsen argued that one of the central functions of indicative conditionals is to culturally transmit information about contingency relationships, which would otherwise have to be tediously acquired by each subject on their own through associative learning.

The probabilistic implementation of inferentialism established by Skovgaard-Olsen et al. (2016) is a descriptive thesis named the *Default and Penalty Hypothesis* (DP). DP posits that participants have the goal of evaluating whether a sufficient reason relation obtains when evaluating P(if A, then C). According to the above explication of the reason relation, this requires at least two things: (a) assessing whether A is positively relevant for C, and (b) assessing the sufficiency of A as a reason for C by means of P(C|A). Moreover, DP postulates that participants make the default assumption that (a) is satisfied, which reduces their task of assessing P(if A, then C) to an assessment of P(C|A). However, when participants are negatively surprised by a violation of this default assumption, such as when they are presented with stimulus materials implementing the negative relevance ($\Delta P < 0$) or irrelevance category ($\Delta P = 0$), they apply a penalty to their estimate of P(if A, then C) as a way of reacting to the conditional's failure to express that A *is a reason for* C. An example would be the conditional 'If Oxford is in England, then Napoleon is dead', which sounds defective to the extent that the antecedent is obviously irrelevant for the consequent, as noted above.

Skovgaard-Olsen et al. (2016) reported empirical evidence in support of DP, showing that P(if A, then C) = P(C|A) only holds when A is positively relevant for C in virtue of raising its probability. When A is *negatively relevant* by lowering C's probability, and when A is *irrelevant* for C by leaving its probability unchanged, violations of the conditional probability hypothesis occurred. These findings were replicated by Skovgaard-Olsen et al. (2017b), who observed an average estimate of P(if A, then C) of .38, along with P(C|A) = 1. Moreover, Skovgaard-Olsen et al. (2017a) found that Cruz et al.'s (2015) finding of an above-chance conformity to the inequality P(Conclusion) ≥ P(Premise) in the uncertain and-to-if inference task only holds for positive relevance. In negative relevance and irrelevance conditions, participants actually perform at below-chance levels. For instance in the irrelevance condition it was found that participants conformed to the inequality in only 54% of the cases, a considerable drop from the 87% observed in the positive relevance condition. Importantly, this drop in conformity to the and-to-if inference across relevance levels was not reflected in participants' conformity to the inequality P(C|A) ≥ P(A,C): 77% and 76% in the positive relevance and irrelevance conditions, respectively. It is not clear how the dissociation between the effect of relevance on the P(Conclusion) ≥ P(Premise) and P(C|A) ≥ P(A,C) can be reconciled under ST's assumption that P(if A, then C) = P(C|A).

Given the theoretical status of the inequality P(Conclusion) ≥ P(Premise), it is critical that we understand the nature of the lack conformity to it under certain relevance conditions. One possibility is that individuals are adhering to ST but just so happen to be committing reasoning errors. Alternatively, it is possible that individuals are in fact adhering to an alternative interpretation of conditionals like DP, under which their responses are not only justified but expected. Unfortunately, this interpretational ambiguity cannot be resolved with the currently available studies, as they only enable an evaluation at the aggregate-group level. Ultimately, we want to be able to establish individual profiles that characterize each participant's reflective attitudes, and use them to evaluate the correctness of their judgments. In order to achieve this goal, we developed a novel experimental paradigm, the Scorekeeping Task, along with a Bayesian mixture model that was tailored to characterize the data coming from it.[97]

Experiments

The scorekeeping task is implemented in three different studies, and used to establish individual profiles according to their classification as followers of the Suppositional Theory (ST) or the Default and Penalty Hypothesis (DP). These profiles were then used to investigate whether participants are committing reasoning errors, relative to their own interpretation of the conditional. In Experiments 1 and 2, we focused on the uncertain and-to-if inference task, whereas Experiment 3 focused on the acceptance of entailment relations. Additionally, we tested whether individuals classified as adhering to ST and DP differed with respect to their interpretation of probabilities (Experiment 1), production of conjunction fallacies (Experiment 1), or argumentative skills (Experiment 2).

7.2 Experiments 1 and 2: Probability and Scorekeeping

The goal of the first two experiments is to use the participants' responses in the Scorekeeping Task in order to establish individual profiles of the participants based on whether they can be classified as following the Suppositional theory (ST) or the Default and Penalty Hypothesis (DP). But due to their similarity, both experiments are reported together. However, it should be highlighted that one of the main motivations of Experiment 2 was to replicate some of the results from Experiment 1. The key differences between the two experiments concern the use of novel scenarios in Phase 4 (instead of the same scenarios from Phase 1), and the type of individual judgments being evaluated in Phase 4 (Experiment 1: conjunction fallacy and interpretation of probabilities; Experiment 2: argumentation skills). Given the similarity of the main results obtained with Experiment 2, we will only present the figures for results from Experiment 1 (for the results of phase 4 of both experiments, and the results of Experiment 2, see Online Supplementary Materials).

7.2.1 Method

Participants

[97] For a detailed discussion of how the Bayesian mixture model differs from previous regression-based approaches, see Online Supplementary Materials.

Participants from the USA, UK, Canada, and Australia took part in these experiments, which were launched over the Internet (via Mechanical Turk) to obtain a large and demographically diverse sample. 354 persons took part in the first Experiment, 552 in the second. Participants were paid a small amount of money for their participation. The following exclusion criteria were used: not having English as native language, completing the experiment in less than 300 seconds, failing to answer two simple SAT comprehension questions correctly in a warm-up phase, and answering ‘not serious at all’ to the question how serious they would take their participation at the beginning of the study. The final samples consisted of 261 and 340 participants, respectively. In Experiment 1, the mean age was 36.53 years, ranging from 20 to 75, 66% were female, 66% indicated that the highest level of education that they had completed was an undergraduate degree or higher. The demographic measures of the participants differed only minimally before and after the exclusion. The demographic variables in Experiment 2 were very similar.

Design

The experiments implemented a within-subject design with two factors varied within participants: relevance (with two levels: positive relevance, irrelevance) and priors (with four levels: HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH).

Materials and Procedure

We used a slightly modified version of 12 of the different scenarios presented in Skovgaard-Olsen et al. (2016) (see Suppl. Materials). They have been pretested to manipulate the reason relations defined above. This allows us to vary the presence and absence of specific reason relation orthogonally to other psychological factors of interest. To illustrate, Table 1 displays target positive relevance and irrelevance conditionals for the Scott scenario:

Table 1. Stimulus Materials, Scott Scenario

Scenario	Scott was just out playing with his friends in the snow. He has now gone inside but is still freezing and takes a bath. As both he and his clothes are very dirty, he is likely to make a mess in the process, which he knows his mother dislikes	
	Positive Relevance	**Irrelevance**
HH	If Scott turns on the warm water, then he will be warm soon	If Scott’s friends are roughly the same age as him, then Scott will turn on the warm water.
HL	If Scott makes an effort to be tidy, then the bathroom will be just as clean as before he took his bath.	If Scott’s friends are roughly the same age as him, then Scott will turn on the cold water.
LH	If Scott bathes in a hot spring, then he will be warm soon.	If Scott’s friends are 10 years older than him, then Scott will turn on the hot water.
LL	If Scott turns on the cold water, then he will soon start to freeze even more.	If Scott’s friends are 10 years older than him, then Scott will turn on the cold water.
Positive Relevance (PO):	mean ΔP = .32	High antecedent: mean P(A) = .70
Irrelevance (IR):	mean ΔP = -.01	Low antecedent: mean P(A) = .15

	High consequent:	mean P(C) = .77
	Low consequent:	mean P(C) = .27

Note. HL: P(A) = High, P(C) = low; LH: P(A) = low, P(C) = high. The bottom rows display the mean values for all 12 scenarios pretested with 725 participants in Skovgaard-Olsen et al. (2017a).

For each scenario we had 8 conditions according to our design (i.e., 4 conditions for positive relevance [i.e., HH, HL, LH, LL], 4 conditions for irrelevance). Each participant worked on one randomly selected (without replacement) scenario for each of the 8 within-subjects conditions such that each participant saw a different scenario for each condition.

Experiments were split into four phases. The precise formulation of all the questions and instructions can be found in the Supplementary Materials. Here we focus on conveying the conceptual ideas.

7.2.1.1 Phase 1: Case Judgments

The first phase contained eight blocks, one for each within-subjects condition. The order of the blocks was randomized anew for each participant and there were no breaks between the blocks. Within each block, the participants were presented with four pages. On the first page, the participants were shown a scenario text like the above Scott scenario.
To introduce the eight within-subjects conditions for the scenario above we, *inter alia*, exploited the fact that the participants assume that Scott's turning on the warm water raises the probability of Scott being warm soon. In the terms introduced above, Scott's turning on the warm water is in other words positively relevant for (or a reason *for*) believing that Scott will be warm soon (positive relevance). In contrast, Scott's friends being roughly the same age as Scott is irrelevant for whether Scott will turn on the warm water (irrelevance). The first sentence in other words leaves the probability of the second sentence unchanged, as verified in Skovgaard-Olsen et al. (2017a). In this study, we use such irrelevance items to present the participants with missing-link conditionals.

The scenario text was repeated on each of the following three pages which measured P(A and C), P(C|A), and P(if A, then C) in random order. Throughout the experiment, participants gave their probability assignments using sliders with values between 0 and 100%. To measure P(C|A), the participants might thus be presented with the following question in an irrelevance condition:

> Suppose Scott's friends are roughly the same age as Scott.
> Under this assumption, how probable is it that the following sentence is true on a scale from 0 to 100%:
> Scott will turn on the warm water.

7.2.1.2 Phase 2: The Scorekeeping Task

In this phase the participants were first presented with a new irrelevance item to be rated in the same way as the items in phase 1. The missing-link conditional took the following form and it was evaluated in the context of a dating scenario describing Stephen's preparations for a date with Sara: 'If Stephen's neighbour prefers to put milk on his cornflakes, then Stephen will wear some of his best clothes on the date'. Then the participants were presented with the following instruction:

> When given the task you just completed, John and Robert responded very differently to some of the scenarios as outlined below.

And it was explained that John and Robert responded in the following way to the "if-then sentence" and the "suppose-sentence" (where the "suppose-sentence" had been identified for the participants as the type of question quoted above for measuring P(C|A)):

> John assigned **99%** to the suppose-sentence and **1%** to the if_then sentence**.**
> Robert assigned **90%** to the suppose-sentence and **90%** to the if_then sentence.

In order to reduce the processing demands of this task, these values were repeated on each of the following four pages along with the irrelevance item. Note that although John and Robert are fictive participants, these values were based on actual data provided by other participants in response to the irrelevance item in previous studies.

As part of the Scorekeeping Task, the participants were instructed to apply a sanction to John or Robert's response based on its adequacy. Given their large divergence, the participants were instructed that at most one of John or Robert's responses could be approved as adequate. Since the experiment was run on Mechanical Turk we exploited the fact that an ecologically valid sanction for the participants would be not to have a task (called a "HIT") approved. Because the approval of HITs on Mechanical Turk determines whether the participants are paid for a completed task (and moreover counts towards their reputation, which determines whether they can participate in future HITs), it is our experience that the participants on Mechanical Turk care a lot about the approval of their HITs. We therefore expected that applying the sanction of not approving either John or Robert's HIT based on its adequacy would be a contextually salient sanction, which the participants would be highly motivated to reason about.

Next, the participants were asked to state the reasons that they could think of which could be given for or against John and Robert’s responses in an open entry question, included for exploratory purposes.

On the two pages that followed, the participants were presented with John’s criticism of Robert and Robert’s criticism of John in random order. Robert made the following complaint about John’s response:

> **Robert's no difference justification**: “There is no difference between the two questions. So why do you give a lower probability to:
> *'IF Stephen’s neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date’*
> than you gave to:
> *'Stephen will wear some of his best clothes on the date'* under the assumption that *'Stephen’s neighbour prefers to put milk on his cornflakes'?*
> This makes no sense!”

John in turn made the following complaint about Robert’s response:

> **John's irrelevance justification**: “Whether *'Stephen’s neighbour prefers to put milk on his cornflakes'* or not is irrelevant for whether *'Stephen will wear some of his best clothes on the date'*. So why do you give such a high probability to: *‘IF Stephen’s neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date'?*
> This makes no sense!”

In each case, the participants were asked to indicate using a binary ‘yes/no’ answer whether they agreed with the statements:

> - John’s irrelevance justification [/Robert’s no difference justification] shows that Robert's [/John’s] response is wrong.
> - Robert [/John] needs to come up with a very good response to John's [/Robert’s] criticism, if his HIT is to be approved.

Finally, after having seen the justifications from both sides, the participants were asked which justification they found most convincing by choosing between the following options presented in random order:

The two justifications are equally convincing
John’s irrelevance justification
Robert’s no difference justification

Moreover, the participants were asked to indicate whose HIT deserves to be approved based on their justifications by selecting one of the following options presented in random order:

None of their HITs should be approved
Robert’s HIT should be approved
John’s HIT should be approved

7.2.1.3 Phase 3: The Uncertain And-to-If Inference

This phase served the purpose of testing the participants’ performance on the uncertain and-to-if inference task under relevance manipulations. Phase 3 was used to measure whether participants' responses to the uncertain and-to-if inference task were consistent with the interpretation of the conditional they had been classified according to.

Phase 3 contained 8 blocks implementing the same within-subjects conditions as phase 1. In Experiment 1, for each participant, the same permutations of scenarios and within-subject conditions that had been randomly generated in phase 1 were displayed again in random order. In Experiment 2, new scenarios were used. First the participants were instructed that they would be presented with short arguments based on the scenario texts. They were told that the premise and the conclusion of the arguments could be uncertain and that it was their task to evaluate their probabilities. On the top of the page the scenario text was placed as a reminder. Below the participants were instructed to read an argument containing the conjunction as a premise and the conditional as a conclusion, employing sentences that they assigned probabilities to in phase 1. Furthermore, the actual value of the probability that they had assigned to the premise in phase 1 was displayed to the participants in a salient blue color. We here illustrate it using the example above from phase 1 of a positive relevance item:

Premise: Scott turns on the warm water AND Scott will be warm soon.
Conclusion: IF Scott’s turns on the warm water, THEN Scott will be warm soon.
You have estimated the probability of the premise as: **90%**
Please rate the probability of the statement in the conclusion on a scale from 0 to 100%.

7.2.1.4 Phase 4: Individual Variation

In the Online Supplementary Materials, further investigations are reported into covariates that would characterize participants classified as interpreting the conditional according to ST and DP such as differences in their argumentative skills (evaluated by an adaption of Kuhn's (1991) task), their interpretation of probabilities, and tendency to commit the conjunction fallacy. The goal of these investigations was to consider the hypotheses that 1) what characterizes DP participants is merely a defective understanding of probabilities, and 2) participants in the DP group pay more attention to reason relations because they possess stronger argumentative skills than ST participants. The first of these is introduced as an alternative hypothesis in Skovgaard-Olsen et al. (2017a), and it echoes results by Tentori, Crupi, and Russo (2013), who found that the participants committing the conjunction fallacy are misled by the degree of confirmation of the added conjunct. However, neither hypothesis could be supported by our results; it therefore appears that the differences we tap into when investigating the opposition between ST and DP are orthogonal to differences in these further variables.

7.2.2 Results and Discussion

Bayesian Mixture Model

In order to investigate the participants' interpretation of the conditional, the probability judgments produced in phase 1 were classified as coming from one of two latent classes using a Bayesian Mixture Model (see Online Supplementary Materials). When individuals follow ST, the generated P(if A, then C) are expected to follow P(C|A) in both the positive relevance and irrelevance conditions. In contrast, when individuals follow DP, the generated P(if A, then C) are expected to follow P(C|A) in the positive relevance condition, and a penalized version of P(C|A) in the irrelevance condition (each participant *i* has a penalty parameter, θ):

$$P(if\ A, then\ C)_{i,j} = \begin{cases} P(C|\ A)_{i,j} +\ \varepsilon_{i,j}, & w_i^{IR} = 0, \\ \theta_i P(C|\ A)_{i,j} +\ \varepsilon_{i,j}, & w_i^{IR} = 1, \end{cases}$$

Figure 1 displays the predictions of these two models for the irrelevance condition. Note that when $\theta = 1$, the ST and DP models coincide, although the implied predictions are not really in accordance with the gist of DP. However, this point turns out not to be of practical import, because since ST is more parsimonious it will be preferred when $\theta = 1$ (see M. D. Lee, 2016).

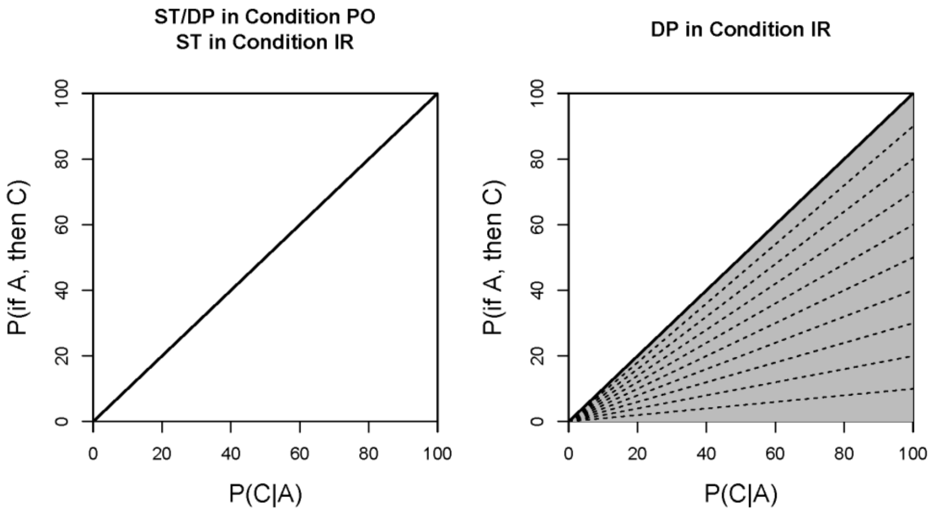


Figure 1. Predictions. The Suppositional Theory (ST) equates P(if A then C) and P(C|A). The Default-Penalty Hypothesis (DP) makes the same prediction only for positive relevance (PO). For irrelevance (IR), it expects a function that lies below the diagonal. Here we assume for our classificatory purposes that the DP predictions in IR correspond to a linear function with a slope between 0 and 1.

In the positive relevance condition, where ST and DP coincide, classifications were made using two classes: One that expects the elicited P(if A, then C) to be equivalent to the elicited P(C|A), as expected by both ST and DP, and a second "saturated" class which establishes one parameter per data point:

$$P(if\ A, then\ C)_{i,j} = \begin{cases} P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 0, 1, \\ \beta_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 2, \end{cases}$$

This second class is used here to exclude individuals whose responses are not in line with either ST or DP. This exclusion constitutes an important step here as we first need to ensure that both models at the very least are able to provide a good account of the data in which they agree, and thus to avoid potential distortions that could be introduced by including data that is at odds with both theoretical accounts (Hilbig & Moshagen, 2014). This focus on a subset of the data establishes an "optimistic testbed" for the two different theoretical accounts in the sense that the testing of predictions is limited to data that both theories can successfully describe.

Phase 1

The individual-level classifications shown in Figure 2 show that the probabilities generated by the majority of individuals in the positive relevance condition were in line with ST/DP (211 out of 261). In contrast, it could be seen based on the irrelevance condition that only a very small group of individuals were in line with ST (52 out of 225), as the vast majority of them followed the predictions of DP (159). The individual data from Experiment 1 shown in the left and central panels of Figure 2 show that the data classified as ST/DP in the positive relevance condition (upper panels) as well as ST and DP in the irrelevance condition (bottom panels) were in line with the model predictions. These results were corroborated by the classification probabilities, as most classifications were far from the cut-off .50 value. There were relatively few classifications that were close to .50 (see Figure 2). Additional support comes from the θ_i estimates obtained when individuals were classified as following DP. In both experiments, these values were far from the upper boundary of 1, where no penalty is imposed and DP converges to ST (Mean = 0.31, SD = 0.24), indicating that the small number of ST adherents is not due to any sort of mimicry from DP.

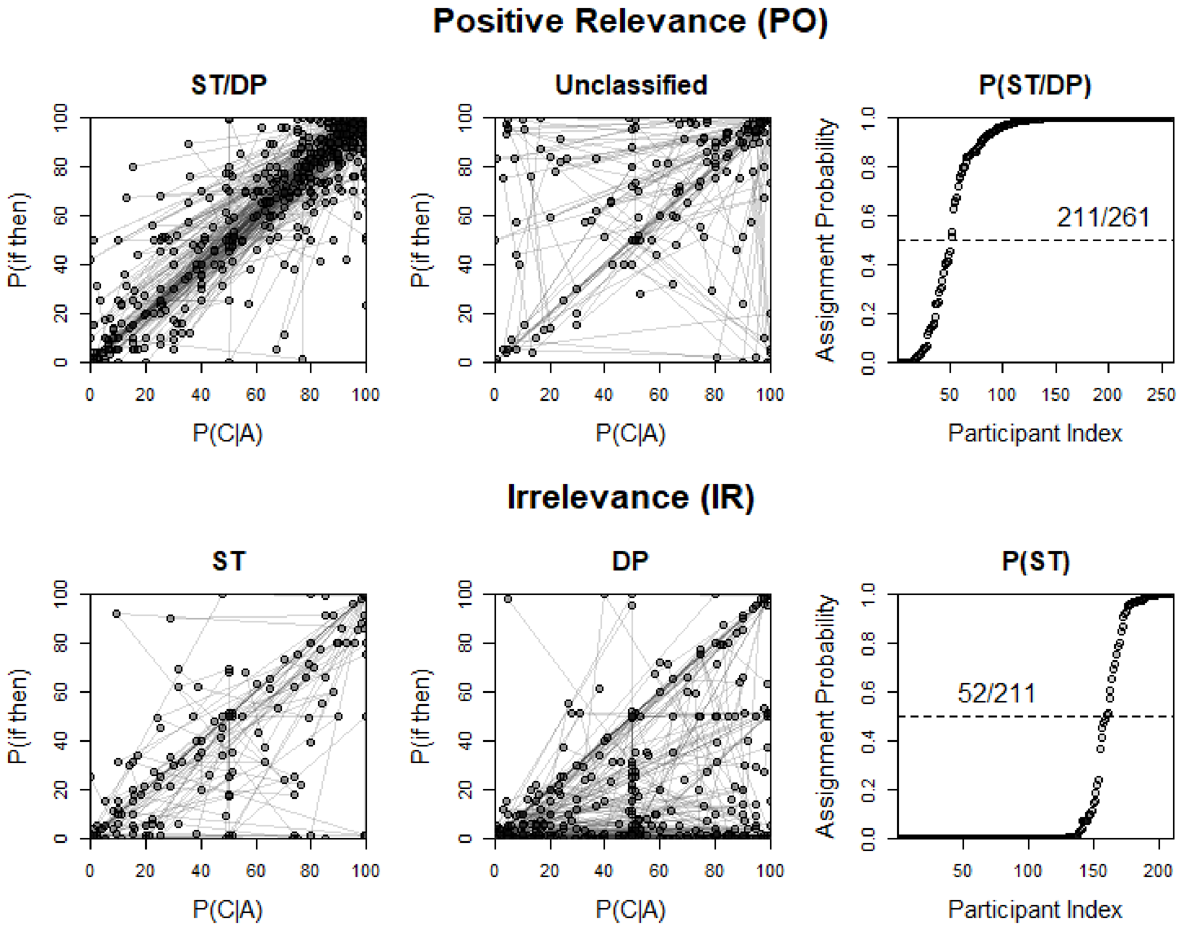


Figure 2. Left and Center Panels: Individual data associated to the phase 1 classifications in Experiment 1. Right Panels: Individuals' posterior classifications (note that in the irrelevance condition, only participants classified as ST/DP in the positive relevance condition were considered).

Phase 2

Next, we classified the participants based on the reflective attitudes the participants' manifested through their behavior on the scorekeeping task. This task was used to commit the participants to an interpretation of the conditional, depending on whether they agreed to criticize John or Robert and sanction them through HIT assignments. If the participants were following the instrumental goal of engaging in suppositional reasoning when assessing the conditional, then they should treat the conditional as expressing a conditional probability and agree with Robert. If the participants were following the instrumental goal of assessing whether a sufficient reason relation obtained, then the irrelevance condition should make the conditionals appear defective and they should agree with John.

In this classification we considered 1) their support for one of the fictive characters and 2) their HIT attribution. Individuals were classified as DP/ST when they judged the fictive character of DP/ST to be most convincing *and* attributed him the HIT.

Table 2. Phase 1 and Phase 2 comparison (Experiment 1)

Phase 1	**ST$_1$ (N = 52)**	**DP$_1$ (N = 159)**	**Unclassified (N = 50)**
Accept Criticism	.67 [,.53, .81]	.85 [.79, .91]	.59 [.43, .74] / .50 [.34, .66]
Assign Burden of Proof	.80 [.72, .86]	.71 [.57, .84]	.49 [.34, .66] / .55 [.39, .70]
Most Convincing*	.96 [.86, 1]	.97 [.92, 1]	.35 [.14, .60]
Approve Hit*	.92 [.80, .99]	.92 [.86, .97]	.62 [.44, .78]
Phase 1 / Phase 2	**ST$_2$ (N = 46)**	**DP$_2$ (N = 132)**	**Unclassified (N = 83)**
ST$_1$ (N = 52)	32	0	20
DP$_1$ (N = 159)	1	125	33
Unclassified (N = 50)	13	7	30

Note. The top rows show the posterior probabilities of ST$_1$ and DP$_1$ participants, following their assigned interpretation for each phase 2 question. In the column 'Unclassified', we report two estimates, corresponding to the subjects that would have been classified as ST/DP in the irrelevance condition (left/right). Rows 'Most Convincing' and 'Approve HIT' indicate the posterior probability that consistent preference was expressed; conditional on the presence of a preference (e.g., subject did not express indifference). The phase 2 classification in the bottom row is based on the participants' responses to who had the most convincing justification, and whose HIT should be approved, after having seen the justification from both sides.

As shown in Table 2, the match between the phase 1 and 2 classifications is large and systematically above .50. Unclassified participants distributed their responses roughly equally across Robert and John. Although the overlap between phase-1 and phase-2 classifications was considerable (157 participants out of 211), it was not perfect. This was mainly due to the circumstance that there was a substantial proportion of the participants (73), who found the two fictive characters equally convincing and a few participants (21), who chose to assign a HIT to neither. But for those who did, their judgments were closely aligned with their phase 1 classification.

Phase 3

We now turn to the participants' conformity to the two inequalities associated with uncertain and-to-if inferences:

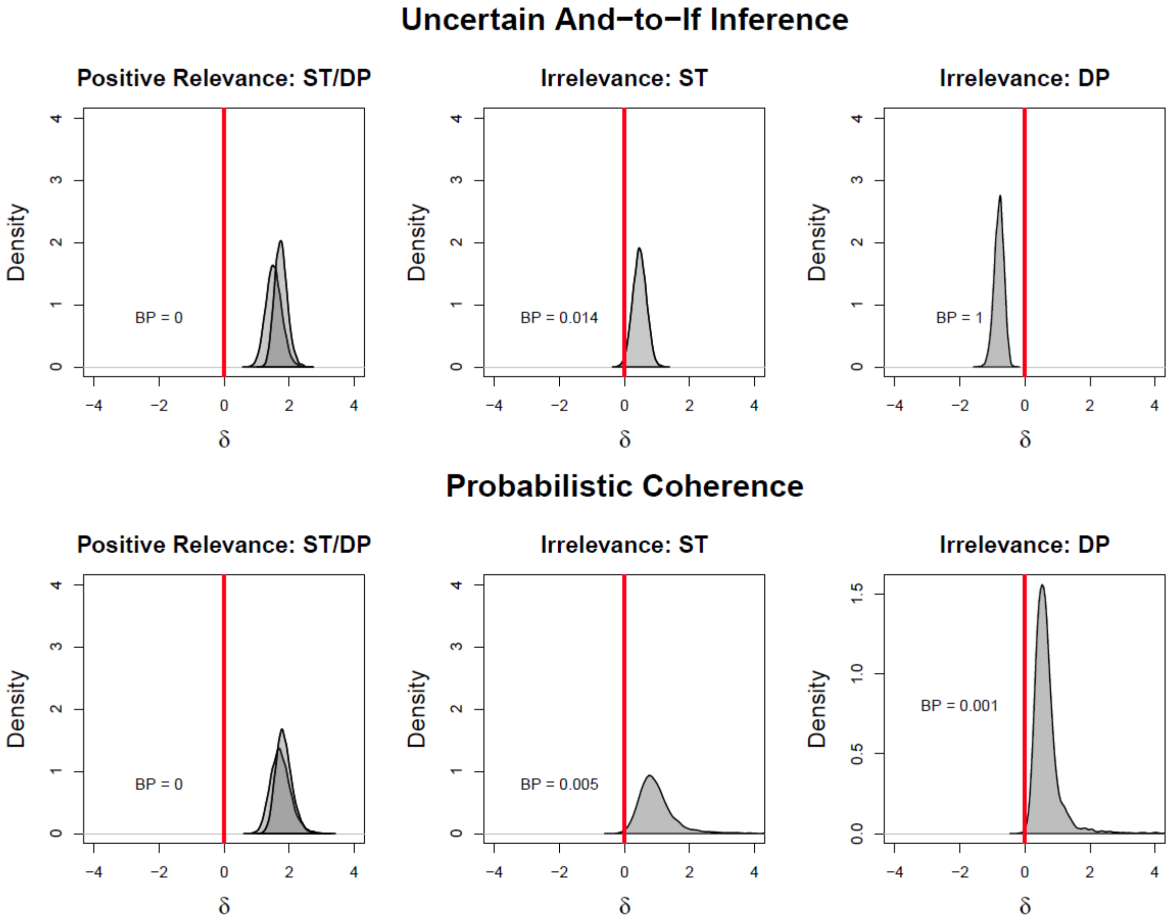


Figure 3. Posterior distributions of the deviations of the tested inequalities from chance-level occurrence (represented on an effect-size scale) in Experiment 1. The vertical lines indicate effect size 0 and BP corresponds to the probability of samples from the posterior distributions taking on values below 0. In the left panels we depict the posterior distributions for participants classified as ST and DP (the latter corresponding to the more peaked distributions) in the positive relevance condition.

$$P(\text{Conclusion}) \geq P(\text{Premise})$$

$$P(C|A) \geq P(A,C)$$

Figure 3 depicts the posterior distributions of these deviations from chance on an effect-size scale, with positive values indicating an above-chance conformity to the inequalities (for details, see Online Supplementary Materials): In the positive relevance condition (left panel), the participants conformed to both inequalities at above-chance levels. This result is represented by the posterior distributions placed with virtually all of their mass above zero (i.e., BP ≈ 0). This pattern of results held for both individuals classified as adhering to ST and DP. However, the posterior distributions for ST are more dispersed due to the small number of participants classified as such. Differences were found in the irrelevance condition, since individuals classified as following ST conformed to both inequalities at above-chance rates, whereas individuals classified as following DP followed P(Conclusion) ≥ P(Premise) at *below*-chance rates. This difference is germane given that P(Conclusion) ≥ P(Premise) is not expected to hold under DP when there is no positive reason relation between the antecedent and the consequent. Note that P(C|A) ≥ P(A,C) is expected to hold across accounts and relevance conditions; this prediction also held empirically.

7.3. Experiment 3: Entailment Judgments

So far, we have been concerned with interpretations of the conditionals that the participants commit to when making probabilistic assessments. This evaluation can be extended to other types of judgments, such as the *acceptance of entailments*. A central empirical adequacy criterion of semantic theories in general is that they respect intuitive entailment judgments (Winter, 2016). Indeed, such judgments make up one of the primary sources of data for semantic theories. The goal of Experiment 3 is to investigate how robust and stable the participants' interpretations of conditionals under different task constraints are.

As previously discussed, individuals following ST are expected to infer a conditional ‘If A, then C’ when using the conjunction ‘A and C’ as a premise. In other words, they are expected to produce *and-to-if inferences.* No such expectation holds for individuals reasoning according to DP, at least in the absence of a reason relation between A and C. In the context of Experiments 1 and 2, we showed that individuals’ classification in the Scorekeeping Task as ST or DP was consistent with whether or not they conformed to the inequality P(if A, then C) ≥ P(A and C) in the uncertain-and-to-if task. This differential conformity has implications for the acceptance of entailments. For instance, it would be inconsistent for reasoners adhering to DP to violate the inequality P(if A, then C) ≥ P(A and C) in the uncertain-and-to-if task while accepting that the conditional ‘if A, then C’ is entailed by the premise ‘A and C’. This consistency requirement follows from general constraints that ensure that probabilistic reasoning is consistent with deductive logic (Joyce, 2004; Oaksford, 2014):

$$A \vDash B \qquad \text{only if} \qquad P(B) \geq P(A)$$

Hence,

$$\text{A and C} \vDash \text{if A, then C} \qquad \text{only if} \qquad P(\text{if A, then C}) \geq P(\text{A and C})$$

In order to evaluate conformity to this consistency requirement, Experiment 3 is comprised of two sessions: The first session is essentially a replication of Experiment 1 that allows us to classify individuals as adhering to ST or DP with the Scorekeeping Task.

In the second session, individuals were presented with different scenarios in which two speakers disagreed on whether a certain conclusion followed from a given premise. We considered three types of inferences under positive relevance and irrelevance conditions:

First, the aforementioned and-to-if inference, that one is expected to follow depending on the interpretation of the conditional adhered to:

$$\text{A and C} \vDash \text{if A, then C.}$$

Specifically, we expect individuals conforming to ST to accept that ‘if A, then C’ is entailed by ‘A and C’, whereas no such acceptance is expected for individuals adhering to DP across relevance conditions. We also considered two other inferences, namely and-to-A inferences, which are uncontroversially valid,

$$\text{A and C} \vDash \text{A,}$$

and A-to-and inferences, which are uncontroversially invalid[98]

$$\text{A} \vDash \text{A and C.}$$

7.3.1 Method

Participants

Experiment 3 was run over Mechanical Turk and used the same exclusion criteria as Experiment 1. A total of 811 people participated in the first session 1. Of these a total of 610 participated in session 2, which was run approximately 10 days later. In addition to the exclusion criteria from Experiment 1, we checked their identity in session 2 by requiring them to provide once again some personal information (e.g., first letter of your favourite colour, first letter of mother’s name) to generate codes like 'AS6G1P', which preserved the anonymity of the participants. In the end, we were left with a final sample of 552 participants, with similar demographic characteristics as in Experiment 1 and 2. Of these, 515 could be classified as following either DP or ST in the Scorekeeping Task. In the analysis below, we focus on these 515 participants (330 DP; 186 ST).

Design

The first session of Experiment 3 had the same design as Experiment 1, with additional questions for prior probabilities. However, in contrast with Experiment 1, the participants were now presented with the Scorekeeping Task as a two-alternative forced-choice task, where they either had to take sides with one of the two fictive characters (i.e., they cannot deem them equally convincing). The second session presented the same eight

[98] We refer to the validity status of these two inferences as uncontroversial given that we do not know of any logical system in which they are assigned a different status.

within-subject conditions as Experiment 1. In addition to the entailment judgments, we also collected the participants' self-reported consistency in session 2 with their judgments in session 1.

Materials and Procedure

For the entailment judgments in session 2, the participants were given the following instructions:

> In the following you are going to see a short conversation, where Louis accuses Samuel of saying two things that cannot both be true. Whether you agree with Samuel's assertions is beside the point. What we are interested in is just the extent to which you agree with Louis that Samuel is saying two things that cannot both be true. When you read the sentences please pay attention to small differences in their content, so that we don't unfairly accuse Samuel of making a mistake.

After a few practice items, the participants were presented with the same randomly selected scenarios as in Experiment 1, and on the three pages that followed, Samuel would assert the premise of each of the three types of inferences described above and deny its conclusion. Consider the following example, using the Scott scenario in Table 1 and one of the irrelevance items:

> **Samuel:**
>
> Scott's friends are roughly the same age as him AND Scott will turn on the warm water.
>
> ... but it would be wrong to think that IF Scott's friends are roughly the same age as him, THEN Scott will turn on the warm water.

To which his interlocutor replied:

> **Louis:** Wait, you've now said two things that can't both be true.

The task of the participants was to indicate the degree to which they agreed or disagreed with Louis' statement on a five-point Likert scale with levels *strongly disagree, disagree, neutral, agree,* and *strongly agree*. Agreeing with Louis in that Samuel had said two things that cannot both be true counts as accepting the corresponding entailment.

7.3.2 Results

Entailment Judgments

The design had replicates for each participant and item. It could therefore not be assumed that the data were independently and identically distributed. Consequently, linear mixed-effects models were used, with crossed random effects for intercepts and slopes by participants and by scenarios (Baayen, Davidson, and Bates, 2008). This analysis was conducted using the statistical programming language R (R Core Team, 2013), and the package brms for mixed-effects models in Bayesian statistics (Bürkner, 2017). In order to examine the rating of entailments for the three types of inferences, we relied on the following models:

> Model **M1** modelled the rating as a function of factor 'inference' (coding the three different types of inferences), factor 'relevance', the factor 'individual classification' (as ST, DP, based on the Scorekeeping Task), and their interactions.
>
> Model **M2** builds upon M1 but without the 'individual classification' factor and its interactions.
>
> Finally, model **M3** builds upon M2 but without the 'relevance' factor and its interactions.

In line with the previous studies, these models were implemented in a Bayesian framework with weakly informative priors, using the R package brms (Bürkner, 2017). Since the responses obtained from the five-point Likert scale are ordinal responses, the responses were modelled as generated by thresholds set on a latent continuous scale with a cumulative likelihood function and a logit link function (Bürkner & Vuorre, 2018). The upper part of Table 5 reports the performance of the models as quantified by the leave-one-out cross validation criterion and WAIC.

Table 5. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1	30307.93	10.04	2.15	30276.2	0.006
M2	30302.11	4.22	0.89	30270.3	0.108
M3	30297.89	0	--	30266.6	0.886
M4	4968.35	4.52	5.22	4964.8	0.095
M5	4963.84	0	--	4960.5	0.905
M6	5118.24	154.41	28.30	5113.7	0.000

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. Weight = Akaike weight of LOO.

As the information criteria indicate, M3 was the winning model within this first cluster of models. This indicates that overall, the entailments the participants accept do not appear to be based on the relevance condition of the items, nor on which interpretation of the conditional the participants committed to in session 1. We thus find Bayes factors in the range of [19, 51] in favour of H_0 when setting coefficients involving the relevance factor in M1 equal to 0. For instance, $b_{\text{PositiveRelevance:ANDIF:ST}}$ = 0.12, 95%-CI [-0.33, 0.57], BF_{H0H1} = 19.47 and $b_{\text{PositiveRelevance}}$ = -0.04, 95%-CI [-0.22, 0.14], BF_{H0H1} = 50.64. Furthermore, we find Bayes factors in the range of [6, 31] in favour of H_0 when setting coefficients involving the individual classification factor in M1 equal to 0.

Examining the posterior predictive distribution of the winning model M3 illustrated in Figure 4, it is clear that most of the participants accept the valid and-to-A inferences, and that most reject the and-to-if inferences to the similar degree to which they reject the invalid A-to-and inferences.

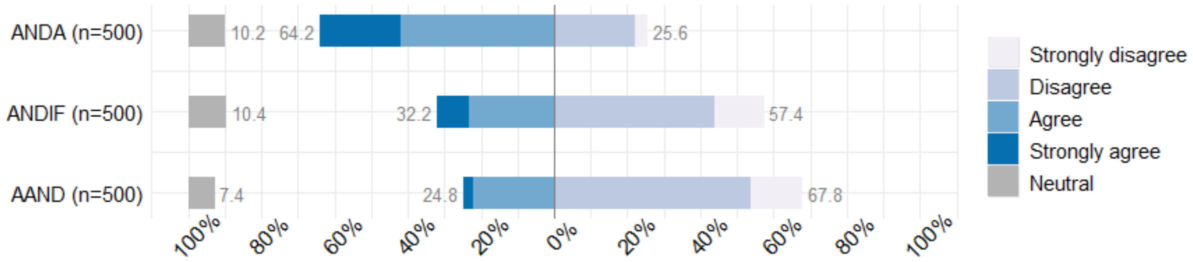


Figure 4. Predictions for Sampling from the Posterior Distribution of M3. Note: The plot shows the relative proportions of the posterior predictions of the winning model (M3). ANDA = and-to-A inference, ANDIF = and-to-if inference, AAND = A-to-and inference.

And-to-if inference

Given that the phase 2 classification does not predict the participants' acceptance of entailments, we turned our focus to the participants' acceptance of and-to-if inferences (i.e., ratings larger than 3) and investigate whether it can be predicted by their acceptance of the invalid A-to-and inference and the valid and-to-A inference. Finally, we also considered the degree to which the participants view themselves in session 2 as being consistent with their judgments in session 1, ca. 10 days earlier. For the participants' own perceived consistency, a factor was formed based on the quantiles low (≤ 40%), middle (41-61%), high (≥ 62%):

Model **M4** described the probability of accepting the and-to-if entailment as a function of the acceptance of the and-to-A inference, the A-to-and inference, the participant's self-reported degree of consistency, and their respective interactions.

M5 builds on M4 but does not include the acceptance of the and-to-A inference factor.

M6 builds on M5 but does not include acceptance of the A-to-and inference factor.

Since acceptance of an entailment is a binary variable, a binominal likelihood function was used with a logit link function and weakly informative priors, using the R package `brms` (Bürkner, 2017). The results shown in the lower part of Table 5 indicate that there is a strong effect of the acceptance of (the invalid) A-to-and inferences on the probability of accepting and-to-if inferences. Figure 5 reports the expectations of the posterior predictions of models M4-M6 weighted by their Akaike weights from Table 5 for a new participant.

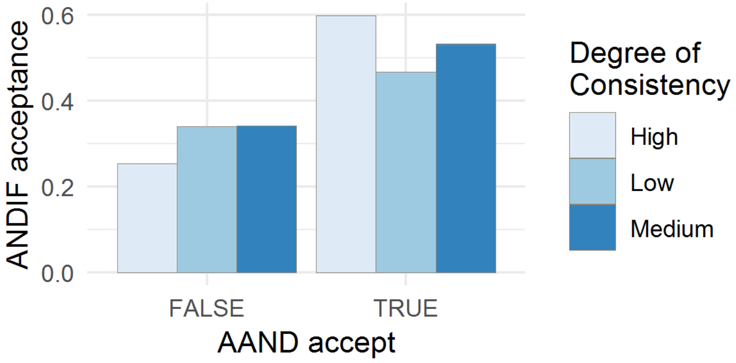


Figure 5. Posterior Predictions for New Participants. Note. The posterior predictions for acceptance of the and-to-if inference (ANDIF) for new participants based on their acceptance of the invalid a-to-and inferences (AAND) and low/middle/high quantiles of perceived consistency across sessions 1 and 2. The posterior predictions of the models have been weighted by the Akaike weight from Table 5.

The effect indicates that the participants are more likely to accept the and-to-if inference if they incorrectly accepted the A-to-and inference ($b_{\text{AAND\_accept}}$ = -0.57, 95%-CI [-0.658, -0.485], BF_{H0H1} = -2.75 * 10^{-26} ≈ 0). Transforming from the logit scale, this gives an increase of 36% chance of accepting the and-if-inference based on accepting the invalid A-to-and inference. In contrast, there is only a weak effect for the acceptance of the and-to-if inference based on acceptance of the valid and-to-A inference ($b_{\text{ANDA\_accept}}$ = 0.09, 95%-CI [-0.001, 0.184], BF_{H0H1} = 17.17), which makes M5 the second most preferred model.

7.3.3 Discussion

Overall, the results show that participants' endorsed interpretation of the conditional in the Scorekeeping Task and their own judgments of internal consistency across the two sessions were poor predictors of accepted entailments. In general, the participants accepted an uncontroversial example of a valid inference rule (A and C ⊨ A?), and rejected an uncontroversial example of an invalid inference rule (A ⊨ A and C?), across relevance conditions. It was found that the participants’ performance with and-to-if inferences (A and C ⊨ if A, then C?) resembled their performance for the invalid A-to-and inferences more than for the valid and-to-A inferences. Moreover, the results indicated that the participants' acceptance of and-to-if entailments was most strongly predicted by their acceptance of the invalid A-to-and entailment.

Applying the Principle of Tolerance amounts to empirically investigating how far we can succeed in reconstructing internally consistent competence models of the participants. Accordingly, the participants classified as adopting ST in session 1 of Experiment 3 were expected to accept the and-to-if entailment in session 2, and the participants conforming to DP in session 1 were expected to reject it across relevance conditions. Instead, what we found was that both groups tended to reject and-to-if entailments to the same degree as they rejected the invalid A-to-and entailments. For the participants following DP, this response pattern is still consistent with their assigned competence model. But for the participants following ST, rejecting the and-to-if entailment looks like an error, and the fact that the acceptance of the and-to-if entailment is best predicted by acceptance of the invalid A-to-and entailment leaves little room for reconstructing the participants’ performance as rational. The problem is that we cannot conceive of a competence model under which the acceptance of A-to-and entailments can be considered as anything but a reasoning error.

7.3.4 Summary of Case Study

The literature on formal systems of reasoning has branched out into a series of competing frameworks. Insofar as psychology seeks to model realistic reasoning performance, psychological investigations need to come to terms with the fact that there is often more than one competence model that could plausibly be applied to the participants’ performance.

In this paper, we put forward a normative and experimental framework for studying reasoning performance in a multiple-norms environment. We applied the Principle of Charity when obtaining independent evidence of the participants’ parameter settings before evaluating their reasoning performance. Using Bayesian mixture modeling we classifed the participants’ interpretations of conditionals at the individual level. Moreover, we elicit the participants’ reflective attitudes through a novel Scorekeeping Task, where the participants commit themselves to a particular interpretation in a case of norm conflicts by criticizing and sanctioning their peers. We apply the Principle of Tolerance by permitting the participants to approach the reasoning tasks with multiple competing formal frameworks while enforcing the requirement that the participants are at least internally consistent in order for them to count as competently implementing any one of them.

In Experiment 1, it was found that two groups of participants could be identified that interpret the indicative conditional differently by either using conditionals to engage in suppositional reasoning (ST) or to express reason relations (DP). DP is by far the largest group, both using the classifications of the participants’ case judgments in phase 1 and the classifications of the participants’ reflective attitudes in the Scorekeeping Task. When the results of the uncertain and-to-if inference task are analyzed relative to these individual profiles across relevance conditions, we find that both groups conform to the theorem of probability theory that $P(C|A) \geq P(A,C)$ at above-chance levels, but only one of the groups conforms to $P(\text{if } A, \text{then } C) \geq P(A,C)$ across relevance conditions. This behaviour matches the interpretations of the conditionals that the participants were assigned to at the individual level.

In addition, the Online Supplementary Materials reports data showing that the alternative hypothesis that the DP participants were following a defective interpretation of probabilities, which would make them more inclined to commit the conjunction fallacy, could not be supported by the results.

Based on the results from Experiment 1, it then appears that what could look like a reasoning error at the group level in an earlier study (Skovgaard-Olsen et al., 2017a) disguises two distinct interpretations of the conditional at the individual level, each of which is consistently followed by different participants in the uncertain and-to-if inference task.

Experiment 2 replicated the main findings from Experiment 1 and showed that they can be generalized to novel items in the irrelevance condition (see Online Supplementary Materials). In Experiment 3, we evaluated the cross-task consistency of our results by conducting an experiment with both the Scorekeeping Task and entailment judgments. Results showed that participants, irrespective of their classification as adhering to ST or DP, largely rejected and-to-if entailments. In fact, the acceptance of such entailments was well predicted by the acceptance of the invalid A-to-and inference. Together, these results suggest that for individuals classified as ST, it is likely that they are committing a reasoning error.

The general tendency to reject the entailment of and-to-if inferences has long-reaching implications as they are valid on many accounts of indicative and counterfactual conditionals, including Pearl's (2000) system, which figures centrally in recent work on causation and counterfactual reasoning (Over, 2017; Lucas & Kemp, 2015). It is possible that prior exposure to irrelevance items in session 1 accounts for why most of the participants allowed for the possibility of 'if A then C' being false while 'A and C' is true in session 2. However, if the ST participants were performing the Ramsey test, then the conditional should be trivially true when considering a situation where the conjunction is true and so it still counts as an error. One possible explanation for these results is that adherence to ST is less stable than adherence to DP.

Another anticipated reaction to these results consists in pointing to pragmatic processes modulating the semantic content postulated by ST. However, these pragmatic processes need to be fleshed out and receive independent validation. In Skovgaard-Olsen et al. (2019), the most popular of such approaches, based on conversational implicatures, was found not to be supported by the results. Instead, Skovgaard-Olsen et al. argue that the data from numerous experiments are most consistent with a conventional implicature interpretation. Conventional implicatures make up a second layer of semantic content as lexicalized parts of the meaning of the sentences in which they occur (Potts, 2007). Since conventional implicatures do not affect the primary truth conditions of these sentences, they are expected to enrich the conditions of rational assertability beyond truth evaluations. Accordingly, if the participants in Experiment 3 interpreted the task as concerning preservation of rational assertability rather than truth preservation, it is possible to account for the results based on a conventional implicature. But in that case, it would be a conventional implicature pointing towards the DP interpretation of the conditional and the interpretation assigned to the ST group would still have been found to be less stable.

7.4 Implications for Rationality Research

Schurz and Hertwig (2019) seek to re-open the discussion of which formal system is the most optimal way of reasoning by comparing reasoning systems in terms of their ability to solve a prediction problem that contributes to the agent's cognitive success across different environments. As part of their argument, Schurz and Hertwig assume that the problem of arbitrating between norms based on conflicting intuitions may be insolvable.

The focus of this paper is not on the evaluative question of which formal system is the most optimal way of reasoning. Instead, we approached the problem of how to assign norm-adherence to participants when multiple conflicting norms are possible—facing the problem of arbitration head-on. The case study illustrates how this normative issue may be approached empirically, and how this can lead to novel, empirical insight. In the final part of the paper, we draw out the key lessons from the case study and set these in the wider context of the role of normative theories in research on human cognition.

Whereas traditional normative research in the psychology of reasoning has largely been focused on developing experimental tasks that have one correct solution so that *absolute* attributions of reasoning errors can be made, this reorientation permits designing tasks where the availability of competing approaches only permit *relative* attributions of reasoning errors based on independent evidence of the participants' own parameter settings (see also Stenning and van Lambalgen, 2008; Elqayam, 2012).

Consequently, we seek to empirically reconstruct the participants' subjective standpoints in order to assess the participants' performance based on their own internal standards. We use empirical data to investigate the extent to which we can use people's normative behaviour towards others to reconstruct internally consistent competence models. In general, normative theories can be evaluated from an *external* perspective by considering which theory is best justified as encoding the correct principles of reasoning, or by attempting to identify a theory-neutral notion of cognitive success (Schurz, 2014). Alternatively, normative theories can be from an *internal* perspective by considering whether the agents committed to a given theory succeed in managing their beliefs in a way that is consistent with their own evaluative standards (Steinberger, 2018). An example would be to identify lack of transitivity in an agent's preferences/choices while presupposing the agent's way of setting up the decision problem. In contrast, reasoning errors in decision making are judged from an external point of view when assessing the parameter settings of the decision problem as the agent construes the decision problem. Examples would be to probe whether the agent takes all of the relevant outcomes into account and assigns them the right probability (Bermudez, 2011, Chap. 3).

Both the internal and external perspectives matter, and both, we argue, are essential to understanding human behaviour. Given the importance of normative considerations, we welcome recent debate about the proper role of normative theories in the study of cognition (e.g., Elqayam & Evans, 2011; Elqayam & Over, 2016). There is much in psychological research practice that can benefit from methodological clarification, and those debates have helped identify areas of confusion. Such confusion should be avoided, but not at the expense of moving normative considerations outside the purview of psychological theory. Rather, it seems essential to understand and employ both the descriptive and the normative in their proper place and the way successful psychological research combines the two.

To be clear: Fallacious is-ought inferences arise when psychologists attempt to infer which theory is best *justified* as a normative theory of reasoning based on the participants' responses themselves (Elqayam & Evans, 2011). This, however, is arguably not what authors in the reasoning literature have sought to do. In particular, Oaksford and Chater (2007) argued that probability theory provides a framework that is *better suited* to the goal of everyday uncertain reasoning (e.g., Oaksford & Chater, 1991), and that *that*, in turn, provides a reason for why participants might construe (and sometimes misconstrue) what experimenters considered to be logical reasoning tasks as probabilistic ones. In other words, the paradigm shift in the reasoning literature from deduction to probabilistic reasoning combined external considerations about what type of reasoning would be efficacious in everyday contexts, that is, an instrumental, normative consideration, with evaluation of participant responses to infer that that kind of reasoning was indeed what participants were, descriptively, engaged in.

Our case study helps clarify this, by showing how the descriptive work of norm attribution is distinct and pursued separately from questions about the foundations for the normative status of those putative 'norms' themselves. What norms people follow is a different question from what makes those 'norms' norms. Hence it is entirely possible to pursue the attribution question non-fallaciously. This matters because, arguably, normative theories have been incredibly valuable to psychology, and it would be detrimental to abandon them. For example, the so-called probabilistic turn in reasoning (or the "new paradigm") has widely been hailed a success (e.g., Evans, 2012), but that 'turn' was directly fuelled by an interest in what participants *should* do, that is, by normative questions.

Normatively motivated research has given rise to tighter, better models than before: Oaksford and Chater's (1994) work prompted the first *quantitative* models of what had traditionally been viewed as 'logical' reasoning tasks, thus providing considerable *descriptive gains* over previous theoretical accounts of these tasks which had merely predicted directional differences across experimental conditions (see Hahn, 2009).

In fact, this is not an isolated, historic coincidence. Closely related to reasoning, the last decade has seen a rise of interest in argumentation within cognitive psychology. Long seen as the purview solely of philosophy and education (for exceptions see, Rips 1998, 2002; Rips et al., 1999), what empirical work there was (see, e.g., Kuhn, 1991; Aufschnaiter et al. 2007) was limited by the lack of resolution in the available normative standards: logic had little to say about everyday informal argument and the extremely limited evaluative framework of the Toulmin model (Toulmin, 1957) afforded only very crude tools for studying argumentation. The Toulmin framework asks simply whether claims are given reasons in support, and whether those reasons have themselves been challenged, but lacks any means to evaluate the quality of those reasons or challenges. Bayesian argumentation has enabled quantitative prediction about very specific factors, such as source reliability, strength of arguments and their interaction, in a way that intersects with large body of work on evidential and causal reasoning (e.g., Pearl, 1988; Griffiths & Tenenbaum, 2005; Hahn & Oaksford, 2009; Fenton, Neil & Lagnado, 2013; Sloman & Lagnado, 2015, and references therein). In other words, developments with respect to normative theories have extended the methodological arsenal of psychologists and the substantive research questions that can be pursued.

Furthermore, this is in no way limited to reasoning or reasoning related areas such as argumentation. Normative considerations are pervasive across cognition from perception, through judgement and decision-making, categorization to various aspects of language processing and language acquisition. Here too, normative models have driven theoretical research, both in terms of questions asked and in terms of methodology (for examples, see e.g., Hahn, 2014 and references therein). For example, ideal observer analysis which has had tremendous success in the study of perception (e.g., Geisler, 2012) draws on the formal tools of probability and decision theory to specify a model of optimal performance given the available input for a task. Behavioural studies then compare actual human performance to the performance of this ideal agent (see, e.g., Geisler, 1989; Legge, Klitz, & Tjan, 1997; Sims, Jacobs, & Knill, 2012). In a process of iterative refinement, human performance and ideal observer are brought into ever closer correspondence by incorporating into the ideal observer details of the human system. Ideal observer analysis is a tool for clarifying mechanism and process that seeks to understand the system as ‘doing the best it can do’ given the available hardware. It combines descriptive and normative by linking up behavioural prediction, mechanistic and functional explanation, in what can be viewed as a methodological formalization of the principle of charity. Many of the most high-profile studies in the field of perception in the last decade fall under this general approach (e.g., Ernst & Banks, 2002; Najemnik & Geisler, 2005; Hillis, Ernst, Banks & Landy, 2002).

Within cognitive psychology, similar programs can be found under the header of bounded rationality or bounded optimality. Howes et al. (2009), for example, stress how rational norms can aid the disambiguation between competing theories and assist in the identification of underlying cognitive universals above and beyond the demand characteristics of experimental tasks. However, probably the most consequential in terms of sheer volume of research has been the advent of the use of optimal models from economic theory as an organizing framework for cognitive neuroscience and neuro-biology (e.g., Glimcher, 2004; Glimcher & Rustichini, 2004; Glimcher et al., 2009; and references therein; Trommershäuser, Maloney, & Landy, 2009, and references therein). Here, what is optimal provides a bound on what is a priori possible, against which actual performance can then be compared in order to – *descriptively* – understand it. This shift, and the flood of research it has prompted, was brought about not by an interest in 'rationality', but by the increasing realization that thinking about neural processes purely in terms of 'reflex'-based approaches is inadequate (Glimcher, 2004).

In the context of all of this research, ranging from neurobiology and neuroscience, through perception to decision-making, reasoning and argumentation, normative and descriptive questions need to be distinguished (else fallacious is-ought inferences may indeed ensue). But it is equally erroneous to think of these questions as entirely separate, as recommendations of 'descriptivism' seem to imply. The claim there seems to be that normative theories such as Bayes' rule may be taken simply descriptively as "computational level theories", stripping them of their 'normative baggage' (Elqayam & Evans, 2011; Elqayam, 2012). Presumably, this intended interpretive switch is expected to leave empirical research not just without loss, but actually improved. What that gain is meant to consist of, is, however, left unclear. More importantly, however, it seems unlikely that present programs could be sustained *without loss*: this is because these recommendations, arguably, misconstrue what computational level theories actually *are*. In Marr's words, a computational level theory involves the following:

> "Its important features are (1) that it contains separate arguments about ***what is computed*** and ***why*** and (2) that the resulting operation is defined uniquely by the constraints it has to satisfy." (Marr, 1982: p. 23)

Normative considerations are essential here. They provide a *functional explanation*, which explicates "what is computed" in terms of inferentially characterized capacities that introduce a criterion for correct/incorrect performance (Cummins, 1983) and specifies an answer to the "why?" question by specifying the *benefits* to the agent of following those recommendations. On such benefits, the normative frameworks of classical logic and probability theory have offered powerful reasons for adherence: probabilistic coherence protects from bets against nature one cannot win, probabilistic coherence coupled with the use of Bayes' rule for belief revision minimizes the inaccuracy of our beliefs (as measured by the Brier score, Pettigrew, 2016), and maximise expected utility (Rosenkrantz, 1992).

While mere "endorsement" of a rule or procedure may suffice (at least in some circumstances) to establish a normative basis (see, e.g., the discussion in Hart, 1994; Corner & Hahn, 2013), such endorsement, in and of itself, provides no basis for the functional level explanation that computational level theories seek to provide. That question is asking why something would be a good thing for me to do, not just whether or not I want to do it. That 'why' is what the 'pragmatic' justification of any putatively normative theory must address. And because that justification is 'external', it can be separated from the internal perspective that norm attribution empirically requires.

The requirements of computational level theories are also not undercut by pointing to linguistics as a role model for a purely descriptive use of *competence models* as Elqayam and Evans (2011) do. The basis of their analogy between linguistics and psychology is the following observation. The study of language has long drawn on competence/performance distinctions to bridge the gap between the utterances a particular grammar might license and those that are observed in actual real-world utterances. In the study of language, research aimed at seeking to identify the competence model (grammar), is entirely distinct from question of whether that competence model is prescriptive or not. "Grammar" in the context of linguistics is not a prescriptive notion embodying a concept of 'good language' but a generative system that allows language users to generate well-formed sentences, where 'well-formedness is relative to specific grammar, and the grammars of different English speakers need not be and will not be exactly the same.

However, 'well-formedness" is itself an inherently normative notion. So Elqayam and Evans (2011) miss the mark when they suggest that "Competence" is not intended to be contrasted with "incompetence," but rather with performance, that is, the instantiation of linguistic competence in actual speech" (p. 239). This makes it sound as if no delineations between competence vs. incompetence (or grammaticality vs. ungrammaticality) are drawn in syntax. This is not true. For much of the past 75 years, the distinction between allowed and disallowed sentences within a language have formed the basic datum of linguistic research. In keeping with this, the most elementary criterion of success for any putative grammar is so-called "descriptive adequacy": that is, the ability to correctly identify the well-formed sentences of the language while rejecting the ill-formed ones. Hence theoretical work on acceptability judgments in descriptive grammar like Schütze (1996), which is continuous with contemporary, experimental syntax (Myers, 2009; Sprouse & Almeida, 2013), contains extensive discussion of "good"/"bad" sentences, degrees of badness, deviances, error, violation, and grammatical/ungrammatical sentences. For example, it is often viewed as an error to reject a sentence containing center-embedding (e.g., "The man who the boy who the students recognized pointed out is a friend of mine") as ungrammatical just because of difficulties with parsing it (Chomsky, 1965).

When theoreticians like Sampson (2007) suggest that linguists should dispense with the grammatical/ungrammatical distinction and turn to a bottom-up approach based on corpus analysis, he is making a radical suggestion in direct opposition to decades of linguistic practice that has unsurprisingly spawned considerable debate (e.g., Kertész & Rákosi, 2008). In this debate Sampson (2007) is immediately contradicted by linguists like Pollum (2007) who state that linguistics is inherently normative and relies on the method of reflective equilibrium. Importantly, it remains common ground in this debate that theoretical linguistics should not return to the prescriptive grammar often associated with the eighteenth or nineteenth century (Beal, 2009). Rather the discussion concerns the use of competence models for the purposes of descriptive grammar, which have an *inherent normative content*.

What separates linguistics from other areas of cognitive science concerned, in one form or other, is primarily that linguists typically spend little time with considerations of *external* justification for the normative notions they employ (but see, e.g., Pereira, 2000; Aylett and Turk, 2004; Levy and Jaeger, 2007, on the rise of normative frameworks such as information theory, or Bergen, Levy and Goodman, 2016 on game theory). However, it is also, arguably, a mistake to think of internal and external justification as entirely unrelated. Crucially, the ‘why’ of functional explanations is also inferentially informative with respect to what it is I want to do, without that inference being a fallacious ought-to-is. The reason such non-fallacious inference from ought to is may be required is because of the *identifiability* problem. *Any* not directly observable ‘theory’ will be under-determined by the data (see e.g., Stanford, 2017). But this general, methodological problem is exacerbated in the context of human behaviour, because any specific behavioural response will be influenced by many factors. As a consequence, actual behaviour will only ever *approximate* a computational level theory, raising the explanatory (and inductive) question of how approximate is approximate enough.

These difficulties are well-illustrated by competence theories in linguistics and psycholinguistics. An underlying grammar is not directly observable and can be identified only via inductively fallible empirical measures: for example, acceptability judgments tracking grammaticality, reaction times, or rating tasks. Crucially, these identification inferences about the competence theory are made entirely without recourse to justificatory concerns. Likewise, in our case studies, we treat the different normative systems participants might be seeking to apply as (mere) competence models that we are seeking to identify, without trying to address questions about their normative status per se.

However, normative concerns *can* be informative for this otherwise entirely descriptive pursuit, because they too can help with the identification problem. Many competence theories will, in principle, explain the same finite set of behavioural data. Considerations other than data fit can provide additional constraints that help prune that set: That it would be useful to act a certain way provides a defeasible piece of evidence in support of the fact that that is what I am, in fact, trying to do. It is not sufficient (that would indeed be erroneous is-to-ought inference) but it is similarly fallacious to hold that such utility consideration have no evidential value. And claiming that it doesn't would be directly at odds with our most basic routines for understanding the utterances and actions of others, not just in science, but in our daily lives. This is what principles of charity encapsulate and throwing away functional considerations is simply throwing away an important methodological tool.

In all of this, the normative work itself needs to be done: some independent reason for why a procedure is normative needs to be explicitly established, and that reason must connect meaningfully with actual goals of the agent. 'Descriptivism' as advocated by Elqayam and colleagues doesn't obviate the need for that: one still needs to do the normative work. And that work may be hard because agents may have multiple epistemic (and non-epistemic) goals. But stepping away from normative theories altogether comes at too heavy a cost.

What is required are not broad brushstroke solutions, but detailed engagement with the issues in the context of particular problems. There is a need to refine the methodological arsenal, not to restrict it. This is what we have sought to provide with the present case study:

What we hope to have shown is that there is a fruitful role that normative theorizing can play in experimental psychology that consists in making internal evaluations of the participants' performance based on competence models assigned on the individual level, even for cases where multiple, conflicting norms can be applied. We thereby directly address the problem of arbitration, which is one of the main practical problems that Elqayam and Evans (2011) point to in the application of norms to empirical investigations of reasoning.

The Scorekeeping Task constitutes a new tool for measuring the participants' reflective attitudes. It is the reflective attitudes that competence theories of human reasoning generally aim to describe (e.g., Macnamara, 1986), very much like how judgments of grammaticality are supposed to reveal our linguistic competence (e.g., Chomsky, 1965). Yet in studies of reasoning, experimental procedures for measuring the participants' considered judgments have been neglected. The central idea behind the Scorekeeping Task is that the participants' norm adherence is revealed by the norms they use to criticize and sanction their peers with. One domain where the Scorekeeping task appears to be particularly promising is decision making under risk and uncertainty, where a considerable amount of theoretical developments has been based on the rejection of certain norms (e.g., Allais, 1953; Birnbaum, 2008; Kahneman & Tversky, 1979). For example, Birnbaum and colleagues have reported a series of '*choice paradoxes'* that reject Cumulative Prospect Theory (for a review, see Birnbaum, 2008). Different accounts attempt to accommodate these paradoxes by attributing them to attention biases, distractions, differential weighting of better/worse outcomes, among other notions (e.g., Cenci et al., 2014; Pandey, 2018). One could use the Scorekeeping Task to determine whether individuals' judgments are consistent with their sanctioning of others' choices. These results should be able to clarify exactly which paradoxes can be attributed to some kind of perceptual/reasoning errors (e.g., violations of stochastic dominance), and which indeed reflect the core principles underlying the comparison of options (e.g., a viewpoint-dependent weighting of outcomes). Important here is the notion that no one-size-fits-all solution is likely to work, given the heterogeneity that is consistently found across individuals (see Regenwetter & Robinson, 2017).

7.5 Conclusion

A normative and empirical framework was put forward in this paper for attributing reasoning errors in cases where there are multiple, conflicting norms that could serve as competence models. A new task was introduced for eliciting the participants' reflective attitudes, and individual profiles of the participants were made, which assessments of correct and incorrect reasoning were made relative to.

In the case study of conditional reasoning, it was seen that at least two interpretations of indicative conditionals could be separated based on the participants' probability assignments, and that the participants consistently followed these interpretations when assigning probabilities to the conclusions of uncertain and-to-if inferences. In a third experiment, it was found, however, that when the participants were tested after a temporal delay in a task eliciting entailment judgments, only one of these two groups of participants showed a consistent pattern by rejecting the entailment from and-to-if just as in their probability assignments in the uncertain and-to-if task. Moreover, participants' own assessment of how consistently they had responded across experimental sessions turned out to be an unreliable guide.

The results thus have repercussions for how possible it is to internally reconstruct consistent competence models of participants when reasoning with conditionals. In short, we demonstrated the utility of our method by showing novel and interesting empirical conclusions for the psychology of reasoning. However, the method itself is entirely general, and can be used in any domain in which normative considerations guide descriptive research (e.g., decision making).

Finally, the case studies of this paper allowed us to clarify both the importance of normative theories to the descriptive understanding of individual's behaviour and to entangle some of the confusions about seemingly fallacious is-to-ought inferences highlighted by the recent literature. Setting aside normative theories in psychology would mean setting aside a rich source of interesting research questions and a central methodological tool. This makes it imperative that psychological research gets the conceptual issues right.

References

Adams, E. (1965). The Logic of Conditionals. *Inquiry*, *8*, 166-97.

Ali, N., Schlottmann, A., Shaw, A., Chater, N., and Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In Oaksford, M. and Chater, N. (Ed.), *Cognition and Conditionals* (pp. 117-134). Oxford: Oxford University Press.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, 21, 503–546.

Arlo-Costa, H. (2007). The Logic of Conditionals. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/.

Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2007). Argumentation and the learning of science. *Contributions from science education research*, 377-388.

Aylett, M., and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language & Speech*, *47*, 31-56.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 340-412.

Ball, L. J. (2013). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *Journal of Cognitive Psychology*, *25*, 183-191.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning*, *19*, 308-328.

Beal, J. C. (2009). Three Hundred Years of Prescriptivism (and Counting). In van Ostade, I. T. and van der Wurff, W. (Eds.), *Current Issues in Late Modern English. Linguistic Insights: Studies in Language and Communication 77* (pp. 35–55). Bern: Peter Lang.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Bermúdez, J. L. (2011). *Decision Theory and Rationality*. Oxford: Oxford University Press.

Bhatt, R. and Pancheva, P. (2006). Conditionals. In Everaert, M. and van Riemsdijk, H. (Eds.), *The Blackwell companion to syntax* 1 (pp. 638–687). Oxford: Blackwell.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4,* 243-249

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115*, 463–501.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Brandom, B. (1994). *Making it Explicit*. Cambridge, Mass.: Harvard University Press.

Buchak, L. (2013). *Risk and Rationality*. Oxford: Oxford University Press.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1-28.

Bürkner, P., & Vuorre, M. (2018, June 23). Ordinal Regression Models in Psychology: A Tutorial. https://doi.org/10.31234/osf.io/x8swp

Carnap, R. (1937). *The Logical Syntax of Language*. London: Kegan Paul.

Cenci, M., Corradini, M., Feduzi, A., & Gheno, A. (2014). Half-full or half-empty? A simple model of decision making under risk. *Journal of Mathematical Psychology, 68,* 1-5.

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition*, *113*(3), 350-364.

Chater, N., Felin, T., Funder, D. C., Gigerenzer, G., Koenderink, J. J., Krueger, J. I., Noble, D., Nordli, S. A., Oaksford, M., Schwartz, B., Stanovich, K. E., and Todd, P. M. (2018). Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, *25*, 793-826.

Chater, N. and Oaksford, M. (2012). Normative Systems: Logic, Probability, and Rational Choice. In K. J. Holyoak and R. G. Morrison (Ed.), *The Oxford Handbook of Thinking and Reasoning* (pp. 11-21). Oxford: Oxford University Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287-291.

Cheng, P.W. (1997). From covariation to causation*:* A causal power theory. *Psychological Review, 104,* 367-405.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cooper, R. P. (2002). *Modeling High-Level Cognitive Processes*. Mahwah, NJ: Erlbaum.

Corner, A., & Hahn, U. (2009). Evaluating science arguments: evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*(3), 199.

Corner, A., & Hahn, U. (2013). Normative theories of argumentation: are some norms better than others?. *Synthese, 190*(16), 3579-3610.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological review, 121*(3), 463.

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology, 89*, 106-133.

Cruz, N., Baratgin, J., Oaksford, M. and Over, D.E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6* (192).

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Douven, I. (2015). *The Epistemology of Indicative Conditionals. Formal and Empirical Approaches*. Cambridge: Cambridge University Press.

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese, 194*, 1541-54.

Edgington, D. (1995a). On Conditionals. *Mind, 104*, 235-327.

Edgington, D. (1995b). Conditionals and the Ramsey Test. *Proceedings of the Aristotelian Society Supplementary Volume*, *69*, 67-86.

Elqayam, S. (2012). Grounded Rationality: Descriptivism in epistemic context. *Synthese*, *189,* 39-49.

Elqayam, S. and Evans, J. St. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*, 233-90.

Elqayam, S. and Over, D. (2016) (Ed.). *From is to ought: The place of normative models in the study of human thought*. Lausanne: Frontiers Media.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.

Evans, J. S. B. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*(1), 5-31.

Evans, J. St. B. T., and Elqayam, S. (2011). Towards a descriptivist psychology of reasoning and decision making. *Behavioral and Brain Sciences, 34*, 275–290.

Evans, J. St. B. T., Handley, S. J., Neilens, H., and Over, D. E. (2007). Thinking about conditionals. A study of individual differences. *Memory & Cognition, 35*, 1772-1784.

Evans, J. St. B. T. and Over, D. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Thompson, V. A. & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology, 6*, 398.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological review*, *96*(2), 267.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771-781.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*(2), 195-204.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. MIT press.

Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science*, *306*(5695), 447-452.

Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (2009). Introduction: A brief history of neuroeconomics. In *Neuroeconomics* (pp. 1-12).

Goodman, N. (1965). *Fact, fiction, and forecast*. Indianapolis: Bobbs-Merrill.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334-384.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Hahn, U. (2009). Explaining more by drawing on less. *Behavioral and Brain Sciences, 32*(1), 90-91.

Hahn, U. (2014). The Bayesian boom: good thing or bad?. *Frontiers in psychology, 5*, 765.

Hahn, U. and Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review, 114*(3), 704-732.

Hart, H.L.A (1994, first edition 1961). *The Concept of Law*, 2nd ed. ed.P. Bulloch and J. Raz . Oxford: Clarendon Press.

Hertwig, R. and Gigerenzer, G. (1999). The ‚Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors. *Journal of Behavioral Decision Making, 12,* 275-305.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138,* 211-237.

Hilbig, B. E., and Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic bulletin & review*, *21*, 1431-1443.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science*, *298*(5598), 1627-1630.

Hilton, D. J. (1995). The Social Context of Reasoning: Conversational Inference and

Rational Judgment. *Psychological Bulletin*, *118*(2), 248-271.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological review*, *116*(4), 717.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences, 34(4), 169-188.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874. http://doi.org/10.1037/a0016979

Joyce, J. M. (2004). Bayesianism. In Miele, A. R. and Rawling, P. (Ed.), *The Oxford Handbook of Rationality* (pp. 132-155). Oxford: Oxford University Press.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Kahneman, D., and Tversky, A. (1996). On the Reality of Cognitive Illusions. *Psychological Review*, *103*(3), 582-591.

Kellen, D., and Klauer, K. C. (2018). Elementary signal detection and threshold theory. In Wagenmakers, E.-J. (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1-39). Vol. V (4th ed.). New York: John Wiley & Sons.

Kertész, A., Rákosi, Cs. (2008). Conservatism vs. Innovation in the (Un)grammaticality Debate. In Kertész, A., Rákosi, Cs. (Eds.), *New Approaches to Linguistic Evidence* (pp. 61-84). Frankfurt am Main: Lang.

Kirk, R. E. (2013). *Experimental Design. Procedures for the Behavioral Sciences.* London: Sage Publications. (*4th Edition*)

Klauer, C. K. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review*, *106*, 216-223.

Kneer, M. and Machery, E. (2019). No Luck for Moral Luck. *Cognition*, *182*.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Krzyżanowska, K. (2015). *Between "If" and "Then": Towards an empirically informed philosophy of conditionals*. PhD dissertation, Groningen University. Retrieved from http://karolinakrzyzanowska.com/pdfs/krzyzanowska-phd-final.pdf

Kuhn, D. (1991). *The Skills of Argument*. Cambridge: Cambridge University Press.

Lee, C. J. (2006). Gricean Charity: The Gricean Turn in Psychology. *Philosophy of the Social Sciences*, *36*, 193-218.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*, 29-41.

Lee, M. D., Steyvers, M., and Miller, B. (2014). A cognitive model for aggregating people's rankings. *PloS one*, *9*.

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an ideal-observer model of reading. *Psychological review*, *104*(3), 524.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), (p. 849-856). Cambridge, MA: MIT Press.

Lucas, C. G. and Kemp, C. (2015). An Improved Probabilistic Account of Counterfactual Reasoning. *Psychological Review*, *122*(4), 700-734.

Macnamara, J. (1986). *A Border Dispute. The Place of Logic in Psychology*. Cambridge, MA.: The MIT Press.

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2018, October 14). Reasons or Rationalisations: Inconsistencies in Endorsing, Articulating and Applying Moral Principles. https://doi.org/10.31234/osf.io/pcsfj

Mercier, H. and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57-74.

Mercier, H. and Sperber, D. (2017). *The Enigma of Reason*. Cambridge, MA.: Harvard University Press.

Myers, J. (2009). Syntactic judgment experiments. *Language & Linguistics Compass*, *3*(1), 406-423.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387.

Nickerson, R. S. (2015). *Conditionals and Reasoning*. Oxford: Oxford University Press.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian Models. *Frontiers in Psychology*, *5* (332), 1-5.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language*, *6*(1), 1-38.

Oaksford, M, and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., and Chater, N. (2010). C*ognition and conditionals: Probability and logic in human thinking*. Oxford, England: Oxford University Press.

Oaksford, M., and Chater, N. (2017). Causal Models and Conditional Reasoning. In M. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 327-346). Oxford: Oxford University Press.

Oberauer, K., Geiger, D., Fischer, K., and Weidenfeld, A. (2007). Two Meanings of If? Individual Differences in the Interpretation of Conditionals. *Quarterly Journal of Experimental Psychology*, *60*, 790-819.

Olsen, N. S. (2014). *Making ranking theory useful for psychology of reasoning*. PhD dissertation, University of Konstanz. Retrieved from http://kops.uni-konstanz.de/handle/123456789/29353.

Over, D. (2017). Causation and the Probability of Causal Conditionals. In Waldmann, M. (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 307-325). Oxford: Oxford University Press.

Over, D. E., and Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In Linden J. Ball, & Valerie A. Thompson (Ed.), *International handbook of thinking and reasoning* (pp. 434-450). Hove, Sussex: Psychology Press

Over, D. and Evans, J. St. B. T. (2003). The Probability of Conditionals: The Psychological Evidence. *Mind and Language*, *18*, 340-58.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Pandey, M. (2018). The opportunity-threat theory of decision-making under risk. *Judgment and Decision Making, 13,* 33.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible reasoning. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Pereira, F. (2000). Formal grammar and information theory: together again?. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1239-1253.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning, 19,* 329-45.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Ed.), *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003).*

Potts, C. (2007). Conventional implicatures, a distinguished class of meanings. In Gillian Ramchand, & Charles Reiss (Eds.). *The Oxford handbook of linguistic interfaces* (pp. 475–501). Oxford: Oxford University Press

Pullum, G.K. (2007). Ungrammaticality, rarity, and corpus use. *Corpus Linguistics and Linguistic Theory*, 3, 33-47.

Ragni, M., Kola, I. and Johnson-Laird, J. (2017). The Wason Selection Task: A Meta-Analysis. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 980-985.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Read, S. (1995). Conditionals and the Ramsey Test. *Proceedings of the Aristotelian Society Supplementary Volume, 69,* 47-65.

Regenwetter, M. & Robinson, M. M. (2017). The construct–behavior gap in behavioural decision research: A challenge beyond replicability. Psychological Review, 12, 533–55

Rescher, N. (2007). *Conditionals*. Cambridge, MA.: The MIT Press.

Rips, L. J. (1998). Reasoning and conversation. *Psychological Review*, *105*(3), 411-441.

Rips, L. J. (2002). Circular reasoning. *Cognitive Science, 26*(6), 767-795.

Rips, L. J., Brem, S. K., & Bailenson, J. N. (1999). Reasoning dialogues. *Current Directions in Psychological Science, 8*(6), 172-177.

Rosenkrantz, R. D. (1992). The justification of induction. *Philos Sci 59*(4), 527–539.

Rott, H. (1986). Ifs, though, and because. *Erkenntnis, 25*, 345–70.

Ryle, G. (1950). ‘I‘, ‘so’, and ‘because’. In M. Black (Ed.), *Philosophical Analysis* (pp. 323–40). Ithaca, NY: Cornell University Press.

Rawls, J. (1971). *A theory of justice*. Cambridge, Mass.: Belknap Press.

Sampson, G.R. (2007). Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*, 3, 1-32.

Schurz, G. (2014). Cognitive success: instrumental justifications of normative systems of reasoning. *Frontiers in Psychology*, 5, 1-16.

Schurz, G. and Hertwig, R. (2019). Cognitive Success: A Consequentialist Account of Rationality in Cognition. *Topics in Cognitive Science*, 1-30.

Schütze, C. T. (1996). *The empirical base of linguistics. Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in psychology, 5* (316).

Skovgaard-Olsen, N. (2016a). Ranking Theory and Conditional Reasoning. *Cognitive Science, 40*, 848-880.

Skovgaard-Olsen, N. (2016b). Motivating the Relevance Approach to Conditionals, *Mind & Language, 31*(5), 555-579.

Skovgaard-Olsen, N. (2017). The problem of logical omniscience, the preface paradox, and doxastic commitments. *Synthese, 194*(3), 917-939.

Skovgaard-Olsen, N., Collins, P., Krzyzanowska, K., Hahn, U., and Klauer, C. K. (2019). Cancellation, Negation, and Rejection. *Cognitive Psychology, 108*, 42-71.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016). The Relevance Effect and Conditionals. *Cognition, 150*, 26-36.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2017a). Relevance and Reason Relations. *Cognitive Science, 41*(5), 1202-1215.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., and Klauer, K. C. (2017b). Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'. *Thinking and Reasoning, 23*(4), 449-482.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology, 66*, 223-247.

Slovic, P. and Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science*, *19*(6), 368-373.

Spohn, W. (1991). A Reason for Explanation: Explanations Provide Stable Reasons. In Spohn, W., van Fraasen, B. C., and Skyrms, B. (Eds.), *Existence and Explanation. Essays Presented in Honor of Karel Lambert* (pp. 165-196). Dordrecht: Kluwer.

Spohn, W. (1993). Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein? In Eckensberger, L. and Gähde, U. (Ed.), *Ethik und Empirie. Zum Zusammenspiel von begrifflicher Analyse und erfarungswissenschaftlicher Forschung in der Ethik* (pp. 151-196). Frankfurt a.M: Suhrkamp.

Spohn, W. (2012). *The Laws of Beliefs*. Oxford: Oxford University press.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science, 37*, 1074–1106.

Sprouse, J. & Almeida, D. (2011). The role of experimental syntax in an integrated cognitive science of language. In C. Boeckx and K. Grohmann (Eds.), *The handbook of Biolinguistics* (pp 181-202). Cambridge: Cambridge University Press.

Stanford, Kyle (2017). Underdetermination of Scientific Theory. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition). Retrieved from <https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/>.

Stein, E. (1996). *Without Good Reason. The Rationality Debate in Philosophy and Cognitive Science*. Oxford: Clarendon Press.

Steinberger, F. (2016). How Tolerant Can You Be? Carnap on Rationality. *Philosophy and Phenomenological Research*, *92*(3), 645-668.

Steinberger, F. (2018). Logical pluralism and logical normativity. *Philosophers Imprint*. Retrieved from https://floriansteinberger.weebly.com/research.html (In press)

Stenning, K. and van Lambalgen, M. (2004). A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, *28*(4), 481-529.

Stenning, K., and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT University Press.

Strawson, P. F. (1986). 'If' and '⊃'. In R. Grandy and R.Warner (Ed.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends* (pp. 229–42). Oxford: Clarendon Press.

Stupple, E. J. N. and Ball, L. J. (2014). The Intersection between Descriptivism and Meliorism in Reasoning Research: Further Proposals in Support of Soft Normativism. *Frontiers in Psychology*, *5*.

Tentori, K., Crupi, V., and Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*, 235–255.

Tentori, K.,Bonini, N., and Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, *28*, 467-77.

Thagard, P. and Nisbett, R. E. (1983). Rationality and Charity. *Philosophy of Science, 50*, 250-67.

Toulmin, S. E. (1957/2003). *The uses of argument*. Cambridge university press.

Trautmann, S. T., and Van De Kuilen, G. (2015). Ambiguity attitudes. In G. Keren, & G. Wu (Ed.) *The Wiley Blackwell handbook of judgment and decision making* (pp. 89-116). Oxford, UK: Wiley.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2009). The expected utility of movement. In *Neuroeconomics* (pp. 95-111).

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

van Rooij, R. and Schulz, K. (2018). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, 1-17.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273-281.

Winter, Y. (2016). *Elements of Formal Semantics*. Edinburgh: Edinburgh University Press.

Yao, G., and Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*, 79–92.

Suppl. Materials:
Bayesian Mixture Model and Individual Variation[99]

The Bayesian mixture model developed here builds on previous regression-based efforts to characterize the different probabilities elicited from individuals (Singmann et al., 2013; Skovgaard-Olsen et al., 2016, 2017a): Generally speaking, for each individual, the elicited values of P(C|A) were used to predict the elicited probability P(if A then C) concerning the same antecedent A and consequent C:

$$\text{P(if A then C)} = \beta_0 + \beta_1 \text{P(C|A)} + \varepsilon,$$

where β_0 and β_1 are the intercept and slope parameters respectively, and ε is the residual term.[100] Parameter estimates were then used to evaluate ST and DP accounts across relevance conditions. Given the conditional probability hypothesis is key to ST, it can be argued that this theoretical account expects β_0 and $\beta_1$to be 0 and 1, respectively. In contrast, the alternative DP account expects β_1 to take on value 1 in the case of positive relevance (PO) but to take on positive values less than one in irrelevance (IR) and negative relevance (NE) conditions, reflecting the penalty that follows from the lack of a (positive) inferential relation between the antecedent A and the consequent C. An evaluation of the two theoretical accounts is then made possible by comparing β_1 estimates across relevance conditions. For example, Skovgaard-Olsen et al. (2016, 2017a) reported $\beta_1$estimates in the irrelevance and negative relevance conditions that were significantly smaller than in the positive relevance condition, in line with the predictions of DP.

The accompanying evaluation of (chance-corrected) probabilistic coherence was based on an approach originally proposed by Evans et al. (2015). As an example, consider the case of the uncertain and-to-if inference, according to which P(Conclusion) ≥ P(Premise) for ST.

[99] Further supplemental materials including all data and analysis scripts are available at: https://osf.io/9fm45/.

[100] For clarity purposes, this description omits subscripts denoting participant and trial, and also glosses over the random-effects structures that enable the estimation of individual differences around group-level means.

Assume that if the elicited probabilities are produced by a pure guessing process, then this process yields probabilities that are uniformly distributed between 0 and 1. It follows that given an elicited value for P(Premise), a guessing-based elicitation would respect P(Conclusion) ≥ P(Premise) with probability 1-P(Premise). Now, consider a dichotomous random variable X_{UAI}, which takes on value 1 when P(Conclusion) ≥ P(Premise) is respected, and 0 when it does not. In order to evaluate whether conformity to P(Conclusion) ≥ P(Premise) occurs at an above-chance rate, one simply has to test whether the difference between X_{UAI} and 1-P(Premise), computed across trials and individuals, is reliably larger than 0. For example, Skovgaard-Olsen et al. (2017a) showed that this difference was significantly above chance in the positive relevance condition, but not in the negative relevance and irrelevance conditions.

Despite its merits, the regression-based approach used so far suffers from important limitations. First, it assumes that the error term ε follows a Normal distribution with mean zero and variance σ_{ε}^2. This error distribution attributes non-zero probability to the occurrence of elicited values outside the 0%-100% scale used. The problem here is not limited to the fact that impossible values are deemed possible by the model, but the fact that this unbounded "error theory" overlooks the important biasing role that errors can have in the occurrence of empirical phenomena such as conservatism, subadditivity, and conjunction/disjunction fallacies (see Costello & Watts, 2014; Hilbert, 2012). For example, for low/high probabilities, errors will systematically lead to elicitations that are biased upwards/downwards. One consequence of these biases is an overestimation of β_0 and an underestimation of β_1 (for a detailed discussion, see Hilbert, 2012).

The second limitation concerns the fact that the adopted regression approach assumes that individuals vary in terms of *degree*, but not in *kind*. Given the notion that individuals can rely on different norms (some might be in line with ST, others with DP), the regression model's tacit assumption that all individuals belong to the same group is ultimately unsatisfactory. For example, Skovgaard-Olsen et al. (2017a) provided evidence in terms of a group-level $\beta_1$estimates below 1 and the occurrence of conformity to P(Conclusion) ≥ P(Premise) at below-chance rates for IR and NE. These results are silent on the actual proportion of individuals that adhere to either ST or DP, and whether the compliance rates with respect to predictions such as the inequality for the uncertain and-to-if inference differs between these two groups.

In order to overcome these limitations, we developed a Bayesian mixture model according to which the predicted relationship between P(if A, then C) and P(C|A) is determined by that individual's adherence to ST or DP. In the positive relevance condition, for individual *i* and a pair *j* of elicited P(if A, then C) and P(C|A) concerning a given antecedent A and consequent C:

$$P(if\ A, then\ C)_{i,j} = \begin{cases} P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 0, 1, \\ \beta_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 2, \end{cases}$$

where $\varepsilon_{i,j}$ come from a *truncated* Normal distribution with mean 0 and variance σ_ε^2 (see the left panel of Figure 1). This distribution is truncated between 0 and 100 in order to limit predictions to the permitted range of responses and to mitigate the biases expected in noisy elicitations (see Costello & Watts, 2014; Hilbert, 2012). The indicator variable w_i^{PO}denotes whether participant *i* in the positive relevance condition follows ST ($w_i^{PO} = 0$), DP ($w_i^{PO} = 1$), or a *saturated model* ($w_i^{PO} = 2$). The latter model can account for any data (it has one parameter $\beta_{i,j}$ per trial pair) and allows us to identify the individuals that cannot be well accounted by either ST or DP (for a discussion, see Hilbig & Moshagen, 2014). In the cases of ST and DP, it is assumed that the individual equates P(if A, then C) and P(C|A).

In the absence of a (positive) reason relation between the antecedent A and consequent C the two accounts make diverging predictions. In the IR condition the predictions are:

$$P(if\ A, then\ C)_{i,j} = \begin{cases} P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{IR} = 0, \\ \theta_i P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{IR} = 1, \end{cases}$$

where θ_i is the discount-penalty parameter of DP-adherent individual *i*. The range of predictions (excluding noise) made by the two models are illustrated in Figure 1.

The mixture model was implemented in a Bayesian framework: In a nutshell, the information (or ignorance) regarding the model parameters is represented by *prior distributions*. The observed data is then used to update our knowledge about the parameters, resulting in *posterior parameter distributions* (Gelman et al., 2014; Kruschke, 2014; M. D. Lee & Wagenmakers, 2014). Based on the posterior probabilities of the indicator variables w_i we can easily classify each individual per condition as adherents of ST, DP, or neither (see M. D. Lee, 2016).

One important aspect of these model-based classifications is that they take into account the flexibility of the two accounts (for a discussion, see M. D. Lee, 2016). As shown in Figure 1, whereas ST is bound to predict that data follow the main diagonal, DP is also able to accommodate data falling along a monotonic function below the main diagonal, with ST being a special case of DP when $\theta = 1$. Given that there is currently no theoretical claim with respect to the shape of this function, we are assuming that it is linear. Due to its greater flexibility, the classification is therefore biased against DP, requiring sufficient evidence from the data in order to justify the additional flexibility.

The key parameters of interest in this analysis are the posterior probabilities of $w_i = 1$ obtained in the positive relevance and irrelevance conditions. In the positive relevance condition, when the mean of this posterior probability was estimated to be below or equal to .50, the individual was classified as following the saturated model. When the mean is estimated to be larger than .50, the individual was classified as following ST/DP. In the irrelevance condition, these same ranges of values led to the ST and DP classifications, respectively. The individual classifications were used to produce different, chance-corrected estimates of probabilistic-coherence phenomena such as:

$$\text{P(Conclusion)} \geq \text{P(Premise)}$$
$$\text{P(C|A)} \geq \text{P(A,C)}$$

Specifically, we estimated how much the observed rate of probabilistic-coherent elicitations deviates from chance, a deviation that was quantified on an effect-size scale. We can then test whether the posterior distribution of these deviations is reliably above or below zero by inspecting whether the value zero is included in their 95% credibility intervals (i.e. $\Phi(K_{i,j}) = 1 -$ P(Premise); Kruschke, 2016).

For participant *i*, the probability that her response to a given item-pair *j* conformed to a given inequality is given by $\Phi(\Delta_i + K_{i,j})$, with $\Phi()$ being the probability function of the standard Normal distribution. Parameter $K_{i,j}$ is a correction term for participant *i* and item-pair *j* such that $\Phi(K_{i,j})$ corresponds to the probability that the responses to a given item-pair were inequality-conforming by chance alone (Singmann, Klauer, & Over, 2014). Parameter Δ_i corresponds to that individual's displacement from chance (i.e., when Δ_i is positive, that individual produces inequality-conforming responses at an above-chance rate). Using a hierarchical framework, these individual parameters were assumed to come from a Normal group-level distribution, with mean μ_Δ and standard deviation σ_Δ. If individuals in general conform to P(Conclusion) ≥ P(Premise) or P(C|A) ≥ P(A,C), then their respective μ_Δ should be consistently above 0 (i.e., the probability of μ_Δ being below 0 should be very small). These parameters were estimated separately for individuals classified as ST and DP in the irrelevance condition.

A very similar hierarchical approach was used to model the relative probability of an individual judging the no-difference justification (in line with ST) as most convincing after having seen both sides, as well as the relative probability attributing the HIT to such justification. We also used the individual classifications to test for differences in theoretically-relevant variables, such as the occurrence of conjunction fallacies (Tversky & Kahneman, 1983), the interpretation of probability (Hertwig & Gigerenzer, 1999), manifestations of argumentative skills (Kuhn, 1991), and demographic variables such as college education and training in probability (see below). The ranking of probability interpretations was analyzed using a Thurstonian model assuming that the probability of a given rank-order corresponds to that probability that a sample from latent distributions (one distribution per interpretation) produces that rank order. These latent distributions are assumed to be Normal with a given mean and variance. We assumed that all distributions are Gaussian and have the same variance (a common assumption in these models, see Kellen & Klauer, 2018). Without loss of generality, we fixed the mean of one of these interpretations to zero. Details on the estimation of these parameters can be found elsewhere (M. D. Lee, Steyvers, & Miller, 2014; Yao & Böckenholt, 1999). The posterior-parameter distributions of the mixture model were estimated via Gibbs sampling using the general-purpose software JAGS (Plummer, 2003). Chain convergence was confirmed via the R-hat statistic and visual inspection.

The phase-1 classifications obtained in Experiment 2 are given in Figure A1, whereas the coherence measures are provided in Figure A2.

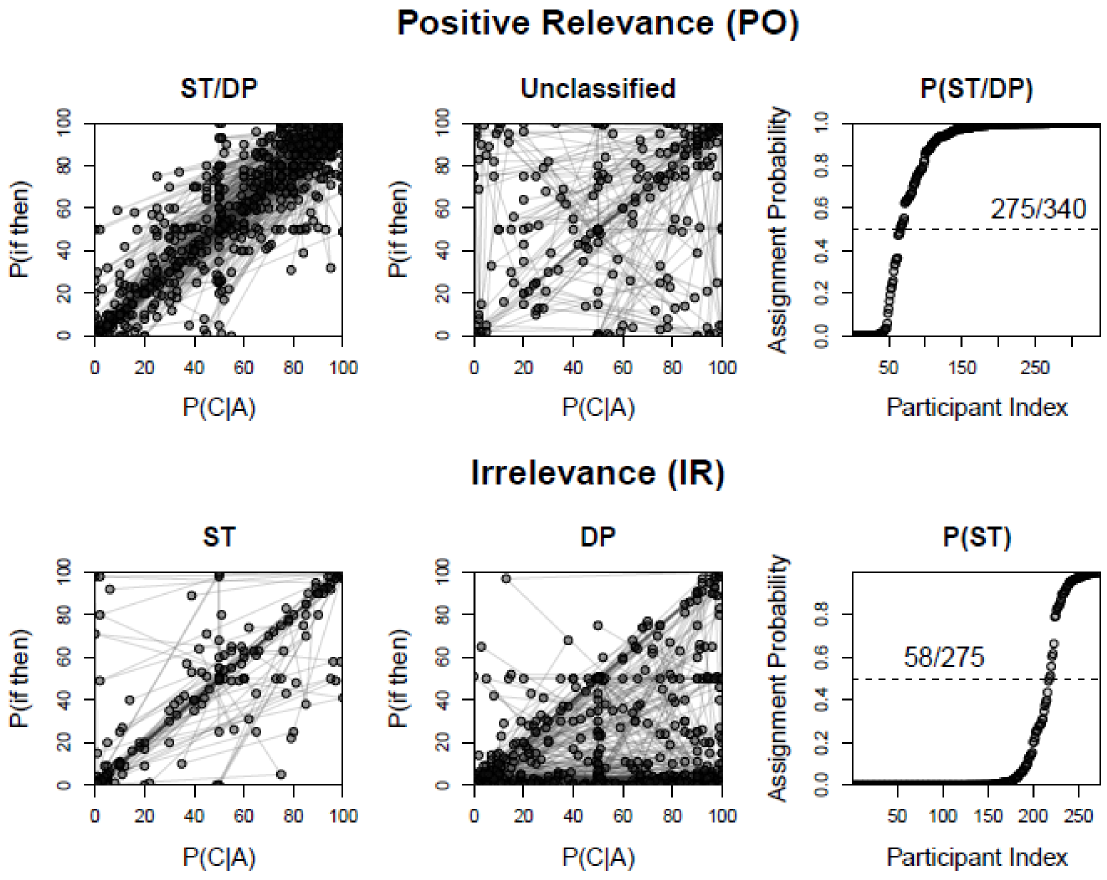


Figure A1. Left and Center Panels: Individual data associated to the phase 1 classifications in Experiment 2. Right Panels: Individuals' posterior classifications (note that in the irrelevance condition, only participants classified as ST/DP in the positive relevance condition were considered).

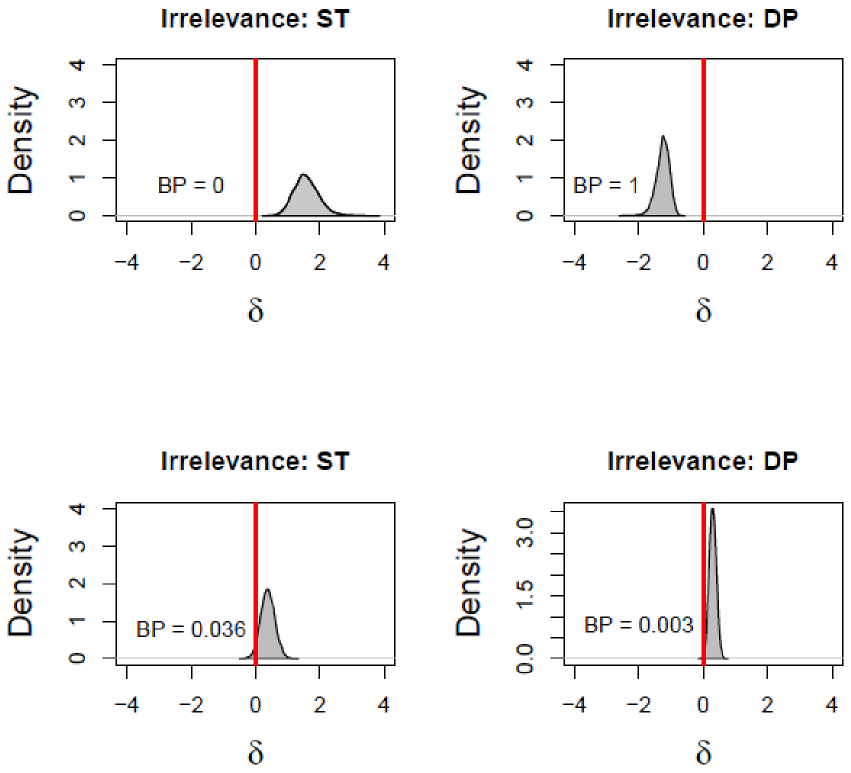


Figure A2. Posterior distributions of the deviations of the tested inequalities from chance-level occurrence (represented on an effect-size scale) in Experiment 2. The vertical lines indicate effect size 0 and BP corresponds to the probability of samples from the posterior distributions taking on values below 0. In the left panels we depict the posterior distributions for participants classified as ST and DP (the latter corresponding to the more peaked distributions).

Phase 4 of Experiments 1 and 2: Individual Variation

Phase 4 served the purpose of testing for further covariates that would characterize participants that were classified as interpreting the conditional according to ST and DP.

In Experiment 1, phase 4 tested for whether the participants differed in their tendency to commit the conjunction fallacy and their interpretation of probability, based on a suggestion in Skovgaard-Olsen et al. (2017a). One possibility is that what distinguishes the ST participants from the DP participants is the latter having a defective understanding of probabilities. This possibility echoes results previously reported by Tentori, Crupi, and Russo (2013), who found that the participants committing the conjunction fallacy are misled by the degree of confirmation of the added conjunct. Participants were presented with four pages separated in two blocks. The first block contained the less well-known Bill version of the conjunction fallacy task presented in Tversky and Kahneman (1983). Following Hertwig and Gigerenzer (1999), the participants were instructed in a second block of phase 4 to help a fictive user named Ludwig to understand the instructions of the previous task. The participants were told that English was not the native language of Ludwig and that Ludwig was a bit uncertain about how to interpret the word 'probability'. The task of the participants was to provide paraphrases of the term 'probability' that would help Ludwig understand the instructions. To do this, the participants were instructed that they should rank-order paraphrases of probability in terms of relative frequencies, propensities, plausibility, and subjective degree of belief according to which one was most adequate and that they could reselect their responses (see the Supplementary Materials).

In Experiment 2, phase 4 evaluated individuals' argumentation skills using an adaption of Kuhn's (1991) task. To classify the participants' responses a coding manual was written based on Kuhn (1991), which three coders applied independently. In this task, the participants are assessed for their level of argumentative skills based on their ability to:

(1) Produce a causal hypothesis about why children fail at school,
(2) Produce genuine evidence stating a correlation or co-variation that would substantiate their claim as opposed to, for instance, providing pseudo-evidence which merely elaborates their own theory through illustrations, and arguments from analogy or general assumptions about human nature,
(3) produce a possible counterargument to their own theory targeting, for instance, its sufficiency or necessity,
(4) recognize the principled possibility of error of their own theory, and
(5) recognize that they are presented with weak, underdetermined evidence, which is compatible with several causal hypotheses instead of reading their own theory into the evidence.

In an extensive coding manual, the coders were instructed how to classify the participants' open-ended responses based on Kuhn's (1991) conceptual distinctions (see Supplementary Materials). Three independent coders classified all of the responses. When there was disagreement, a simple majority rule was used.

Phase 4 (Experiment 1)

We estimated the occurrence of the conjunction fallacy in the context of the Bill case (Kahneman & Tversky, 1983), and evaluated participants' interpretation of probabilities

(Gigerenzer & Hertwig, 1999). With respect to the occurrence of the conjunction fallacy, the rate at which it occurred was high, but similar across individuals adhering to ST (.43 [.26, .59]) and DP (.48 [.40, .55]). Finally, the ranked interpretations of probabilities were analyzed using a Thurstonian model that characterizes ranks as samples from latent distributions with different means (M. D. Lee, Steyvers, & Miller, 2013; Yao & Böckenholt, 1999). The posterior latent means associated to each interpretation of probabilities are reported in Table 3. Overall, the interpretation of probabilities as relative frequencies was found to be the most adequate, although the considerable overlap observed (in particular among the few individuals adhering to ST) precludes any clear-cut conclusions. In any case, there is no indication that individuals committing to ST and DP hold very different interpretations of probabilities, such as a shift of the DP participants towards an interpretation in terms of plausibility.

Table 3. Latent Means of the Different Interpretations of Probability in Experiment 1

Interpretation	ST	DP
Plausibility	0	0
Frequency	-0.34 [-0.77, 0.10]	-0.47 [-0.83, -0.11]
Degrees of Belief	0.48 [0.03, 0.93]	0.59 [0.24, 0.96]
Propensity	-0.29 [-0.72, 0.13]	-0.12 [-0.48, 0.23]

Note. Lower values are associated with higher ranks (the top rank is 1). The mean of 'plausible' interpretation was fixed to zero without any loss of generality. Values inside the square brackets correspond to the 95% credibility intervals.

Phase 4 (Experiment 2)

We investigated whether there were any differences between the individuals classified as ST and DP based on their argumentative skills using our adaptation of Kuhn's (1991) task. To test the agreement of the classifications of argumentative skills by our three coders, the intraclass coefficient (ICC) was computed. A substantial agreement among the coders was found: ICC(2, 1) = .669 with 95% CI(.579, .739), $F(331, 662) = 8.105$, $p < .001$.

For the phase 1 classification of Experiment 2, the posterior probabilities associated to the occurrence of each single argumentative behavior are slightly higher for DP than ST. However, their respective 95% credibility intervals overlap. In order to pool the information quantified by each of these posterior probabilities, we will rely on the '*encompassing prior approach'* proposed by Klugkist and Hoijtink (2007) and Myung, Karabatsos, and Iverson (2008). According to this approach, the support for a given inequality (e.g., values in Condition 1 are larger than in Condition 2) provided by the data can be quantified by contrasting the probabilities that such inequalities are observed when taking samples from the prior and posterior distributions, respectively. In the present case, when we sample probabilities of observing the argumentative behaviors from their respective prior distributions, the probability that *all* sampled values from DP are larger than the sampled values from ST is only $.50^5 \approx .03$. When sampling from the posterior distributions, this probability is roughly .66. This difference suggests that individuals classified as adhering to DP manifesting more argumentative behaviors than their ST counterparts becomes roughly 21 times more likely in light of the data (when compared with a competing hypothesis that imposes no pattern whatsoever).

However, as Table 4 also shows, these differences in argumentative scores found for the phase 1 classification were not found in the phase 2 classifications, with the hypothesis of higher argumentative skills for DP adherents only becoming twice as likely in light of the data (i.e., there is only anecdotal evidence in support of the hypothesis).

Table 4. Probability of Argumentative Behaviors in Kuhn's (1991) Task (Experiment 2)

	ST_1	DP_1	ST_2	DP_2
Generate Alternative Theory	.78 [.66, .89]	.92 [.87, .95]	.86 [.77, .93]	.90 [.84, .94]
Recognizing Possibility of Own Error	.52 [.38, .66]	.71 [.64, .77]	.65 [.53, .76]	.69 [.62, .76]
Evaluate Underdetermined Evidence	.15 [.07, .26]	.23 [.17, .29]	.14 [.07, .23]	.20 [.14, .27]
Provide Genuine Evidence for Own Theory	.50 [.36, .65]	.69 [.62, .75]	.66 [.55, .77]	.67 [.60, 74]
Generate Possible Counterevidence	.40 [.27, .55]	.46 [.39, .53]	.49 [.38, .61]	.44 [.36, .52]

Note. Posterior probabilities and credibility intervals for the phase 1 classification (ST_1, DP_1) and phase 2 classification (ST_2, DP_2). The evidence variable was recoded such that it shows the median posterior probability that the indexed group succeeded in providing genuine evidence for their causal claim. The counterevidence variable was recoded such that it displays the median posterior probability that the indexed group succeeded in providing strong or weak possible counterevidence against their own theory. See the Supplementary Materials.

Demographics

In terms of demographics, we were interested in checking whether the individuals classified as adhering to ST and DP differed in terms of college education, and in terms of any previous training in probability theory. In the case of individuals classified as ST using phase1 responses in Experiments 1/2, the posterior probabilities of having college education and training in probability theory were .50 [.33, 67] / .62 [.48, .75] and .29 [.15, .45] / .34 [.21, .48], respectively. The analogous probabilities for adherents of DP were similar, .68 [.61, 75] / .73 [.66, .79] and .41[.34, .49] / .36 [.29, .43].

Discussion

In phase 4 in Experiment 1, it was found that the alternative hypothesis could not be supported by the results that the DP participants were following a defective interpretation of probabilities, which would make them more inclined to commit the conjunction fallacy.

Moreover, we did not find any systematic differences in whether the participants classified as following ST or DP had received probabilistic training. We therefore continue to interpret DP as representing a genuine inferential interpretation of the indicative conditional and as not just the result of erroneous probability assignments.

Finally, phase 4 of Experiment 2 also investigated the hypothesis that DP would possess stronger argumentative skills than ST, due to their increased focus on reason relations, using Kuhn's (1991) argumentation task, but found little to no support.

It is telling that we find the systematic differences that we do in the way participants classified as following ST or DP perform on the uncertain and-to-if inference task, in spite of the fact that these groups did not generally differ in their tendency to commit the conjunction fallacy (Experiment 1), nor in the degree to which they had received college education or probability training. Given the size of our samples, we should have been able to detect differences in these variables, if there were any of reasonable size. It therefore appears that the differences we tap into when investigating the opposition between ST and DP are orthogonal to the differences in these further variables.

Chapter 8:
Norm Conflicts and Epistemic Modals[101]

Niels Skovgaard-Olsen,
John Cantwell

Statements containing epistemic modals (e.g., “by winter 2022 most European countries may have the Covid-19 pandemic under control”) are common expressions of epistemic uncertainty. In this paper, previous published findings (Knobe & Yalcin, 2014; Khoo & Phillips, 2018) on the opposition between Contextualism and Relativism for epistemic modals are re-examined. It is found that these findings contain a substantial degree of individual variation. To investigate whether participants differ in their interpretation of epistemic modals, an experiment with multiple phases and sessions is used to classify participants according to the three semantic theories of Relativism, Contextualism, and Objectivism. Through this study, some of the first empirical evidence for the kind of truth-value shifts postulated by semantic Relativism is presented. It is furthermore found that participants’ disagreement judgments match their truth evaluations and that participants are capable of distinguishing between truth and justification. In a second experimental session, it is investigated whether participants thus classified follow the norm of retraction which Relativism uses to account for argumentation with epistemic modals. Here the results are less favorable for Relativism. In a second experiment, these results are replicated and the normative beliefs of participants concerning the norm of retraction are investigated following work on measuring norms by Bicchieri (2017). Again, it is found that on average participants show no strong preferences concerning the norm of retraction for epistemic modals. Yet, it

[101] This chapter is under review:

Skovgaard-Olsen, N. and John Cantwell (*in review, Journal of Experimental Psychology: Learning, Memory, and Cognition*). Norm Conflicts and Epistemic Modals.

Authors’ Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com). Supplementary Materials: https://osf.io/g6vym/

was found that participants who had committed to Objectivism and had training in logics applied the norm of retraction to might-statements. These results present a substantial challenge to the account of argumentation with epistemic modals presented in MacFarlane (2014), as discussed.

8.1 Introduction

Epistemic modals are a collection of linguistic expressions primarily used to express varying degrees of certainty, uncertainty or ignorance, which concern possibilities that are not excluded by what is known (Portner, 2009; Lassiter, 2017). Typical examples are might-modals, as in “in 2022 the status of Covid-19 might shift from pandemic to endemic”, must-modals, as in “from the rapidity of its spread, omicron must be more infectious than delta”, and various modalities for expressing likelihood or probability, as in “it is likely that omicron leads to less severe hospitalizations”. While their use is ubiquitous in all forms of discourse, they have proven something of a conundrum for theoreticians trying to account for their meaning, raising a host of fundamental issues relating to objectivity, subjectivity, truth-relativity, and context-dependence in discourse.

Traditionally, these issues have been investigated in linguistics and philosophy. Recently, psychologists have taken an interest in epistemic modality as well, with some arguing that epistemic possibilities should be made the foundation of psychology of reasoning through mental model theory (Hinterecker et al. 2016; Johnson-Laird & Ragni, 2019), and other psychologists taking a more critical perspective (Oaksford et al., 2019; Over, 2022).

Occasionally, these semantic issues make their appearance in the political discourse. A good example is when Dr. Fauci in March 2020 made the following statement about the use of the anti-malarial drug hydroxychloroquine for the treatment of Covid-19: “What I’m saying is that it might — it might be effective. I’m not saying that it isn’t. It might be effective.”[102] As of July 2020, Dr. Fauci concluded that all of the randomized, controlled clinical trials had consistently shown that hydroxychloroquine was not effective against Covid-19. The different semantic views reviewed below (Contextualism, Relativism, and Objectivism), differ in their truth evaluations and on whether this correction implies that the previously asserted might-statement should later be retracted, as illustrated in Table 1.

[102] www.washingtonpost.com

Table 1. Norm Conflict with Epistemic Modals

	Contextualism	Relativism	Objectivism
Was Dr. Fauci's statement true at 03.2020?	Yes	Yes	No
Is Dr. Fauci's statement true in 07.2020?	Yes	No	No
Should the might statement be retracted?	No	Yes	Yes
Is Dr. Fauci at fault for making the assertion?	No	No	No

Note. The table displays evaluations of a might statement as uttered at 03.2020.

Below we will elaborate on how these different judgments arise. For the moment, we just present them as an exhibit of how differences in semantic interpretation can affect the normative reactions to everyday events.

Through our experiments, we investigate whether the norm conflict highlighted by Table 1 maps onto individual variation in the semantic interpretations of epistemic modals. To do so, we make use of an experimental design developed in Skovgaard-Olsen et al. (2019) to study individual variation in case of norm conflicts in cognitive psychology, which will be introduced below. But first we introduce the semantic opposition between the three views. Next, we reanalyze previous published findings on epistemic modals and show that they are compatible with individual variation in participants' interpretations. Finally, we use these results to motivate our empirical studies.

Three Opposing Views

The basic problem can easily be appreciated. Different people have access to different evidence and so know, are uncertain, and ignorant of different things. So different people will be prone to describe one and the same situation using different – indeed apparently contradictory – epistemic modals. Consider a simple example. Bill and Doris have arrived in an art museum with three big exhibition halls, A, B, and C. They want to see the new Picasso, which happens to be in hall C, though Bill and Doris do not know this. Since their state of knowledge does not exclude any of the three possibilities, it is epistemically possible for them that the new Picasso is in hall A, possible that it is in hall B, and possible that it is in hall C. Bill says to Doris:

(1) The Picasso might be in hall A. (*warranted assertion*)

Meanwhile, Anne and Charlie are also at the museum. They have already searched room A for the new Picasso without finding it, and Anne, somewhat flustered, says to Charlie:

(2) The Picasso can't be in hall A; it must be in either B or C. (*counterclaim*)

On the surface, (1) seems to contradict (2); for, at the very least, the claim that the Picasso both might and can't be in hall A is a contradiction. If then Bill's and Anne's claims are jointly contradictory, then one of the claims is false. Now, if one of them is wrong it is presumably Bill; for he said that the Picasso might be in hall A even though it is in hall C (as we know but he doesn't).

The problem is that Bill's claim seems perfectly warranted. He knows that the museum has a Picasso even though he doesn't know where. Of course, given what he knows, Bill is not in a position to assert that the Picasso is in hall A, that would be an epistemic lapse on his part. But on any account, knowing nothing more than that the Picasso is in one of the halls, Bill should accept that the Picasso might be in hall A, just as he should accept that it might be in B or C. So, Bill seems to be doing nothing wrong – indeed seems to be doing exactly what he should be doing – yet allegedly by doing so asserts a falsity. That is puzzling; for it is odd that one can reach a falsity by drawing the conclusion that one should draw based on correct (though incomplete) information.

There are a variety of semantic theories that attempt to explain the apparent mystery of might-modals. In our experiments, we focus on Contextualism, Relativism, and Objectivism, which we introduce below. Following a standard approach in linguistics (Heim & Kratzer, 1998; Portner, 2009), these views are explicated by formulating truth conditions: states of affairs that make might-statements uttered at a context true or false. These semantic theories were introduced to capture specific norms of language use, which end up having widely different implications for argumentation with epistemic modals, as we shall see. In Appendix 1, the formal details of these views are reviewed, and we more carefully state how subtle differences in the semantic definitions are related to conflicting norms of language use.

Contextualism

Until recently the dominant explanatory framework in philosophy and linguistics was Contextualism (Hacking, 1967; Teller, 1972; Kratzer, 1977; DeRose, 1991). According to Contextualism, the truth conditions for assertions of sentences like (1) and (2) depend on what is known in the *context of use*. Accordingly, the content of an epistemically modal sentence varies with what is known in the context in which it is used. In Bill's context, his claim amounts to the claim that for all he (Bill) knows, the Picasso is in hall A. In Anne's context, her assertion amounts to the claim that she knows that the Picasso is not in hall A. These two claims do not contradict each other; indeed, they are both true: the puzzle has been removed. Note, by way of contrast, what would happen if the two couples had met, and Bill, addressing Anne, had asserted (1). It would seem perfectly legitimate for Anne to reply:

(3) No, not true, the Picasso can't be in hall A. We searched hall A thoroughly and it isn't there, it must be in either B or C. (*disagreement in joint context*)

Anne's seemingly reasonable response seems to be a flat denial of Bill's claim; so, it no longer makes sense to treat Bill's claim as a claim about what he knows. This can be explained by the shift in context. According to Contextualism, once Anne and Charlie have joined the context, Bill's assertion amounts to the claim that for all they (Bill, Doris, Anne and Charlie) jointly know, the Picasso is in hall A, and as Anne and Charlie know that it is not in hall A, Bill's claim is just not true. Whereas (1) and (2) did not contradict each other when asserted in different contexts, they do contradict each other when asserted in the same context.

By allowing flexibility in what counts as known in a context of use, Contextualism obtains considerable degrees of freedom for explaining both what can be plausibly asserted in a given context and how such assertions can be assessed by the audience in the context. Of course, critics are liable to see this flexibility as a weakness, since it can be interpreted as vagueness concerning what demarcates the context of use and how it changes (Egan & Weatherson, 2011; MacFarlane, 2014). Yet, even if we permit this flexibility, there remain cases that are hard for Contextualism to explain. For consider if Bill, having learned that Anne and Charlie are also looking for the Picasso, said the following:

(4) Have you checked hall A? The Picasso might be there. (*assertion without common knowledge*)

By his initial question Bill plausibly makes it clear that he does not exclude the possibility that Anne and Charlie have checked hall A and might know that the Picasso isn't there. So, Bill clearly doesn't intend his might-statement to be a claim about what they all jointly know. However, Anne's response (3) that flatly denies Bill's might-statement still seems to be perfectly legitimate. This presents a problem for Contextualism. If "The Picasso might be there" in Bill's mouth means that no one in the present company has excluded the possibility that the Picasso is in hall A, then Bill has no business in making the claim. However, if it instead means that he has not excluded the possibility that the Picasso is in hall A, then Anne has no business contradicting him; for he is merely making a claim about himself. Either way Contextualism has problems finding a contextually determined body of knowledge that at the same time makes both the initial assertion and the subsequent assessment plausible. Perceived problems with the contextualist account have opened for alternatives,[103] which is now an active research area in formal semantics and linguistics (see e.g., Egan & Weatherson, 2011) that is ripe for empirical investigation. Since Contextualism treats 'might-*p*' as a claim to the effect that *p* is consistent with what is known by a contextually determined group of people, for-all-we-know statements will feature in our experiments below.

Relativism

One influential alternative is assessment Relativism, or Relativism for short (Egan, Hawthorne, & Weatherson, 2005; Kölbel, 2003, 2015a, 2015b; Egan, 2007; Weatherson, 2009; MacFarlane 2011, 2014). While Contextualism holds that the content of a might claim depends on what is known in the *context of use*, Relativism denies this. Instead, the relativist holds that the truth value of a might-claim depends on what the person assessing the claim knows: truth must be relativised to the *context of assessment*, shifting the emphasis from the speaker to the 'listener' (assessor). For the speaker, of course, the context of use and the context of assessment is the same. So, when Bill asserts (1) his claim is true-for-him, but the same claim is false-for-Anne. Hence, Bill can properly make his claim, and Anne can properly deny it.

There are further examples where Relativism seems to do better than Contextualism in accounting for certain kinds of intuitions regarding plausible exchanges involving epistemic modals. Well-known examples are cases of *eavesdropping* and, importantly, *retractions*. Both are discussed further below and will feature centrally in our experiments. To specify the norm of retraction, MacFarlane (2014, p. 108) distinguishes between the context at which the original assertion was made, c_1, and the context at which the retraction takes place, c_2. The norm is then formulated as follows (where *p* is a placeholder for statements like "The Picasso might be in hall A").

> **Retraction Rule:** An Agent in context c_2 is required to retract an (unretracted) assertion of *p* made at c_1 if *p* is not true as used at c_1 and assessed from c_2.

Retracting is here not the same as admitting fault since the assertion may have been reasonable in the context in which it was made. Yet, since the assertion is not true as assessed from c_2 its conversational effects need to be undone, according to Relativism. To illustrate retractions: Upon hearing Anne's reply (3) to Bill's initial assertion (1), Bill will realise that

[103] There are, however, extensions of the contextualist paradigm that seek to address these issues in a broadly contextualist way. See, for instance, Kratzer's discussion in (2012, p. 100) and von Fintel and Gillies (2008, 2011).

the Picasso is not in hall A. He should then be willing to retract his earlier claim by conceding something like:

(5) Oh, then I guess I was wrong. (*retraction*)

For the relativist such a retraction makes sense, because while (1) was true-for-Bill when he made his assertion, it is no longer true-for-Bill after Bill has learned that the Picasso is not in hall A. By contrast, the contextualist faces difficulties explaining why such a retraction would be reasonable: Bill's initial claim was a true claim about what he knew at the time, so why retract? Relativism is attractive to the extent that it is better than Contextualism at explaining our intuitions in these cases. Whether these intuitions are shared by ordinary people will be examined in our experiments below.

However, relativising truth to a context of assessment opens a bundle of foundational problems having to do with the regulatory or normative role of truth in a discourse. For instance, we posed the problem of might-modals in terms of the apparent conflict involved in holding that Bill seems to be fully justified in believing and asserting (1) even though the resulting claim is false; based only on correct information he draws the conclusion that he should draw, but the conclusion is false. Relativism resolves this issue by letting (1) be true-for-Bill when it is asserted. So, from Bill's perspective he is saying something true. Yet, on the relativist account, it is false-for-Anne. Indeed, given that 'we' (the readers and writers of this paper) know that the Picasso is not in hall A, what Bill said is also false-for-us. Accordingly, Relativism introduces relative truth values in accounting for a fragment of natural language, which can shift with changes to information states. Whether this controversial theoretical innovation can be substantiated empirically remains to be seen.

Objectivism

Cantwell (forthcoming) has suggested an analysis that embraces the puzzle: Bill is fully justified, but wrong. With the idea that there is nothing inherently wrong about having a false might-belief or asserting a false might-modal. This might seem contradictory at first, but 'being wrong' is ambiguous in this context. It can mean that one has a belief or made an assertion that one shouldn't have, or it can simply mean that one has a belief or made an assertion that is false. The former is a normative claim, the latter is a purely descriptive claim about the semantic status of a belief or assertion.

By contrast, in normal factual discourse the normative and descriptive uses of 'being wrong' seldom come apart; for although false factual beliefs are often excusable, we expect people who realise that they have formed a false belief to make some adjustment to their belief-forming habits; at the very least not to make the same mistake again. However, with might-modals, things are different: the next time Bill finds himself uncertain about which of various scenarios obtain we expect him to conclude that each scenario might obtain, which, in effect, is to make the same 'mistake' again. If it is indeed not a mistake, then drawing an erroneous might-conclusion is different from making a factual mistake.

In Cantwell (forthcoming) this way of understanding epistemic modals is dubbed Objectivism. The account holds that there is an objective sense in which someone like Bill is fully justified in his might-judgment yet allows that there is an objective sense in which might claims have a truth value which does not vary with either the context of use or the context of assessment. Applied to the present example: as the Picasso is in hall C, there is an objective sense in which it can't be in hall A, and in this sense it is false that it might be in hall A. On this 'objectivist' analysis, Bill is fully justified in asserting either (1) or (4), and Anne is fully justified in responding with (3). Upon hearing Anne's response (and so learning that the Picasso is not in hall A), Bill is in a position to judge that his pervious assertion was false. To the extent to which (5) expresses this (and only this: no admission of wrongdoing), Bill is in a position to assert (5).

Reanalyzing Published Findings

Previously published findings on the opposition between Contextualism and Relativism gives a rather unclear picture with results not clearly favoring either theory. We will illustrate this point by considering the studies of Knobe and Yalcin (2014) and Khoo and Phillips (2018). Since the authors helpfully made their original data sets available, we will briefly reanalyze their results with an eye to individual variation. The papers perform statistical comparisons based on aggregated statistics like the following:

Table 2. Summary Statistics of Previously Published Findings

	Knobe & Yalcin (2014)			**Khoo & Phillips (2018)**	
	Exp2		**Exp3**	**Exp1**	
Condition	*True*	*False*	*Retract*	*Assessment*	*Utterance*
Factual	M = 2.03, SD = 1.83	M = 6.77, SD = 0.62	M = 6.53, SD = 0.94	M = 5.89, SD = 1.39	M = 6.03, SD = 1.27
Modal	M = 4.86, SD = 2.34	M = 3.19, SD = 2.31	M = 4.04, SD = 1.63	M = 4.65, SD = 2.14	M = 4.13, SD = 2.06
Indexical	//		//	M = 5.67, SD = 1.40	M = 2.64, SD = 1.97
DV	"statement is true/false?"		"should retract?"	"At least one of the claims must be false?"	
Scale	7-point Likert Scale from 1 ("completely disagree") to 7 ("completely agree")				

Note. In Knobe and Yalcin (2014): 'Modal' = might. In Khoo and Phillips (2018): 'Modal' = could; 'Assessment' = two speakers make conflicting assessments of assertion by third party; 'Utterance' = two speakers make conflicting utterances. 'DV' = dependent variable.

Table 2 displays data concerning whether the statement is true or false, whether the speaker should retract it after a change in information, and whether one of two opposing statements must be false on a Likert-scale from 1-7. In these studies, modal statements (might, could) are contrasted with factual statements and indexical statements (e.g., “I have had breakfast”) to obtain two contrasting baselines. For factual statements, the expectation in Knobe and Yalcin (2014) is that the statement is false and that it should be retracted, and the expectation in Khoo and Phillips (2018) is that participants will agree that one of two contrary statements must be false. For indexical statements, the expectation is that there is no real conflict between contrary statements, and that participants accordingly should disagree with the assessment that at least one of them must be false.

When looking at these data, it is striking that the means for the epistemic modals are at the midpoint of the scales and that epistemic modals have the largest standard deviations. In contrast, the means of the baseline categories are located at more extreme points on the 7 point Likert-scale. Both papers analyzed the results using statistics at the group level. Khoo and Phillips (2018) used a combination of ANOVA and t-tests; Knobe and Yalcin (2014) used ordinal regression. Refitting ordinal regression models with the factors Condition × Sentence, using the statistical programming language R (R Core Team, 2015) and the R package brms for Bayesian statistics (Bürkner, 2017), yielded the posterior predictive predictions shown in Figure 1 below.[104] Figure 1 confirms that there was a large degree of individual variation in these results for sentences containing epistemic modals, specifically. These patterns of individual variation were not examined in the respective papers, however.

[104] See the osf project page for R-scripts: https://osf.io/g6vym/.

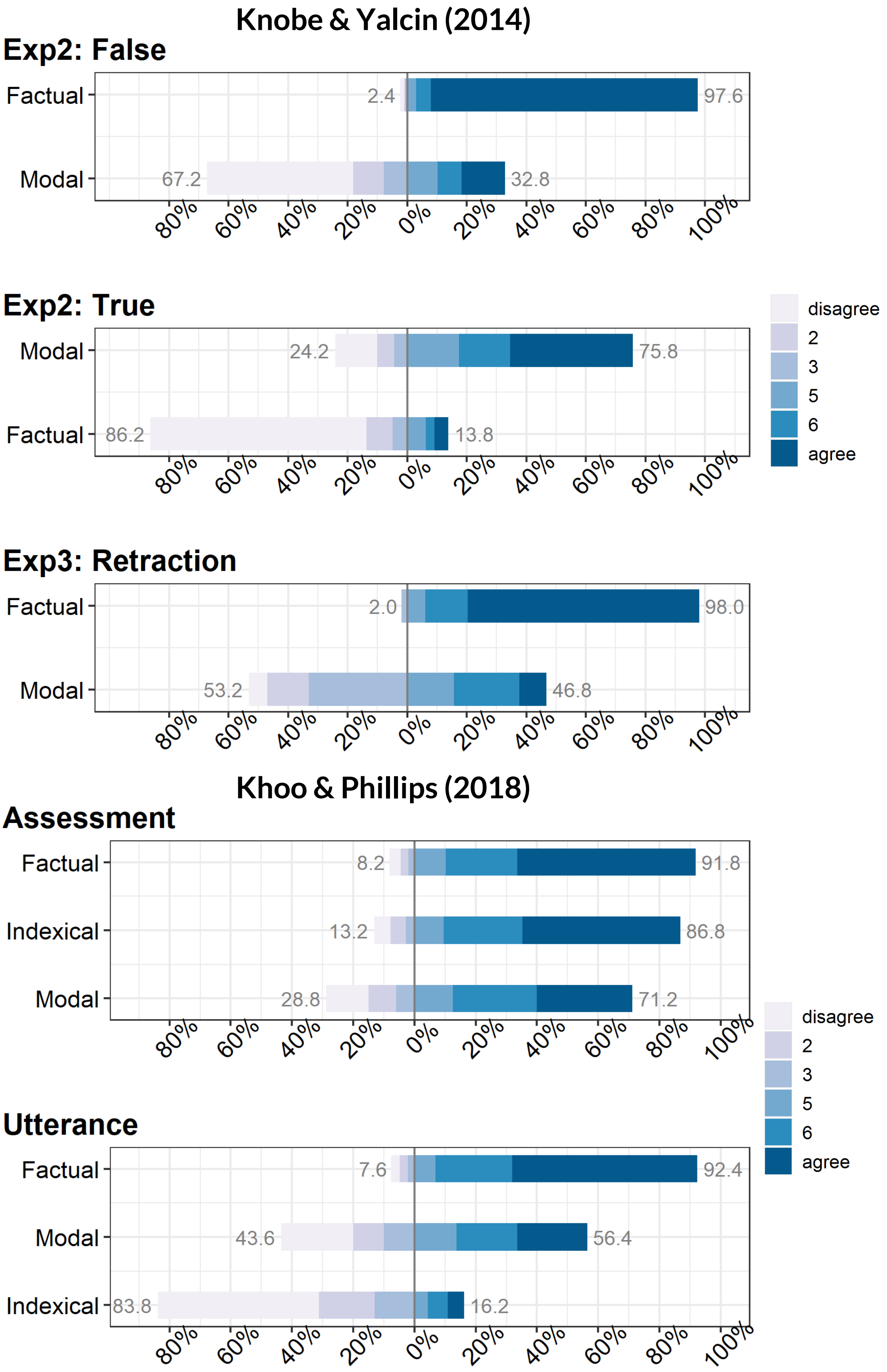


Figure 1. Sampling from the Posterior Predictive Distributions. Predictions based on 500 random samples from the posterior distributions, given the participants did not select the neutral '4' category. For Knobe & Yalcin (2014, Exp2): Condition = True vs. False; Sentence = Modal vs. Factual. For Khoo & Phillips (2018): Condition = Assessment vs. Utterance; Sentence = Factual vs. Modal vs. Indexical.

In these plots, values from 1-3 indicate disagreement with the presented statements and values from 5-7 represent agreement. As shown in Figure 1, in each case participants were divided between agreeing (5-7) and disagreeing (1-3) with the presented statements for epistemic modals. Applying statistics based on central tendencies to the data of Table 2 at the aggregate level is liable to misrepresent the distributions in the face of such split tendencies of agreeing and disagreeing. One of the purposes of our experiments is therefore to make a more targeted investigation of individual variation in the interpretation of epistemic modals. At issue is the following underlying assumption:

> (U) There is a uniform interpretation of expressions like epistemic modals. Only one of conflicting semantic theories can be descriptively adequate. If semantic theories like Contextualism, Relativism, and Objectivism are incompatible, then at most one of them can be descriptively correct.

Through our experiments, we will investigate whether (U) is correct as applied to epistemic modals. In doing so, we apply the individual profiling approach and Scorekeeping Task of Skovgaard-Olsen et al. (2019), which we review below. This leads to our first hypothesis.

> (H_1) There is individual variation in the interpretation of might-statements.

Both Khoo and Phillips (2018) and Cantwell (forthcoming) consider that a contributing factor to these mixed results is that participants may have had a difficulty separating truth from justification. On all accounts, the assertability of 'might *p*' is determined by the evidence available at the context of utterance. For participants conflating truth judgments with whether a statement is 'acceptable' or 'assertible', the judgment 'true' may be produced in apparent agreement with Contextualism, which only takes information accessible in the context of utterance into account. This conflation hypothesis is accordingly a further alternative hypothesis that we will test in the experiments that follow.

> (H_2) Participants conflate assessments of truth and justification.

Both Katz and Salerno (2017) and Khoo and Phillips (2018) argue that data on disagreement concerning epistemic modals is more fundamental than data concerning truth evaluations and retraction. Both studies thus use participants' judgments about whether two contrary statements can both be true as a measure of adherence to Relativism. Neither study found support for the central prediction of Relativism that the statements "might *p*" and "might not-

p” cannot both be true together, and thus constitute genuine cases of disagreement. But, on the other hand, Contextualism was not favoured either.[105]

8.1.1 The Present Studies

One goal of Experiment 1 was to directly probe whether evidence could be obtained for relativistic truth-values, which is the central innovation of MacFarlane (2014) to account for subjective language. As emphasized by Wright (2008: 180): “Correctness varies with context of assessment: that claim is the very heart and soul of truth-relativism. So that is what we need to see evidenced in linguistic practice if linguistic practice is to provide evidence for relativism”. At the same time, Wright (ibid.) emphasizes that it must be shown that these evaluations of correctness (or truth) dissociate from evaluations of justifications, as we have seen. Hence, evidence of shifts in truth values across different contexts of evaluations, while controlling for the difference between truth and justification, would constitute direct evidence for the relativistic interpretation of epistemic modals. This leads to our third hypothesis.

> (H_3) Shifts in truth-value judgments of might-statements occur across different contexts of evaluations, when controlling for the difference between truth and justification.

In Experiment 1, we investigated whether such truth-value shifts occur and made them the basis for our classification of participants as following Relativism. In addition, we investigated the type of disagreement data which Katz and Salerno (2017) and Khoo and Phillips (2018) take to be fundamental to the debate between Contextualism and Relativism.

Given the possibility of individual variation raised by our reanalysis above, a further objective of Experiment 1 was to probe whether the population could consistently be classified as a mixture distribution of diverging interpretations of epistemic modals. This leads to our fourth research hypothesis as a more precise version of (H_1).

> (H_4) Participants can be classified as following a mixture distribution of diverging interpretations of epistemic modals, where the profiles assigned based on truth evaluations are consistent with participants’ evaluations of disagreement.

[105] It is moreover likely that the Katz and Salerno (2017) study contains a similar degree of individual variation as the two other studies we have looked at above, which was not captured by the logistic regression analyses reported at the group level. In study 2, Katz and Salerno (2017) thus found that participants had mostly a ca. 50% chance of responding ‘1’ on a binary variable across conditions and items. The Katz and Salerno (2017) thus presents further motivation for investigating patterns of individual variation.

Finally, we tested whether participants thus classified were consistently following their assigned interpretation of epistemic modals when examining the type of retraction data emphasized in both MacFarlane (2014) and Knobe and Yalcin (2014) as being central to the opposition between Contextualism and Relativism.

Norm Conflict Experiments and the Problem of Arbitration

The research question into individual variation and mixture distributions of diverging interpretations of epistemic modals is motivated by a more general problem affecting the use of norms in empirical studies of rationality. Since many tasks have multiple norms that could be applied to them, Elqayam and Evans (2011) argue that the problem of arbitrating between competing norms is one of the main problems in cognitive psychology preventing the application of norms of judgment and decision making in experimental research. In contrast, the present experiments aim to show that by classifying participants according to different competence profiles based on divergent norms, the possibility of competing norms (like in Table 1) can be utilized for studying individual variation.

Through our experiments, we show how this problem of arbitration can be handled via Bayesian computational modelling and innovations in the experimental design, which were first employed in the context of conditionals and the psychology of reasoning in Skovgaard-Olsen et al. (2019). Table 3 outlines how this approach applies to Experiment 1.

Table 3. Individual-Profiling via Norm Conflicts

Experiment 1	Goal	Method
Session 1 *Phase 1*	Classify participants based on truth-value judgments according to Contextualism, Relativism, and Objectivism.	Bayesian mixture model.
Phase 2	Classify participants based on their reflective attitudes elicited by the Scorekeeping Task.	Bayesian latent class analysis.
Session 2 *Phase 3*	Probe whether participants follow the norm of retraction based on their individual profiles.	Bayesian mixed linear models.

Note. Details on the Bayesian mixture model and the latent class analysis can be found in Appendix 2. See our osf project page for R-scripts of all the analyses: https://osf.io/g6vym/.

The Bayesian mixture and latent class models are explained below and in Appendix 2. On this approach, participants are classified in Session 1 according to their truth-value judgments and reflective attitudes in situations of norm conflicts. Via mixture modelling and latent class analysis, it is probed whether individual profiles of the participants can be established that follow Relativism, Contextualism, and Objectivism, respectively.

Participants' reflective attitudes are elicited via the Scorekeeping Task, where participants are put in the position of a scorekeeper. The scorekeeper has to assess the performance of fictive participants, who have produced incompatible responses to the task that the participants have just completed. Through this task, participants commit to an interpretation of epistemic modals by criticizing and sanctioning their fictive peers based on their mutual criticism. A comparison is then made between participants' reflective attitudes with their own case judgments to investigate how well they match.

In session 2, it was probed whether participants consistently followed the assigned profiles to apply the associated norms in a novel task. For the present case of norm conflicts with epistemic modals, this involves investigating whether participants follow the norm of retraction. Participants classified according to Objectivism and Relativism are predicted to enforce this norm, whereas participants following Contextualism are predicted to flout it.

8.2 Experiment 1: Individual Profiling

8.2.1 Session 1

The goal of Session 1 was to analyze participants' truth value judgments as coming from a mixture distribution of truth assignments by Relativism, Objectivism, and Contextualism with the purpose of classifying participants accordingly. In a second step, the participants were presented with the Scorekeeping Task to establish latent classes in a second classification. In a third and final step, the match of the two classifications was probed. For both classifications, it holds that the classifications were mutually exclusive, which means that a participant could at most be assigned to one mixture group (phase 1) and one latent class (phase 2), respectively. But the two assignments need not be the same.

8.2.1.1 Method

Participants

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 780 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk from the USA, UK, Canada, and Australia. They were paid a small amount of money for their participation (on average 6$ per hour) and told that there would be in addition be a 1$ bonus, if they answered accurately and participated in the second half one week later. The following *a priori* exclusion criteria were used: not having English as native language, completing the task in more than 2 standard deviations below or above the mean completion time, failing to answer at least one of two simple SAT comprehension questions correctly in a warm-up phase, and answering ‘not seriously at all’ to the question 'How seriously do you take your participation' at the beginning of the study. Since some of these exclusion criteria were overlapping, the final sample for session 1 consisted of 540 participants. Mean age was 39.42 years, ranging from 18 to 78. 48.15 % of the participants self-identified as male, 51.67% self-identified as female, and one person preferred not to identify with either gender. 79.44 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. Applying the exclusion criteria had a minimal effect on the demographic variables.

Design

Phase 1 of the experiment had a within-subject design with two factors: Sentence (with three levels: factual vs. might vs. know) and type of dependent variable, DV (with four levels explained below: justified vs. true$_{t1}$ vs. true$_{t2}$ vs. true$_{both}$). Since all four dependent variables were presented on every trial, and we wanted four trial replications for each cell of the design, each participant in total went through 48 within-subject conditions.

Procedure Shared by Session 1 and 2

The four trial replications of the three Sentence within-subject conditions were randomly assigned to 12 different scenarios. Random assignment was performed without replacement such that each participant saw a different scenario for each condition. This ensured that the mapping of condition to scenario was counterbalanced across participants preventing confounds of condition and content. To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002). Participants were then asked to supply their Mechanical Worker ID so that their responses in Session 1 could be matched with Session 2, one week later.

Materials and Procedure Specific to Session 1

Phase 1: The Eavesdropper Task

The list of the 12 scenarios used can be found on the *osf* page.[106] The scenarios used were inspired by the stimulus materials of Knobe and Yalcin (2014) and Katz and Salerno (2017), but further were created to increase the number of items. Since the task involved time indices printed in writing and illustrated on analog clocks, participants first saw three practice

[106] https://osf.io/g6vym/?view_only=9425e07e499949b9971baffd1c5519a5

items to ensure that they had understood the setup. In addition, participants were debriefed after the practice items to make sure that they paid careful attention to such distinctions as whether a statement (marked in blue) was true before learning about an additional fact and whether the statement was true now, after the additional fact had been learned.

In the first phase, participants were presented with the 12 scenarios in random order, displayed in the format illustrated below. These scenarios concern so-called "eavesdropper" cases, where a person, who is not taking part in a conversation, possesses additional factual information, which can be used to evaluate the assertions of the speakers. In this task, participants are first presented with the information used as the basis for the assertions by the speakers (at t_1), and then there is a continuation, where the factual information of the eavesdropper is added (at t_2), as illustrated in Table 4 below.

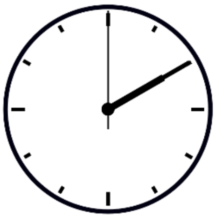
Table 4. Temporal Format of the Eavesdropper Task

Statement (S)	Continuation (Fact)
p / *might p /* *for all we know, p*	Not *p*
14:10 (t_1)	14:12 (t_2)
Is the statement justified?	*Was the statement true at 14:10?* *Is the statement true now at 14:12?* *Is it possible for both the statement and the continuation to be true at the same time?*

This temporal format was used to manipulate changes in information states, which figures centrally in the semantic opposition between Contextualism, Relativism, and Objectivism (Egan & Weatherson, 2011).

To emphasize certain details of the scenario, some phrases were highlighted in blue and red (as indicated below). Here is an example of a scenario:

> Heth and Jed are at the coffee shop together waiting for Fred, who will be reciting poetry later that night. A man is approaching, but Jed has poor eyesight and can only see the bare outlines of the man and says:
>
> **14:10:** "That person MIGHT be Fred."
>
> [/"FOR ALL WE KNOW, that person is Fred"]
>
> [/"That person is Fred"]



Participants were then asked four questions in the following order. First, they were asked whether they agreed/disagreed with the statement that "The speaker is justified in making the blue assertion". Next, the time displayed on the analog clock shifted by two minutes, and they were presented with the following continuation of the scenario for the remaining three questions:

> **Continuation (2 min later):**
> Cindy works for security at the coffee shop and is watching the room by video camera from a remote location. Her camera has a close-up of the approaching man, and while eavesdropping on Heth and Jed's conversation Cindy says to herself:
> **14:12**: "That person can't be Fred. I know Fred has a scar on his face."

Participants were then asked to indicate whether they agreed/disagreed with the following statements after being presented with the continuation:

> "Before learning about the continuation, the blue statement was true at 14:10".
> "After learning about the continuation, the blue statement made at 14:10 is true now at 14:12."
> "It is possible in this case for both the claim marked in blue and the statement made in the continuation to be true (at the same time)."

Each of the four DVs was presented as indicator variables ("yes" vs. "no") to the participants.

Phase 2: The Scorekeeping Task

After completing this task for the 12 different scenarios, participants were presented with the Scorekeeping Task following Skovgaard-Olsen et al. (2019). In the Scorekeeping Task, participants were told that when given the task that they had just completed, John, Robert, and Simon responded very differently. Participants were then presented with one of the scenarios that they had responded to in the Might condition. After having seen the first part of the scenario *before* the continuation, participants could click "Continue" to see the next part *after* the continuation. By further "Continue" clicks, participants could elicit the three responses by John, Robert, and Simon in random order, as shown in Table 5.

Table 5. The Scorekeeping Task, the Three Conflicting Responses

	John responded: At 14:10 the blue statement was true. After learning about the continuation at 14:12 the blue statement is no longer true.
	Robert responded: At 14:10 the blue statement was true. After learning about the continuation at 14:12 the blue statement is still true.
	Simon responded: At 14:10 the blue statement was not true. After learning about the continuation at 14:12 the blue statement is still not true.

Note. The response of John represents Relativism, the response of Robert represents Contextualism, and the response of Simon represents Objectivism. As above, various phrases were highlighted in red and blue to make the processing easier for participants.

Participants were then instructed that "Robert, John, and Simon cannot all be right!" and that they would see each of their responses repeated along with a criticism by the two other persons. Participants were told that their task was to decide based on these mutual criticisms whose response was the most adequate and whose "HIT" should be approved.

On Mechanical Turk, tasks are described as 'HITs' (Human Intelligence Tasks) to participants. Since participants are financially rewarded by approvals of HITs, and build up a reputation on this basis, the approval of HITs was used as an ecologically valid sanctioning measure in Skovgaard-Olsen et al. (2019) that participants are motivated to reason about. Here we adopt this measure as well.

The three criticisms were presented in random order on three consecutive pages. On each page, participants were first presented with the scenario and the response being criticized for repetition. Next, they were presented with the corresponding criticism of the response, by the two other parties, as illustrated in Table 6 below, and asked whether they found the given criticism compelling ("yes" vs. "no"). As above, each bit of information on the page was elicited by the participant by pressing "Continue". Screenshots of the pages can be obtained at the *osf* project page.[107]

[107] https://osf.io/g6vym/

Table 6. The Scorekeeping Task, the Mutual Criticisms

 	Robert and Simon's criticism of John: "How can you both say that the blue statement was true at 14:10 and is false at 14:12? That makes no sense!"
 	John and Simon's criticism of Robert: "How can you say that the blue statement is true at 14:12 after the continuation has become known? That makes no sense!"
 	John and Robert's criticism of Simon: "How can you say that the blue statement was false at 14:10 before the continuation became known? That makes no sense!"

Note. The response of John represents Relativism, the response of Robert represents Contextualism, and the response of Simon represents Objectivism. As above, various phrases were highlighted in red and blue to make the processing easier for the participants.

Finally, participants were presented with a page in which all three responses were repeated with the instruction that they should indicate whose "HIT" on Mechanical Turk should be approved after having seen John, Robert, and Simon's mutual criticism. Out of the following list displayed in random order, participants could select one option for approval: 1) Simon's HIT, 2) John's HIT, 3) Robert's HIT. Participants were then asked a few demographic questions.

8.2.1.2 Results and Discussion

Phase 1 Judgments

Table 7 displays some initial descriptive statistics, which report the central tendencies of the four dependent variables across the three types of sentences at the aggregate level.

Table 7. Descriptive Statistics of Phase 1

	Justified	True at t_1	True at t_2	Both True
Factual	M = 0.75 (0.43)	M = 0.39 (0.49)	M = 0.12 (0.33)	M = 0.17 (0.37)
Might	M = 0.91 (0.28)	M = 0.71 (0.45)	M = 0.21 (0.4)	M = 0.34 (0.47)
For-all-we-know	M = 0.89 (0.31)	M = 0.71 (0.45)	M = 0.27 (0.44)	M = 0.41 (0.49)

Note. Standard deviations are reported in paratheses.

To investigate whether these data conceal individual variation (H_1), the analysis below models the data as coming from a mixture distribution with three mixture components (H_4).

Bayesian Mixture Model

To classify participants into three mixture groups, the prior recommendations and Bayesian mixture models in Lee and Wagenmarkers (2014) were followed using the R-package `rjags` (Plummer, 2019).[108] On this approach, information or ignorance regarding the model parameters is represented by *prior distributions*. The observed data is then used to update our knowledge about the parameters, resulting in *posterior parameter distributions* (Kruschke, 2014; Lee & Wagenmakers, 2014; Skovgaard-Olsen et al. 2019; Skovgaard-Olsen 2019). A Gibbs sampler is used to estimate these posterior distributions by means of Monte Carlo-Markov chains. The general principle is that the posterior distribution of the model parameters is their prior probability times the data likelihood (Li et al., 2018). Accordingly, a Bayesian model is specified in terms of a likelihood function of the data and prior distributions of the parameters. In the mixture model below, the likelihood function for participants' responses to the binary dependent variables in phase 1 was modelled by a Binomial distribution.

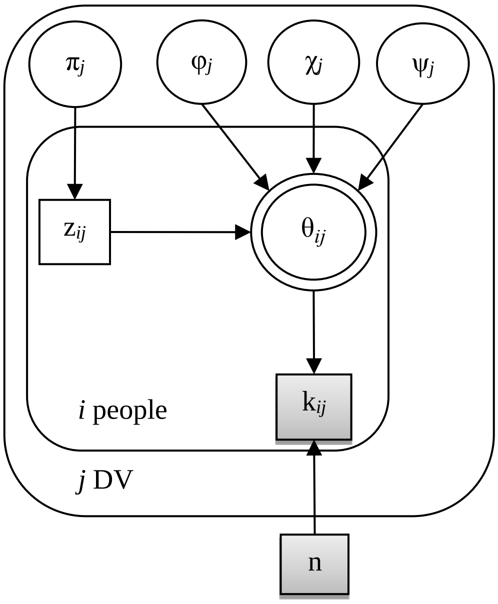
In phase 1, participants were classified as following Contextualism, Relativism, and Objectivism based on their truth-value judgments at two time points (t_1, t_2). For the experimental task illustrated above, t_1 is 14:10 on the analog clock and t_2 is 14:12, after the additional information ($\neg p$) of the eavesdropper has been revealed (see Table 4).

For this classification, it was assumed that participants' binary responses came from a mixture distribution consisting of three mixture groups: participants who had a preference to say 'yes' (with a binominal rate parameter, $\varphi_{j,}$ such that $\varphi_j \in [0.6, 1.0]$), participants who were responding at chance (with a binominal rate parameter, $\chi_{j,}$ such that $\chi_j \in [0.4, 0.6]$), participants who had a preference against saying 'yes' (with a binominal rate parameter, $\psi_{j,}$ such that $\psi_j \in [0, 0.4]$). These mixture groups were in turn used to classify participants as following three interpretations of epistemic modals (Contextualism, Relativism, and Objectivism) based on their truth evaluations of might-sentences.

Of the 540 participants who took part in the first session, 433 (80.19%) participants could be classified on this basis as following either Contextualism, Relativism, or Objectivism by applying the mixture model in Table 8 (*upper half*) and the classification of interpretations of epistemic modals based on it (*lower half*).

[108] In addition, the following extension of Lee and Wagenmarkers' (2014) model in Chapter 6 to three latent classes was used: https://uol.de/en/lcs/probabilistic-programming/webchurch-and-openbugs/bcm-ch061-latent-class-model-exam-scores.

Table 8. Bayesian Mixture Model



π_j ~ Dirichlet(1, 1, 1)

z_{ij} ~ Categorical(π_j)

φ_j ~ beta(1, 1)T(0.6, 1)

χ_j ~ beta(1, 1)T(0.4, .6)

ψ_j ~ beta(1, 1)T(0, 0.4)

$$\theta_{ij} \leftarrow \begin{cases} \varphi_j & \mathit{if}\ z_{ij} = 3 \\ \chi_j & \mathit{if}\ z_{ij} = 2 \\ \psi_j & \mathit{if}\ z_{ij} = 1 \end{cases}$$

k_{ij} ~ Binominal(θ_{ij}, n)

		S = "Might p"	
	Contextualism	**Relativism**	**Objectivism**
DV$_1$: Was S true at t_1?	P("yes") > 0.6	P("yes") > 0.6	P("yes") < 0.4
DV$_2$: Is S true after learning that ¬p at t_2?	P("yes") > 0.6	P("yes") < 0.4	P("yes") < 0.4

Note. beta(1,1)T(0.6, 1) indicates that the beta-distribution with the shape-parameters $\alpha = 1$ and $\beta = 1$ is truncated to only take values from the interval [0.6, 1]. DV ∈ {"was S true at t_1?", "is S true at t_2 after learning ¬p?", "was the speaker justified at t_1 in asserting S?", "can both S and ¬p be true at t_2?"}, see Table 9. As shown in the lower half, a participant was classified as adhering to Contextualism, Relativism, and Objectivism based on the binominal rate parameters that were estimated based on their responses to DV$_1$ and DV$_2$. E.g., if subject$_1$ was estimated to have a binominal rate parameter > 0.6 for both DV$_1$ and DV$_2$ based on the four trial replications, then subject$_1$ was classified as adhering to Contextualism.

The priors for the binominal rate parameters were given by three beta distributions (one for each mixture component). A categorical variable, z_{ij}, with a non-informative Dirichlet prior, determined which mixture group participant *i* was assigned to. Based on the estimated value of the categorical variables z_{ij}, every participant was classified as following the truth evaluation of Contextualism, Relativism, or Objectivism for epistemic modals, according to the predictions in Table 9.

In addition to truth-value judgments of whether an uttered statement was true before and after learning an additional fact, participants also provided responses to three further dependent variables that were inspired by previous studies. Table 9 provides an overview of all these dependent variables and outlines the contrasting predictions of Contextualism, Relativism, and Objectivism based on these measures.

Table 9. Classification and Predictions

Statement (S) made at t_1:	**Factual *p***			**Might *p***			**For all we know *p***		
	C	**R**	**O**	**C**	**R**	**O**	**C**	**R**	**O**
DV$_{\text{True Before}}$: Was S true at t_1?	0	0	0	1	1	0	1	1	1
DV$_{\text{True After}}$: Is S true after learning that ¬p at t_2?	0	0	0	1	0	0	1	1	1
DV$_{\text{Justified}}$: Was the speaker justified in asserting S at t_1?	1	1	1	1	1	1	1	1	1
DV$_{\text{Both True}}$: Can both S and ¬p be true at t_2?	0	0	0	1	0	0	1	1	1

Note: S ∈ {factual *p*, might *p*, for-all-we-know *p*} 'C' = Contextualism; 'R' = Relativism; 'O' = Objectivism. 't_1' = time of utterance of S. 't_2' = time of evaluation of evaluation of the statement, S, after learning that ¬*p*.

Since there were four dependent variables for each type of Sentence (might *p*, for-all-we-know *p,* and factual *p*), 12 binominal rate parameters were estimated for a given participant based on four trial replications with unique scenarios.

In Knobe and Yalcin (2014), the focus was on truth value judgments at t_2 (in addition to retraction judgments, which we will return to in Session 2). In Khoo and Phillips (2018), the focus was on incompatibility judgments at t_2. Table 9 integrates both these judgments and additionally makes a classification based on possible shifts in truth value judgments by measuring truth values at both t_1 and t_2 to test (H_3).

Furthermore, Table 9 controls for the possibility of conflating truth and justification by measuring both to probe (H_2), the conflation hypothesis (Khoo & Phillips 2018; Cantwell, forthcoming). Finally, Table 9 includes a subjective ("for all we know *p*") and an objective ("factual *p*") baseline to compare might-statements with. As can be seen from Table 9, the predictions for these baselines remain invariant across the three types of interpretations of might-statements. We here follow the semantic treatment of "for all *α* knows *p*" given in MacFarlane (2014, p. 265), according to which the statement is true iff *p* is not excluded by what *α* knows at t_1 at the context of utterance. According to this definition, further information acquired at t_2 in the eavesdropper cases should not have an impact of the truth value of for-all-we-know statements uttered at t_1, since the statement restricts its scope to what was known back then. On this basis, the following hypothesis can be formulated.

> (H_5) Only participants classified as following Contextualism treat might-statements and for-all-we-know statements alike.

Moreover, the justification of might-statements is also non-diagnostic w.r.t. these three interpretations. Rather, the distinguishing features concern: 1) whether both the might-statement (S) and the revealed fact in the continuation (¬*p*) can both be true at the same time, and 2) whether the might-statement is true at the two time points (t_1 before ¬p was revealed, and t_2 after ¬*p* was revealed). Only Contextualism permits might-statements and ¬*p* to be true at the same time. Only Contextualism allows might-statements to remain true at t_2 after ¬*p* has been revealed. In contrast, Relativism posits as its distinguishing feature that a shift in truth values of might-statements occurs between t_1 and t_2. The distinguishing feature of Objectivism is to deny that the might-statements were true at t_1 before ¬*p* was revealed. These predictions permit us to test (H_4) that participants can be characterized by a mixture distribution of different interpretations of might-statements.

Testing (H_4) by Analysing the Estimated Parameters

The identified mixture groups differ on the binominal rate parameters by which participants produce 'yes' responses in the four trial replications for each of the four DV in Table 9. To analyze these rate parameters, four linear mixed effects models were fitted via the R-package `brms` (Bürkner, 2017). As random effects, the models controlled for differences in intercepts and slopes for the effect of the Sentence and DV factors across participants in the same way but differed in their fixed effects as follows:

(M1) a maximal model that treats the estimated rate parameters as a function of the Sentence factor (factual vs. might vs. for-all-we-know), the Group factor (Relativism vs. Contextualism vs. Objectivism vs. Unclassified), the DV factor (true before vs. true after vs. justified vs. both true) and their three and two-way interactions.

(M2) a model that is obtained from the maximal model (M1) by removing the three-way interaction.

(M3) a model that is obtained from (M2) by removing the two-way Sentence:Group interaction.

(M4) a model that is obtained from (M3) by removing the two-way Group:DV interaction.

This cluster of models was selected to systematically test the hypothesis that the type of semantic interpretation of might-statements (encoded via the group-factor) had targeted influences on the type of dependent variable participants were presented with (encoded via the DV factor) and the type of sentence they were presented with. As such, interactions between the group factors and the DV and sentence factors indicate qualitative differences in the interpretations of these sentences, as a function of assigned competence profile. The maximal extent of these individual differences would be found by support of a three-way interaction, which was included in (M1). The further models strip this maximal model of interactions with the group factor to test whether models that include less qualitative differences could outperform (M1).

Quantifying Strength of Evidence

Hypotheses concerning the presence/absence of effects are tested here and below via Bayes factors. In this way, evidence in favour of, e.g., the H_0 that there is no difference in posterior medians between two types of sentences can be quantified in terms of Bayes factors, where classical significance testing would only have permitted us to conclude that H_0 could not be rejected (Wagenmakers et al., 2018). To be able to quantify the strength of evidence both against and in favour of H_0, we rely on the following qualitative interpretation of Bayes factors (Lee & Wagenmarkers, 2014): (Anecdotal evidence for H_0) $\frac{1}{3}$ < BF_{H1H0} < 1, (Moderate evidence for H_0) $\frac{1}{10}$ < BF_{H1H0} < $\frac{1}{3}$, (Strong evidence for H_0) $\frac{1}{30}$ < BF_{H1H0} < $\frac{1}{10}$, (Very Strong evidence for H_0) $\frac{1}{100}$ < BF_{H1H0} < $\frac{1}{30}$, (Extreme evidence for H_0) BF_{H1H0} < $\frac{1}{100}$. Values above 1 indicative evidence in favour of H_1 since this scale is mirrored by applying the following ratio: $BF_{H1H0} = \frac{1}{BF_{H0H1}}$.

The performance of the statistical models is quantified by the leave-one-out cross validation criterion. To estimate the out-of-sample predictive adequacy of the models, the data is divided into subsets called “folds”, and the models are trained on a subset of the data and tested for its ability to predict observations left out (Vehati et al., 2017, 2019; McElreath, 2020). Below, we report both the leave-one-out information criterion (LOOIC), where each fold leaves out one observation, resulting in the maximum number of folds, and an information criterion (kfoldic), which is based on a 10-fold cross-validation that refits the model by dividing the data into 10 subsets.[109]

8.2.1.3 The Results

Lower LOOIC values indicate a better fit in the light of the parsimony vs. fit trade-off.

[109] To make the estimation of LOOIC computationally feasible, Pareto-smoothed importance sampling is used as an approximation. The reliability of this approximation is assessed via Pareto-k values, which should ideally be below 0.7 (Vehati et al., 2017, 2019; McElreath, 2020). Since some data points had Pareto-k values above 0.7, the 10-fold cross-validation information criterion was included in Table 10 as well.

Table 10. Model Comparison

	LOOIC	Δelpd	SE	kfoldic	Weight
M1	-3552.3	0	--	-3410.6	0.74
M2	-3180.5	-185.9	21.1	-3114.9	0.14
M3	-3099.9	-226.2	23.4	-3037.8	0.03
M4	-2678.0	-437.2	32.8	-2578.2	0.10

Note. 'elpd' = expected log predictive density. elpd is a measure of out-of-sample predictive adequacy. LOOIC = -2*elpd. The weights are stacking weights based on kfoldic.

The cross-validation criterion in Table 10 clearly favoured M1 indicating evidence for the three-way interaction and group differences in the DV type and sentence examined. Figure 2 presents the estimated posterior probabilities of answering 'yes' for the four DVs across the three mixture groups (Objectivism, Relativism, and Contextualism) based on M1. As outlined in Table 8, participants' truth evaluations for might-statements at t_1 and t_2 were used to define group membership. So, minimally these should differ on the classification. The further qualitative differences in Figure 2 indicate that the classifications thus formed are predictive of other judgments of disagreement in support of (H_4).

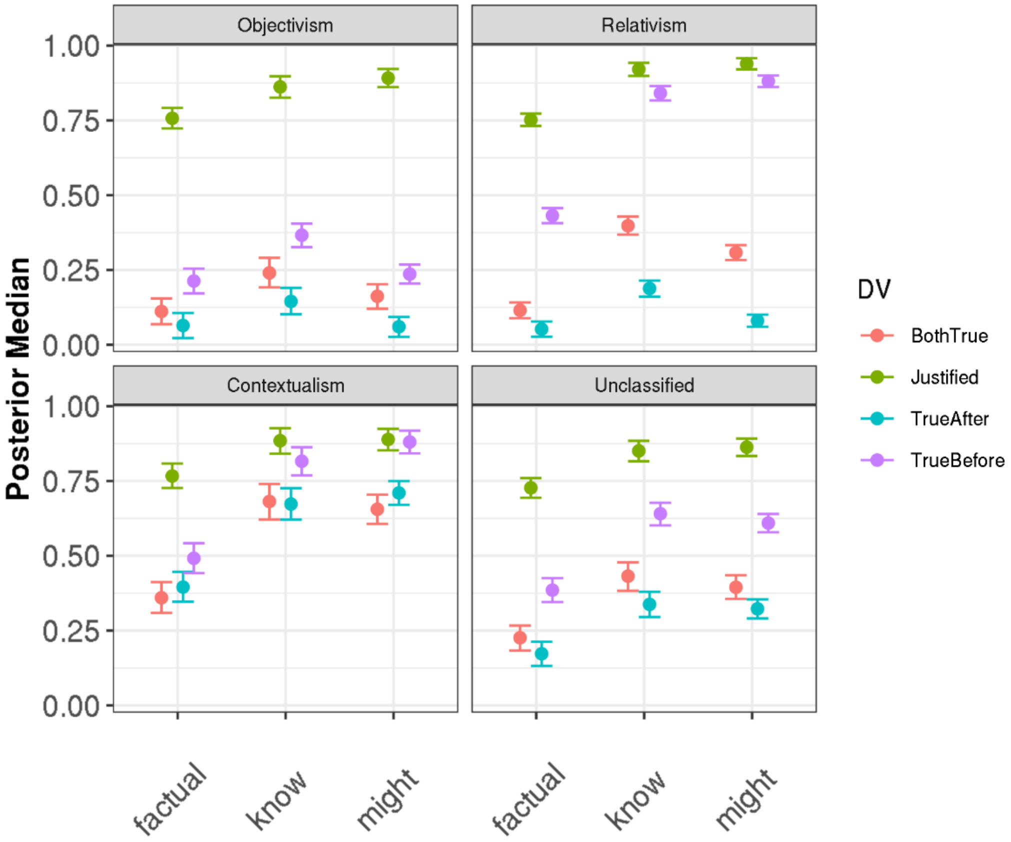


Figure 2. Phase 1 Classification. Parameter estimates of posterior probabilities of answering "yes" for the four DVs in Session 1 for the three mixture groups. In total, based on the mixture model in Appendix 2, 540 participants were classified into four groups: Contextualism (69), Objectivism (98), Relativism (266), and Unclassified (107). 'DV' = dependent variable.

Figure 2 shows that the distinctive patterns of Contextualism, Relativism, and Objectivism concerning might-statements outlined in Table 9 are preserved in the estimated posterior median probabilities of answering "yes" for participants classified as belonging to each of these three mixture groups.[110]

The posterior medians of the estimated parameters were further analyzed via the R-packages `emmeans` (Lenth, 2020) and `bayestestR` (Makowski et al., 2019). Focusing on might-statements, the strong signature differences in whether might statements were true before the fact not-*p* became known (Contextualism, $\tilde{x}$ = .88, 95% HPD[111] [.84, .92]; Relativism, $\tilde{x}$ = .88, 95% HPD [.86, .90], Objectivism, $\tilde{x}$ = .24, 95% HPD [.20, .27]) and after (Contextualism, $\tilde{x}$ = .71, 95% HPD [.67, .75], Relativism, $\tilde{x}$ = .08, 95% HPD [.06, .10], Objectivism, $\tilde{x}$ = .06, 95% HPD [.03, .09]) were used as the basis for the classifications. For Relativism, a stark shift in truth values of might-statements was thus observed in support of (H_3), which spanned most of the probability scale ($\tilde{x}$ = -.80, 95% HPD [-.83, -.77]).

Based on this classification, the research hypothesis that relativists, contextualists, and objectivists differ on whether two contrary might-statements can both be true was tested. It was found that contextualists had a higher posterior median probability of accepting both of the contrary might-statements as true than relativists ($b_{\text{Relativism - Contextualism}}$ = -.35, 95% CI [-.40, -.29], BF_{H1H0} > 100) and objectivists ($b_{\text{Objectivism - Contextualism}}$ = -.49, 95% CI [-.56, -.43], BF_{H1H0} > 100).

To test the alternative hypothesis that participants conflate truth and justification, we focus on the factual statements as the objective baseline and compare the contrast between $DV_{\text{justified}}$ and $DV_{\text{True Before}}$ across the three mixture groups. It was found for Contextualism (b = .28, 95% CI [.22, .33], BF_{H1H0} > 100), Relativism (b = .32, 95% CI [.29, .35], BF_{H1H0} > 100), and Objectivism (b = .54, 95% CI [.50, .59], BF_{H1H0} > 100) that the posterior probability of factual statements being justified before not-p was learned was higher than the probability of factual statements being true. The results thus allow us to reject the hypothesis, (H_2), that participants conflate these two types of evaluation (Khoo & Phillips, 2018; Cantwell, forthcoming).

[110] The unclassified participants exhibit a pattern that resembles Relativism, but which sits in between the other classifications.

[111] As a point estimate, the median of the posterior distribution is used here. The 95% HPD interval indicates an 95% interval in which all points have a higher probability density than points outside the interval. For the contrasts below, we report the 95 % credible interval, which has the interpretation that there is a 95% probability that the effect falls within the displayed range, given the data.

According to the predictions in Table 9, a distinguishing feature of Contextualism is to treat might-statements and for-all-we-know statements alike – as both expressing contextual features of the present epistemic state. In contrast, it was found across all three mixture groups that participants followed the trend of treating for-all-we-know statements in a qualitatively similar way to might-statements, in spite of the stark differences in the interpretations of the latter. This finding is surprising and contradicts (H_5) – moreover, it was consistent over several pilot studies. It may indicate that participants could not detect a difference in meaning between might-statements and for-all-we-know statements in a within-subject comparison, although participants across mixture groups displayed sharply differing responses patterns concerning both types of statements.

8.2.1.4 Phase 2: The Scorekeeping Task

Table 11 displays descriptive statistics of the central tendencies at the aggregate level.

Table 11. Descriptive Statistics for Phase 2

	Criticism of	HIT approval of
Relativism	M = 0.30 (0.46)	M = 0.58 (0.49)
Contextualism	M = 0.66 (0.47)	M = 0.23 (0.42)
Objectivism	M = 0.67 (0.47)	M = 0.18 (0.39)

Note. Standard deviations are reported in paratheses.

To investigate whether these data conceal individual variation as a function of the phase 1 classification, the analysis below applies a latent class analysis to identify three latent classes.

Latent Class Analysis

A latent class analysis (Mair, 2018) was applied to participants' responses in the Scorekeeping task. Following Li et al. (2018), the latent class analysis was implemented in a Bayesian framework by sampling from the posterior distribution via a Gibbs sampler. This analysis follows a similar pattern as the mixture model in phase 1, but applies a different model, which is illustrated in Appendix 2 and explained in Li et al. (2018).

A solution with three latent classes was fitted to the data and posterior item response probabilities for the different items in the Scorekeeping Task were estimated for each of the latent classes along with the frequency of the classes in the population.

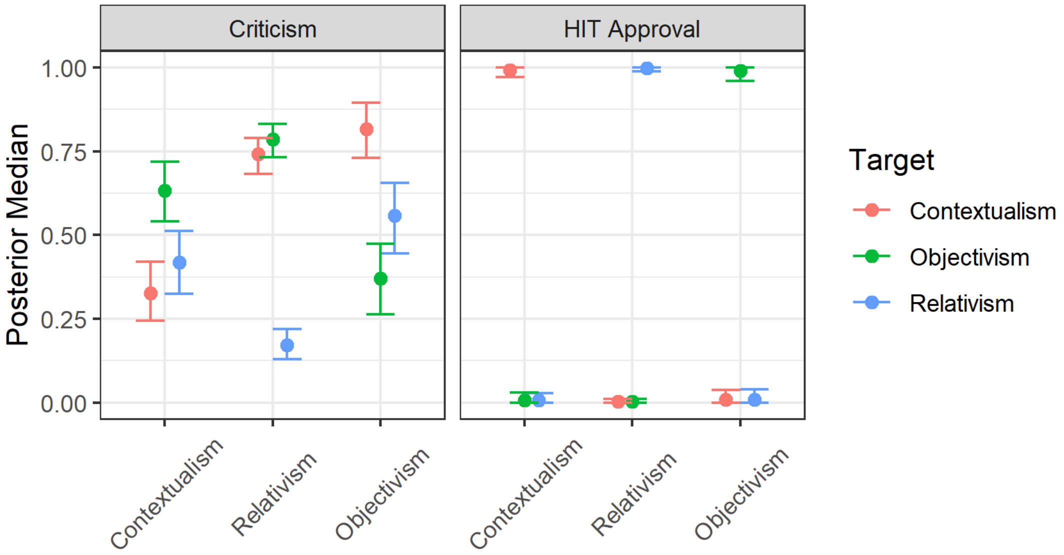


Figure 3. Phase 2 Classification. Posterior Median estimates based on the three latent classes in the Scorekeeping Task, displayed on the x-axis. In this analysis, 253 participants were classified as following Relativism, 79 as following Objectivism, and 101 as following Contextualism. Only the 433 (80.19%) participants, who were classified as belonging to either of the latter three mixture groups in phase 1 entered the latent class analysis. The parameter estimates indicate the posterior median probabilities of a) finding the criticism of the respective view compelling (yes vs. no coded as 1 vs. 0), b) approving the HIT of the respective view after having seen the mutual criticism of all sides, and c) of being classified with the respective semantic interpretation in phase 1. The target labels in the legend indicate which semantic interpretation is criticized, approved, and assigned in phase 1, respectively. The error-bars indicate 95% HPD.

The latent class identification in Figure 3 was based on the posterior probabilities of both agreeing with the criticism of Relativism, Objectivism, and Contextualism, respectively, and the corresponding HIT approvals (after learning the criticism of all sides). As shown, the posterior median HIT approvals in the phase 2 classification were above 99% for all latent classes, and the posterior median probability of accepting criticism of the assigned view was 17.12%, 32.72%, and 37.06% for Relativism, Contextualism, and Objectivism, respectively.

Table 12 shows for each of the three mixture groups above, how their responses were mapped onto the three latent classes of the Scorekeeping Task:

Table 12. Comparison between Mixture Groups and Latent Classes

	Relativism$_2$ (N=253)	Objectivism$_2$ (N=79)	Contextualism$_2$ (N=101)
Relativism$_1$ (N=266)	**199 (75%)**	19 (7%)	48 (18%)
Objectivism$_1$ (N=98)	38 (39%)	**49 (50%)**	11 (11%)
Contextualism$_1$ (N=69)	16 (23%)	11 (16%)	**42 (61%)**

Note. ‘R’ = Relativism, ‘O’ = Objectivism, ‘C’ = Contextualism. The phase 1 and phase 2 classifications are indicated via subscripts: ‘Relativism$_1$’ = phase1; ‘Relativism$_2$’ = phase2. Only the 433 (80.19%) participants, who were classified as belonging to either of the latter three mixture groups in phase 1 entered the latent class analysis based on participants’ performance in the Scorekeeping Task. The mode classify-cation of latent classes is marked in bold for each of the three mixture groups in the upper part of the table.

As Table 12 shows, the central tendency of the three mixture groups was to follow their assigned interpretation of might statements in the Scorekeeping Task. 61% or above of the participants in the mixture groups for Relativism and Contextualism did this. For the mixture group of Objectivism, it was only ca. 50%, however.

These results indicate a considerable degree of correspondence between participants' case judgments in Phase 1 and their reflective attitudes after having been exposed to the mutual criticism of the fictive participants in Phase 2. The exception is Objectivism where a considerable proportion shifted to Relativism in phase 2. On both classifications, Relativism is the most widespread view.

It is possible to achieve a higher correspondence between the two classifications by including the phase 1 classification as a covariate in the latent class analysis for the phase 2 classification. In an earlier frequentist analysis,[112] we did this and found that 88.41% of the contextualists from phase 1 were assigned to the latent class of Contextualism. The price of such a classification is, however, that some participants are included in the latent class of Contextualism, who approve the HIT of Relativism. We therefore prefer the present latent class analysis, which is purer in that participants are only assigned to a given latent class if they have close to 1.0 posterior probability of approving the corresponding HIT. Based on both these analyses, we conclude that participants, who respond in accordance with Objectivism in the Eavesdropper Task are less stable in their response tendencies than the other participants, since a sizeable proportion shifted to Relativism in the Scorekeeping Task.

Moving forward to the assessment of consistency in application of the norm of retraction vis-à-vis the assigned semantic interpretation, we follow Skovgaard-Olsen et al. (2019) in using the profiles established by the Scorekeeping Task as a basis for the comparison. The reason is that these latent classes indicate participants' reflective attitudes after they have been exposed to multiple ways of completing the same task and when they are put in a situation, where they have to commit to one of these interpretations.[113]

8.2.2 Session 2

The Norm of Retraction

A week later, the same participants from Session 1 were invited to participate in a task probing how participants would react to violations of the norm of retraction. Session 2

[112] See the osf project page: https://osf.io/g6vym/

[113] An earlier analysis that includes both the phase 1 and phase 2 classifications can be found on the osf project page: https://osf.io/g6vym/

investigated whether participants classified by the latent classes in Session 1 would react to norm violations of the norm of retraction in accordance with the interpretation of epistemic modals assigned. This leads to our sixth hypothesis.

> (H_6) Only participants classified according to Objectivism and Relativism enforce the norm of retraction for might-statements.

To investigate whether participants thus classified followed this line of response, Session 2 presented the same participants from Session 1 one week later with a series of items, where the norm of retraction was violated for factual-statements, might-statements, and for-all-we-know statements. It was then investigated how participants reacted to these potential norm violations. The factual statements and the for-all-we-know statements were used as baselines for cases, where the norm of retraction applies or does not apply, respectively, to test for differences between the two and might-statements across the phase 2 classification from Session 1.

In the psychology of reasoning, participants who have received instruction in logic are routinely excluded from participation to tap into participants' natural linguistic competence (see e.g., Evans, 2002; Klauer et al., 2010). In our experiments, we decided instead to include previous exposure to logical training as a covariate in the models to examine empirically whether it played a role for participants' adherence to the norm of retraction. We did this, because we expected that participants, who had received prior training in logic, would be more likely to exhibit consistent competence profiles when presented with questions that test subtle differences in the truth, justification, and retraction of might-statements than participants, who had not received such training. By including this co-variate in our analysis, we could investigate whether the degree of internal consistency between session 1 and 2 is higher for participants who had received prior logical training (irrespectively of their adherence to Relativism, Contextualism, or Objectivism). This leads to our sixth hypothesis.

> (H_7) Prior instruction in logic facilitates consistent competence profiles.

8.2.2.1 Method

Participants

Only participants who had taken part in Session 1 and had not been excluded by the exclusion criteria in Session 1, were invited to participate in Session 2 one week later. Out of the 540 invited participants from Session 1, 461 participants (85.37%) took part in Session 2.

The participants were paid on average 6$ an hour for their participation and an additional bonus of 1$ for having taken part in both sessions.

Participants who indicated that English was not their native language, that they would not take their participation seriously, or who completed the task in less than 240s or more than 3600s were excluded. The final sample consisted of 438 participants whose responses from both sessions could be identified. Mean age was 39.40 years, ranging from 18 to 78. 47.95% of the participants self-identified as male, 51.83% self-identified as female, and one participant chose not to identify with either gender. 81.05% of the participants indicated that the highest level of educated that they had completed was an undergraduate degree or higher. 27.40% indicated that they had previously received prior training in logic.

Design

The experiment had a within-subject design with one factor: Sentence (with three levels: factual vs. might vs. for-all-we-know). To allow for six trial replications for each cell of the design, each participant in total went through 18 trials in random order.

Materials and Procedure Specific to Session 2

The within-subject conditions were randomly assigned to 18 different scenarios (see the *osf* project page[114] for sample scenarios). The participants first went through two practice trials and were then instructed to pay attention to small details in the sentences highlighted in blue. For each of the 18 scenarios, the participants were first presented with the first part of the scenarios (the content above the text in bold in the example below) and could then click "continue" for the presentation of the rest of the scenario (the content beginning with the text in bold). One of the 18 scenarios was as follows:

> During a murder trial the defense lawyer presents evidence that his client did not commit the murder but he is also challenged by the blood found at the crime scene. At the trial, the defense lawyer says:
>
> > "FOR ALL WE KNOW, the blood was planted on the scene of crime"
> > [/"The blood MIGHT be planted on the scene of crime"]
> > [/"The blood WAS planted on the scene of crime"]
>
> The prosecutor has the forensic lab conduct a test that can rule out that the blood was planted on the scene of the crime.

[114] https://osf.io/g6vym/

In court, the defense lawyer reacts to the forensic report by saying:

Oh really? I guess then the prosecutor is right. But I still maintain it was true that:

“FOR ALL WE KNOW, the blood was planted on the scene of crime”
[/”The blood MIGHT be planted on the scene of crime”]
[/”The blood WAS planted on the scene of crime”]

when I said it. I stand by my claim and refuse to take it back.

The second occurrence of the speaker’s claim was highlighted in blue, and the task of the participants was to indicate whether they agreed/disagreed (“yes, I agree” vs. “no, I disagree”) with the following statement: “The defense lawyer should have taken back his claim by saying ‘Oh, then I guess I was wrong’”.

The participants completed the 18 within-subject conditions in random order with a unique assignment of the 18 scenarios to each condition. After completing the task, participants were asked a few demographic questions.

8.2.2.2 Results and Discussion

Session 2 served the goal to measure whether participants react to violations of the norm of retraction for might-statements as a function of their phase 2 classification (H_6), and whether they had received previous logical training (H_7). Accordingly, the main dependent variable was a binary variable (“yes” vs. “no”) indicating whether the speaker had violated the norm of retraction by refusing to take back their assertion after the fact $\neg p$ was revealed.

Using the statistical programming language R (R Core Team, 2015), mixed linear models with a binominal likelihood function were applied to participants’ responses of violations of the retraction norm across the three types of sentences (factual vs. might vs. for-all-we-know). This analysis was conducted using R-package `brms` for mixed-effects models in Bayesian statistics (Bürkner, 2017). A model with random intercepts for participants and items was fitted for each of the two classifications from Session 1. These models permitted the phase 2 classifications to interact with the type of sentence evaluated (factual vs. might vs. for-all-we-know) and included participants’ logical training (yes vs. no) as a covariate along with its interactions with the Sentence factor.

To get an overview over the effects, we can inspect whether the 95% HDI intervals of contrasts in the parameter estimates include zero (marked by the dashed line in Figure 4).

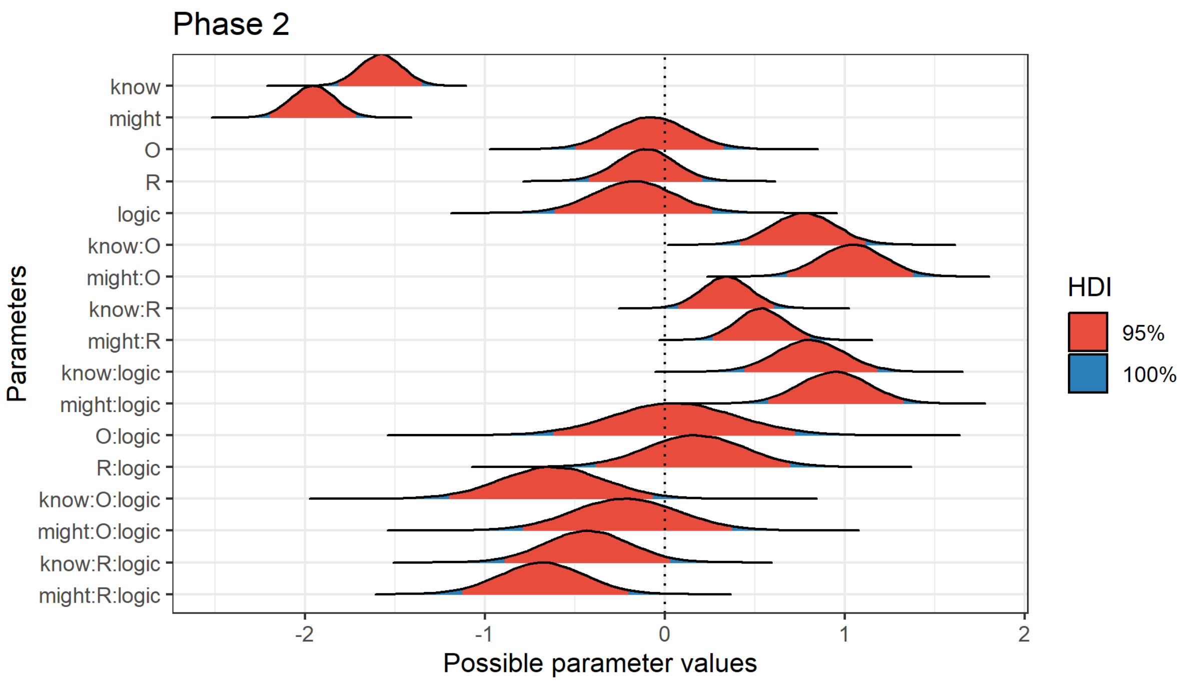


Figure 4. Parameter Estimates and 95% HDI intervals. The plot displays the contrast effects for phase 2. Reference levels: Factual sentences, Contextualism, and no prior training in logic. 'C' = Contextualism, 'O' = Objectivism, 'R' = Relativism.

As the plot in Figure 4 shows, there were credible effects of three and two-ways interaction effects between training in logic, the phase 2 classifications, and the three types of sentences. We will therefore break down the effects by reporting Bayes factors of contrast effects and with a plot of the posterior predictions (Figure 5) as a visual aid.

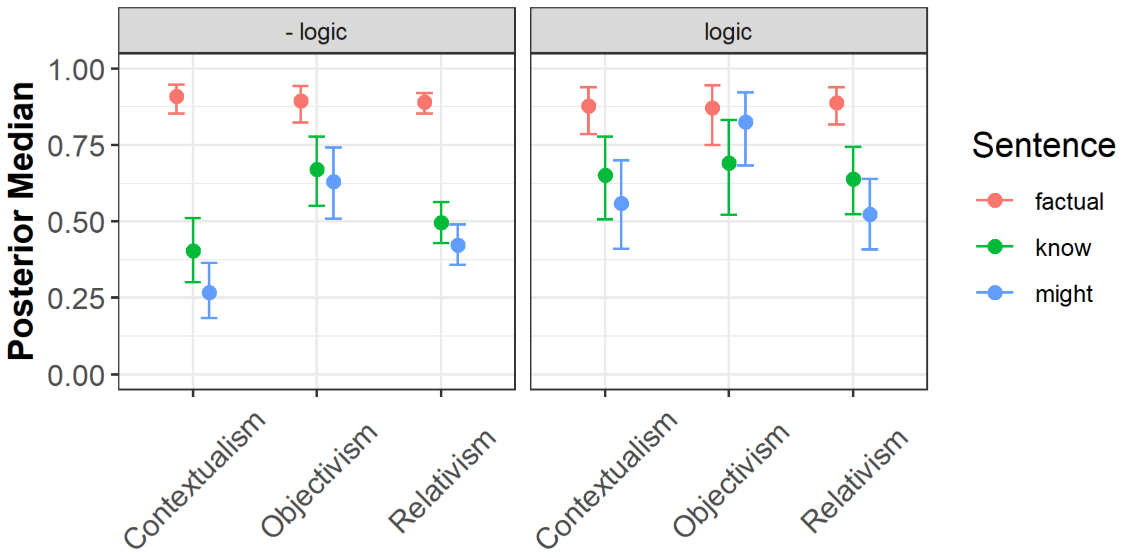


Figure 5. Session 2: Norm of Retraction. Posterior medians of the probability of objecting to violations of the norm of retraction for factual statements, for-all-we-know statements, and might-statements across the phase 2 classification from session 1. The left-side displays the posterior predictions for participants who had not previously received instruction in logic; the right displays the posterior predictions who had received instruction in logic. The error-bars display 95% credible intervals.

Across the phase 2 classification, there were credible effects of having a higher posterior probability of applying the norm of retraction to factual statements than for-all-we-know statements and might-statements ($BF_{H1H0} > 100$). However, in one instance, this effect was qualified by an interaction with the factor of whether participants had received prior training in logic. For participants classified as following Objectivism, who had received prior training in logic, the contrast between factual and for-all-we-know statements was less strong than for the other participants ($b_{factual – know|Objectivism,logic} = .63$, 95% HDI [.25, 1.01], $BF_{H1H0} = 27.05$), and no credible difference was found for these participants between factual and might statements ($b_{factual – might|Objectivism,logic} = .19$, 95% HDI [-.21, .58], $BF_{H1H0} = .22$).

A second finding was that the norm of retraction was applied more strongly to might-statements for participants classified as following Objectivism in phase 2 than for participants classified as following either Contextualism or Relativism. However, the strength of this effect was also qualified by the interaction with prior training in logic.

For participants, who had received no prior training in logic, there was thus very compelling evidence for objectivists applying the norm of retraction more strongly to might statements than contextualists ($b_{Contextualism – Objectivism| - logic} = -.96$, 95% HDI [-1.36, -.56], $BF_{H1H0} > 100$), and moderate evidence for objectivists applying it more strongly than relativists ($b_{Objectivism – Relativism| - logic} = .53$, 95% HDI [.18, .86], $BF_{H1H0} = 8.92$).

In contrast, for participants, who had not received prior training in logic, there was moderate evidence for objectivists applying the norm of retraction more strongly to might statements than contextualists ($b_{Contextualism – Objectivism|logic} = -.76$, 95% HDI [-1.39, -.22], $BF_{H1H0} = 5.51$), and strong evidence for objectivists applying it more strongly than relativists ($b_{Objectivism – Relativism|logic} = .88$, 95% HDI [.34, 1.42], $BF_{H1H0} = 14.03$).

Objectivists and Relativists were predicted to apply the norm of retraction to might-statements, and Contextualists were predicted to not apply the norm of retraction (H_6). This prediction was followed by objectivists both with and without prior training in logic, but more strongly, if they had prior training in logic. In contrast, contextualists only followed the prediction of not applying the norm of retraction to might-statements if they had received no prior training in logic. Moreover, relativists did not follow the prediction concerning the norm of retraction as applied to might-statements. Accordingly, it was only for participants classified as adhering to Objectivism that prior training in logic facilitated following the theory’s predictions. Hence, both (H_6) and (H_7) only received partial support and primarily from the participants classified as objectivists. On the other hand, (H_7) was challenged by the finding that participants classified as contextualists displayed more consistency with their assigned profile, if they had received no prior training in logic.

A further finding was that aside for participants, who followed Contextualism and had not received prior training in logic, credible contrasts between the application of the norm of retraction to might-statements and for-all-we-know statements were not found ($BF_{H1H0} < 3$). For Objectivism and Relativism such differences would, however, had been expected, given that only Contextualism treats the two types of statements as having the same meaning. Accordingly, (H_5) could not be supported by our results.

8.2.3 Summary of Findings from Session 1 and 2

In sum, in Experiment 1 it was found that there were individual differences in interpretations of epistemic modals in line with the re-analysis of published findings in the Introduction (H_1). In Session 1, it was found that participants both differed in their truth evaluations according to Relativism, Contextualism, and Objectivism, and that these truth evaluations were matched by incompatibility judgments underlying disagreement with epistemic modal statements. It was thereby shown that participants could be classified as following a mixture distribution of diverging interpretations of epistemic modals (H_4).

In each case, it was found that participants' truth evaluations could be distinguished from their evaluations of justification, which rules out a central alternative hypothesis of these findings (H_2). Finally, it was found that, of the three interpretations of epistemic modals, the one represented by Relativism was by far the most widespread. This in turn showed that the kind of shifts in truth value judgments of might-statements across different contexts of evaluation predicted by Relativism was frequently occurring (H_3).

Nevertheless, the findings in Experiment 1 also presented two puzzles. Contrary to (H_5), it was found that all latent classes of participants interpreted epistemic modals as similar in meaning to "for-all-we-know" statements in both session 1 and 2. Yet, such an equivalence is only sanctioned by Contextualism (see MacFarlane, 2014, pp. 265).

In addition, it was found in session 2 that the norm of retraction was applied to might-statements most strongly by participants, who had received training in logic and were classified as objectivists. Thus, only for these participants could a facilitation effect from prior instruction in logic be supported (H_7). In contrast, the large subgroup of participants, who were classified as following Relativism showed no strong preferences towards applying the norm of retraction to epistemic modals, counter (H_6). This in turn poses a central explanatory challenge to the account of Relativism in MacFarlane (2014), which treats the norm of retraction as the main diagnostic criterion for separating Relativism from different variants of Contextualism.

8.3 Experiment 2: Norm of Retraction

The goal of Experiment 2 was to reexamine these two puzzles from Experiment 1. First, to better implement the baseline, the time index of for-all-we-know statements was more clearly marked via the formulation "As of Monday, for all we know…".[115] By associating the updated information state with a specific day of the week, this manipulation permitted a simpler integration of the two time points in the for-all-we-know sentences themselves. With one group of participants, we tested the effects of these procedural changes, which leads to our first hypothesis for Experiment 2.

> (H_8) Procedural changes to the baseline of for-all-we-know sentences should make participants less inclined to apply the norm of retraction to these statements.

Second, with a second group of participants, the diagnostic criteria from Bicchieri and Chavez (2013) and Bicchieri (2017) were adopted for testing the empirical reality of the norm of retraction. Bicchieri (2017) investigates ways to measure norms empirically and argues that social norms in general have multiple components: 1) they involve a conditional preference to comply by the norm, which 2) is guided by empirical and normative beliefs about whether other people follow the norm and whether the norm is sanctioned. Each of these components can be measured empirically. The empirical beliefs of participants concern whether the majority of people actually follow the norm in question and the associated risk of being sanctioned by violations. The normative beliefs concern whether participants believe that the majority of people *should* follow the norm and whether norm violations *should* be sanctioned.

In Bicchieri and Chavez (2013), these diagnostic criteria were applied to measure participants' motivation for behaving in a self-serving fashion in a bargaining game. In Experiment 2, we applied some of these diagnostic criteria to examine whether a further group of participants believed that most participants thought that the norm of retraction should be applied to epistemic modal statements and what the probability of being sanctioned for its violation would be. This leads to our final research hypothesis.

[115] We are here grateful to Max Kölbel for this suggestion.

(H_9) Bicchieri's diagnostic criteria for measuring norms match participants' reactions to direct violations of the norm of retraction.

Like Experiment 1, a baseline with factual statements was used, where the norm of retraction uncontroversially holds. Similarly, we also included prior logic training as a covariate to test for possible facilitation effects of prior logic training.

8.3.1 Method

Participants

The same sampling procedure and exclusion criteria were followed as Session 1 of Experiment 1. Unlike Experiment 1, Experiment 2 only had one session. A total of 292 people completed the experiment. After applying the *a priori* exclusion criteria, the final sample consisted of 207 people. Mean age was 41.19 years, ranging from 21 to 74. 50.73% of the participants self-identified as male, 48.79% self-identified as female, and one participant chose not to identify with either gender. 76.33% indicated that the highest level of education that they had completed was an undergraduate degree or higher. 30.92% indicated that they had received prior training in logic.

Design

The experiment had a mixed design with two factors. Sentence was a within-subject factor (with three levels: factual vs. might vs. know). The dependent variable (DV) factor was a mixed factor with levels distributed across two groups of participants (Group 1: DV_{agree}, Group 2: $DV_{sanction}$, $DV_{majority}$). To allow for six trial replications for each level of the Sentence factor, each participant went through 18 within-subject conditions in total.

Materials and Procedure

The same materials and procedures of Session 2 of Experiment 1 were applied to Group 1 in Experiment 2. An exception was that the continuation of the scenarios across t_1 and t_2 was split into two separate days (Monday vs. Wednesday) and that the for-all-we-know statements explicitly highlighted their time stamp via the formulation "As of Monday, FOR ALL WE KNOW, …".

The only difference between Group 1 and Group 2 was that instead of being asked whether they agreed/disagreed that the speaker should have taken the statement back (DV_{agree}), two new DVs based on Bicchieri (2017) were used.

The first ($DV_{majority}$) asked participants whether they agreed/disagreed with the statement that the majority of participants in the study think that the speaker should have taken back the claim by saying “Oh, then I guess I was wrong”. The second ($DV_{sanction}$) asked participants to rate on a scale from 0 to 100% the risk of being criticized for not taking back the statement (which was highlighted in blue to the participants).

Accordingly, the first dependent variable (DV_{agree}) was the same as in Session 2 of Experiment 1, but its formulation had changed. With one group of participants, we investigated these changes to the formulation of DV_{agree}. With a second group of participants, we presented the two new dependent variables from Bicchieri (2017). We decided to present these two new dependent variables in a separate group, to prevent carry-over effects of the dependent variable in the first group to these two new dependent variables.

8.3.2 Results and Discussion

Given the design, there were replicates for each participant and scenario. Accordingly, mixed-effects models, with random effects for intercepts of participants and of items were applied. This analysis was conducted using the statistical programming language R (R Core Team, 2015) and the R-package `brms` or mixed-effects models in Bayesian statistics (Bürkner, 2017). Since the accept and majority dependent variables were dichotomous, a Bernoulli likelihood function with a probit link function was used to model these variables (M_1). A Gaussian likelihood function was used for the continuous %-criticism dependent variable (M_2). Since the same group of participants had responded to both the majority and agree questions, the Sentence (factual vs. might vs. for all-we-know) and DV factors (agree vs. majority) were allowed to interact for this group. Both models included random intercepts and slopes for participants and items.

Following Experiment 1, we also fitted a pair of models that included interactions based on whether participants had previously received instruction in logic. However, unlike in Experiment 1, we did not find credible effects for interactions with prior training in logic with Bayes factors stronger than anecdotal evidence ($BF_{H1H0} < 3$) and the model fitting criteria did not indicate an improvement in the fit by including the interactions.[116] Indeed, in all cases but one, evidence of different strength was found in favor of the H_0 of no effects involving the logic factor ($0.02 < BF_{H1H0} < 0.60$).

[116] M_1 without Sentence * DV * logic interaction (LOOIC = 3272.0); M_1 with Sentence * DV * logic interaction (LOOIC = 3273.2). M_2 without Sentence*logic interaction (LOOIC = -653.0); M_2 with Sentence*logic interaction (LOOIC = -653.4).

We therefore decided in favor of the models without the interaction effects with prior training in logic and display their posterior predictions in Figure 6 below.

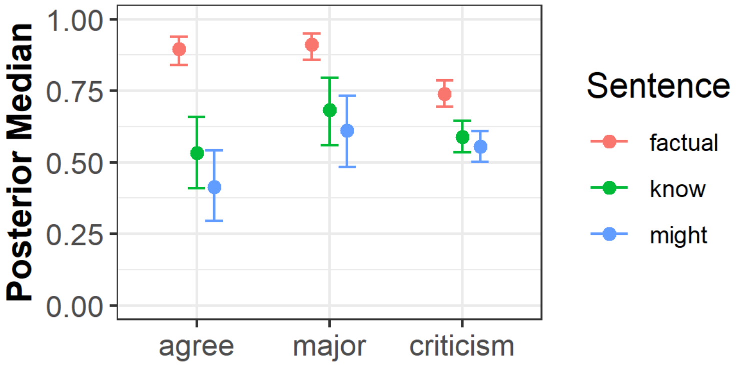


Figure 6. Norm of Retraction. Across various sentence types, Figure 6 displays the posterior predicted median probabilities of 1) agreeing that breaching the norm of retraction is a norm violation (“agree”), 2) judging that the majority of participants would think that breaching the norm of retraction is a norm violation (“major”), and 3) the risk of being criticized for failing to comply with the norm of retraction (“criticism”). The error-bars represent 95% CI intervals.

For all three dependent variables, there was a tendency for the ratings to follow the following rank-ordering: factual > know/might with Bayes factors for the contrasts (might – factual; for-all-we-know – factual) indicating very strong evidence in both models (BF_{H1H0} > 100). In addition, it was found that the data did not support a difference between might-statements and for-all-we-know statements for all three dependent variables (BF_{H1H0} < 3). Indeed, positive evidence in favor of a lack of a difference between these two types of sentences was found for participants’ judgments about whether the majority people would think that breaching the norm of retraction is a norm violation (BF_{H1H0} = 0.11) and for participants’ assessments of the risk of being criticized for not taking back the respective statements (BF_{H1H0} = 0.053).

The results displayed in Figure 6 are in line with the findings of Experiment 1, but the explicit inclusion of the time index “As of Monday” in the for-all-we-know statements lead to a better implementation of the baseline (H_8). For Objectivism and Relativism, it poses a problem that the central tendency from Experiment 1 could be replicated that participants lack strong preferences for applying the norm of retraction to might-statements.

What Experiment 2 adds is that this lack of inclination was matched by their judgments about whether violations are sanctioned and whether the majority of participants think that the norm of retraction should be applied to might-statements (H_9). Given Bicchieri’s (2017) classification of norms, this indicates that participants’ empirical beliefs concerning the application of the norm of retraction are lacking.

In Experiment 2, strong evidence in favor of interactions with prior logical training were not found. In Experiment 1 such interactions were found. But we also saw that these interactions depended on which interpretation of might-statements that participants had been assigned to, which is a factor that the simpler experimental design of Experiment 2 did not permit us to investigate. Thus, abstracting away from latent profile of interpretation of might expressions, Experiment 2 suggests that no general facilitation of performance with the norm of retraction based on prior training of logic was found, contrary to (H_7).

8.4 General Discussion

Truth-conditional semantics developed into a flourishing theoretical framework by accounting for the compositionality of sentences describing the outer world. Eventually, it was enriched to handle context sensitivity (Kaplan, 1989) and counterfactual alternatives to the course of history (Lewis, 1973). One aspect that escaped it was, however, subjective language as involved in taste judgments and uncertainty (Lasersohn, 2017). In a recent discussion, several theoreticians have proposed to extend the toolkit of formal semantics with relative truth values to fill this gap (Egan, Hawthrone, & Weatherson; Kölbel, 2009, 2015a, 2015b; Egan, 2007; MacFarlane, 2011, 2014; Lasersohn, 2017). In MacFarlane (2011, 2014), this takes the form of making the truth values of a range of expressions (e.g., conditionals, subjective taste predicates, epistemic modals, and knowledge ascriptions) relative to a context of assessment.

In contrast, one of the main theoretical alternatives, Contextualism, makes the truth value of sentences containing epistemic modals dependent on the information state at the context of utterance (Hacking, 1967; DeRose, 1991; Kratzer, 1977, 2012; von Fintel & Gillies, 2008, 2011). On this view, epistemic modals are used to make statements *about* a particular body of evidence available at the context of utterance. For Relativism, on the other hand, the content of the statements expressed is not about any body of evidence in particular but has a truth value that varies depending on the evidence available at the context in which it is assessed (Khoo & Phillips, 2018).

A key diagnostic case for arbitrating in this dispute is eavesdropping cases (MacFarlane, 2011). By placing the speaker and the interlocutor at contexts with diverging information states, these cases dissociate the influence of the context of utterance from that of the context of assessments (Katz & Salerno, 2017). As such, eavesdropping cases directly address the need for semantic values that are a function of contexts of assessments. What's more, eavesdropper cases characterize a situation in which the interlocutor knows more than the speaker, and yet chooses to evaluate the speaker's claim for its truth/falsity in the light of information that is only accessible to the interlocutor.

If it could be empirically substantiated that competent English speakers evaluate statements containing epistemic modals in this way, then it would constitute a direct challenge to the core claim of the contextualist that only the information available at the context of utterance matters for the truth of epistemic modals. As Katz and Salerno (2017, p. 142) note:

> Movement from the default contextualist position to a more exotic relativist framework then requires forceful motivation. In that spirit the relativist emphasizes empirical data that allegedly only she can accommodate.

Consequently, it was investigated in several recent papers whether empirical data supports such a shift from Contextualism to Relativism. For instance, in Knobe and Yalcin (2014), it was investigated whether Relativism adequately characterizes data on truth values and retraction. In Katz and Salerno (2017) and Khoo and Phillips (2018), it was investigated whether Relativism adequately characterizes data on the compatibility/incompatibility of two contrary truth value assignments. In each case, it was found that the evidence in favor of Relativism was lacking. But the empirical picture was complicated by the fact that the results did not support standard Contextualism either.

We have already seen how some of this evidence relies on reporting statistical analyses at the group level. When re-examining their results, it was found that they contained considerable variation at the individual level, as we have seen. Consequently, in the present paper we set out to investigate individual variation in the interpretation of epistemic modals by conducting a study with two test sessions. In the first, participants were classified according to three semantic interpretations of epistemic modals at the individual level based on both their truth value assignments. In the second, their adherence to the norm of retraction was investigated. For the remainder, we present some of the open questions raised by our results. We will focus on argumentation with epistemic modals. Finally, we consider how our results relate to other theories in psychology that deal with epistemic modals.

Shifting Truth Values

Historically, the notion of relative truth has had a bad reputation in that it became mingled up with general debates about postmodernism, the science wars, and social constructivism (Boghossian, 2007). Indeed, the very notion has often been suspected of being circular, although the details of the argument are not that simple (Nozick, 2001). It took technical refinements to show that a respectable formal notion of relative truth could be explicated via relativity to contexts of assessment (MacFarlane, 2014). This enabled use of the notion as a theoretical construct to account for a fragment of natural language. But the empirical adequacy of this proposal still needs to be established, since the norms in question are supposed to account for our linguistic competence.

The radicality of introducing relative truth values for this purpose consists in permitting both that the truth values of the same content can diverge between multiple interlocutors and can shift across time for the same interlocutor. Since the content expressed stays invariant, this notion is importantly different from the more common notion that the same sentence can express different contents in the mouth of different speakers (MacFarlane, 2014). One salient example is sentences containing indexicals and other context-sensitive terms (e.g., “I am hungry”). For the contextualist, sentences containing epistemic modals (e.g., might, must) express a similar kind of context sensitivity. In contrast, for the relativist, such sentences can express the same content across different contexts and yet vary in truth values for different speakers as well as shift truth values for the same speaker across time. As emphasized by Wright (2008), it is crucial that direct evidence be had for shifting truth values across different context of assessment for the empirical case for Relativism. Through our experimental investigations, we have been able to supply some of the first empirical evidence for shifting truth values.

MacFarlane (2014, p. 107) is skeptical of the prospect of seeking direct evidence for Relativism in this way. The reason is that the semantic theory of Relativism employs a technical notion of truth that applies to utterances at contexts, whereas he takes it that ordinary speakers use a monadic truth predicate that only applies to the propositional content expressed by sentences. In contrast, Kölbel (2015b, pp. 7) puts forward a principle that connects relativistic semantics with measurable data, which presupposes that competent language users *can* judge the truth of potential utterances of sentences. In Experiment 1 (phase 1), we attempted to collect such data by referring to a token statement made at t_1 and asking participants whether that statement *was* true when it was made and whether the statement made at t_1 *is* true after the continuation had been learned at t_2 (see Table 4). Using this set-up, we were able to find qualitatively distinct patterns in participants’ truth assignments matching the predictions of the competing theories.

Both Khoo and Phillips (2018) and Cantwell (forthcoming) consider the possibility that mixed results in previous studies may have been facilitated by a difficulty in separating truth and justification by the participants. We therefore measured whether participants thought that the speaker was justified in making the statement as a control in Experiment 1 to test this alternative hypothesis. The results indicate that irrespectively of which semantic interpretation participants adopt of epistemic modals, they are capable of distinguishing between truth and justification for the various sentences examined.

Against prior expectations, it was, however, found that for each of these three interpretations, the evaluation of for-all-we-know statements and epistemic modals were aligned. *A priori* such an alignment would only have been expected for participants classified as following Contextualism (MacFarlane, 2011, 2014). Yet, it was found that participants differed as sharply in their interpretations of for-all-we-know statements as they did for epistemic modals. One hypothesis is that a clearer demarcation of the time indices of the points of evaluation would help participants to draw a distinction between the two types of statements. Support for this hypothesis could be obtained in Experiment 2, where t_1 and t_2 were separated into distinct days and the for-all-we-know statements included the explicit qualification "As of Monday, for all we know…".

In Experiment 1, it was found that participants classified as following Relativism and Objectivism extended their tendency to evaluate past tokens of "for all we know, *p*" as asserted by a speaker at t_1 (14:10), by their own present state of knowledge at t_2 (14:12), just as they would with might-statements. Possibly, these participants were misled by that "might *p*" sounds similar in meaning to "for all we know, *p*" when presented interchangeably in a within-subject design, although their truth evaluation of the former did not fit for the latter. Future studies will have to look into this possibility.[117]

Argumentation with Epistemic Modals

A large part of the debate over Relativism concerns argumentation with subjective expressions such as epistemic modals. Views accordingly differ on whether a Trump supporter uttering "Trump might have won the 2020 election" and a Democrat denying this are in genuine disagreement, or whether they are merely talking past one another by expressing features of their own subjective state of uncertainty (see the essays in Egan & Weatherson, 2011). A central motivation for Relativism is to allow that two speakers may be in a state of genuine disagreement about the same content but where each is correct according to their own perspective (Kölbel, 2009).

[117] A further possibility includes using ‚for all I know' instead ‚for all we know' as a baseline. We originally decided against this option, since the position of Solipsistic Contextualism, which narrows the context down to just consist of the speaker, is rarely advocated. Rather, the more wide-spread view is that the context includes the joint information state of a group of interlocutors (MacFarlane, 2014). Yet, results in Kneer (forthcoming) indicate that this other baseline would have been easier to implement.

At the same time, Relativism aims at securing a basis for argumentation via its norm of retraction (MacFarlane, 2011, 2014): although the truth values of statements containing epistemic modals change with the information state of the context of assessment, statements that come out as false at any given time need to be retracted. Consequently, if the Trump-supporter above changes her information state by taking the outcome of the 2020 election at face value, then her earlier might-statement would have to be retracted. Yet, if the information states of neither the Democrat nor the Trump-supporter changes, they may continue to be in a state of "faultless" disagreement as far as the might-statement goes (Kölbel, 2009).

When examining the norm of retraction in Experiments 1 and 2, high-stakes scenarios were used where a speaker makes a public statement in, e.g., media outlets, the court, or in scientific disputes. Still, it was found that participants in general did not have strong inclinations to apply the norm of retraction to epistemic modals, contrary to the predictions of Relativism and Objectivism.

After having examined patterns of individual variation, we can thus agree with Knobe and Yalcin (2014) that strong support for the norm of retraction as applying to epistemic modals is, in general, not found. But it was, however, found that the shifting truth evaluations of Relativism characterized the largest latent class of participants in Experiment 1, and that these participants consistently stood by these judgments when challenged by contrary views in the Scorekeeping task. Moreover, it was found that these truth evaluations were matched within this sub-group by corresponding incompatibility judgments. So, whereas Katz and Salerno (2017) and Khoo and Phillips (2018) find at the group-level that participants do not follow the predictions of Relativism of treating two contrary epistemic modal statements with different evidence accessible as incompatible, support for this prediction was found in Experiment 1 for the largest subgroup of participants.

Based on our results, it is likely that ordinary people on average do not apply the norm of retraction to epistemic modals, because they regard them as a type of hedged assertions, which are less costly to make. Williamson (2020, p. 11) gives expression to this intuition:

> When the detective rightly asserts 'Smith must be guilty', she does not regard her earlier assertion of 'Smith may be innocent' as in any way wrong. It was not a *mistake* made on the basis of incomplete but misleading evidence. That was the point of saying 'Smith may be innocent' rather than 'Smith is innocent'.

In support, it was found in Experiment 2 that participants tend not only to lack strong preferences to apply the norm of retraction to epistemic modals themselves, but that they also accurately judge that the majority lack similar preferences, and that the chances are low for being sanctioned, if one refused to retract a previous might-statement. So, while the norm of retraction provides an interesting normative justification in MacFarlane (2011, 2014) for Relativism as applied to epistemic modals – in terms of the foundation for argumentation it provides – it is not a norm that participants in general have strong normative views about for might-statements. It is only for factual statements that prescriptions of retraction appear to be an important norm to most of the participants we investigated.

However, our study of individual differences permits us to qualify this negative assessment in one important aspect. The results indicate that for participants who have had training in logic and committed to Objectivism in the Scorekeeping task, the norm of retraction was applied to the same degree to might-statements as factual assertions. So at least for this minority, the norm of retraction finds application to epistemic modals.

Our findings leave us with a general explanatory challenge: how is argumentation with epistemic modals possible, given that it is not based on the norm of retraction as MacFarlane (2014) has argued? Since epistemic modals are used to state alternative hypotheses in science, it would be strange if no norms of argumentation applied to them.

Addressing this first challenge is further complicated through a second challenge raised by another topic we encountered. Based on the results of Experiment 1, as well as our re-analysis of published findings, we have found grounds for questioning the following widely shared assumption:

> (U) There is a uniform interpretation of expressions like epistemic modals. Only one of conflicting semantic theories can be descriptively adequate. If semantic theories like Contextualism, Relativism, and Objectivism are incompatible, then at most one of them can be descriptively correct.

Rejecting (U) implies that to the extent that argumentation over epistemic modals occurs, it is complicated by the circumstance that different speakers may apply different interpretations for their semantic evaluation. This lack of a semantic foundation for a shared standard does not necessarily render argumentation over epistemic modals impossible, just as people can engage in argumentation over moral judgments despite wide disagreement on underlying moral standards. But it likely renders argumentation over epistemic modals more fragile and complicates its theoretical explication.

Nevertheless, it is possible to identify components of argumentation with epistemic modals on which contextualists, relativists, and objectivists can all agree. In appendix 1, Table

A1 we present an overview over such cases, which shows that there are cases which regulate which assertions can be made, and which challenges are appropriate, on which interlocutors can agree, even if they differ on the underlying interpretation of epistemic modals. For instance, if *p* is not excluded by the shared information state, relativists, contextualists, and objectivists can all agree that "might-*p*" is assertable and that the speaker is warranted in rejecting challenges to her assertion of might-*p*. Moreover, the interlocutor can use factual information to challenge the speaker to update the shared information state so that *p* is excluded and might-*p* is not warrantedly assertable. *Vice versa*, if the shared information state had already excluded *p* at the outset, then might-*p* would not be assertable. In which case, interlocutors can warrantedly challenge assertions of might-*p*, or alternatively, use factual evidence to challenge the speaker to update the shared information state so that might-*p* becomes assertable.

Thus, despite the uncertainty of whether one is communicating with a person who evaluates might-statements according to Contextualism, Relativism, or Objectivism, it is possible to follow these rules for arguing over might-statements and to coordinate which challenges need to be addressed by the speaker. By following these rules, the Democrat and the Trump-supporter can argue over whether it is correct to assert "Trump might have won the 2020 election". But as this example illustrates, often the root of the dispute is not so much the might-statement itself but a dispute about whether an update to the underlying information state with factual information should be performed. In their case, the root of the dispute concerns factual information about the 2020-election.

What this example illustrates is that there is an important asymmetry between arguing over epistemic modals and arguing over other types of subjective expressions, like taste judgments. The discourse over epistemic modals parasitizes on factual discourse, whereas taste judgments only have subjective standards of taste to rely on. For this reason, argumentation with epistemic modals have the additional resources of being able to dissolve into a dispute over which updates with factual information are mandated. For taste judgments, this option is absent and for this reason a norm of retraction of the kind that MacFarlane (2014) introduces *may* prove more important for argumentation over taste judgments than for argumentation with epistemic modals.[118]

As part of their argument in favor of the norm of retraction, MacFarlane (2014) and Cantwell (forthcoming) interpret the norm as not implying an admission that the speaker was at fault for having made the assertion at t_1. Rather, the retraction is taken as an admission that

[118] Results in Kneer (2021) cast doubt on this possibility, however.

the statement made at t_1 is semantically false at t_2 after not-*p* has become known. The first type of assessment concerns how well the speaker performed in using the might expression; the second concerns the correctness of the *content* uttered (Kölbel, 2015a, fn. 5). This distinction is subtle, and it therefore remains to be seen whether future studies can find stronger support for the norm of retraction, given other formulations of retractions.

Since epistemic modals have played a role in a recent controversy concerning probability and mental model theory concerning epistemic possibilities as the foundation of psychology of reasoning (Hinterecker et al. 2016, Johnson-Laird & Ragni, 2019, Oaksford, Over et al., 2019), we finally consider relations between these theories and our results.

Comparisons to Theories in Psychology

On popular Bayesian approaches to reasoning, generalized quantifiers (e.g., 'most') have been analyzed probabilistically (Oaksford & Chater, 2007). Complementary, Lassiter (2017) has analyzed 'possible' as a gradable, scalar concept, which can be modeled probabilistically, instead of being an all-or-nothing notion captured by existential quantification. As a possible ordering of epistemic modals, Lassiter (p. 152) considers the following order of probabilistic thresholds ($p(\varphi) > \theta_i$):

$$\theta_{possible} < \theta_{might} < \theta_{likely} < \theta_{must} < \theta_{certain}$$

Normally, 'possible' and 'might' are treated as equivalent (e.g., with $\theta_{possible} = \theta_{might} = 0$), but Lassiter (2017) presents experimental evidence indicating that the former is weaker than the latter (i.e., $\theta_{might} > 0$). This observation is important because the main theory of modal reasoning in psychology (i.e., mental model theory) has explicated a semantics for 'possible' (Johnson-Laird & Ragni, 2019), but not for 'might', as far as we are aware.

In this paper, we have focused on a single epistemic modal without considering issues of gradeability and varying strength of probabilistic information. A natural follow-up would therefore be to investigate whether qualitatively distinct patterns of individual differences persist once further epistemic modals are tested, and the strength of the probabilistic evidence is varied. Such work could complement other recent studies investigating the link between updating assumptions about source reliability and use of constructions with might-expressions to hedge assertions and weaken the speaker's commitments (Collins & Hahn, 2020).

On the other hand, a scalar semantics for epistemic modals would also have to account for our data concerning truth value judgments. On a scalar semantics, binary truth values can be recovered via the thresholds; values above the threshold corresponding to 'true' and those below corresponding to ‘false’ (Lassiter, 2017, Ch. 1). The separation between truth and justification in our data could then be accounted for by holding that participants’ judgments of justification are based on probabilities and probabilistic thresholds introduce binary distinctions between true and false. But now consider the puzzle raised by our results.

For some participants, there was a high probability that might-*p* was justified at t_1, might-*p* is evaluated as having met the threshold at t_1, and continues to be evaluated as meeting the threshold at t_2, although not-*p* is known as a fact (contextualists). For other participants, there was a high probability that might-*p* was justified at t_1, might-*p* is evaluated as having met the threshold at t_1, and might-*p* is evaluated as not meeting the threshold at t_2 (relativists). For yet other participants, there was a high probability that might-*p* was justified at t_1, might-*p* is evaluated as not having met the threshold at t_1, and continues to be evaluated as not meeting the threshold at t_2 (objectivists). Since might-*p* retains a high probability of justification for all participants, the participants would have to be modelled as differing in whether the threshold for binary truth judgements is fixed by the information state at t_1 (contextualists), it shifts between t_1 and t_2 to track the context of assessment (relativists), or whether it is fixed by the known facts at t_2 (objectivists). While it is conceivable that the scalar semantics of Lassiter (2017) could be extended with a theory of individual variation in thresholds and their context-dependence on information states, it is also clear that the theory would need to be extended in some way to be able to account for our results.

As Lassiter (2017, pp. 5-6, 157) notes, no one has yet a complete theory about the context-sensitive interpretations of scalar expressions in general; including a complete theory of the contextual factors that influence the probabilistic thresholds of epistemic modals. But the challenge from our results is different. It suggests that for different participants, different information states influence how the thresholds are set in the same context, if a probabilistic threshold-account of our data on truth value judgments is to be viable.

Turning to mental model theory, we find a very different outlook on epistemic modals. Mental model theory is based on the idea that reasoning depends on models of possibilities. Accordingly, mental model theory presupposes that participant parse natural language into representations of the possibilities that the sentences assert, as illustrated in Table 13 below. For compound sentences (e.g., ‘if p then q’), the semantic meaning is explicated via a conjunction of possibilities, and so if participants are to fully process it (and not take heuristic shortcuts like only processing the first model), they should arrive at all the conjuncts listed.

Table 13. Parsing of Natural Language in MMT

Statements	Possibilities added to the Mental Model	Status
p	$\Diamond p$	Fact
p and *q*	$\Diamond(p, q)$	Fact
p or *q*	$\Diamond(p, q) \land \Diamond(p, \neg q) \land \Diamond(\neg p, q)$	Epistemic possibilities
If *p* then *q*	$\Diamond(p, q) \land \Diamond(\neg p, q) \land \Diamond(\neg p, \neg q)$	Epistemic possibilities
It is possible that *p*	$\Diamond p \land \Diamond\neg p$	Epistemic possibilities
It is not possible that *p*	$\Diamond\neg p$	Fact

Note. The table states the fully explicit models in logical notation. In addition, mental model theory also has a performance theory for heuristic processing, where participants only consider the first possibility. ‘$\Diamond p$’ = it is possible that *p*. ‘$\Diamond(p, q)$’ = it is possible that *p* and *q*. ‘$\Box p$’ = it is necessary that *p*. ‘$\land$ ‘ = logical conjunction. For comparison, see e.g., Johnson-Laird and Ragni (2019, Table 2). When only one possibility is added to the mental model of the premises by a statement, this possibility acquires the status of being a fact (Khemlani et al., 2018).

These conjunctions are thought to be exhaustive and so every other combination is treated as impossible. Factual statements, like ‘*p*’ and ‘*p* and *q*’, only assert one possibility (Khemlani et al., 2018). Recently, mental model theory has been extended to apply to sentences containing epistemic modals (Johnson-Laird & Ragni, 2019). Accordingly, ‘it is possible that *p*’ presupposes the following conjunction of possibilities: ‘*p* is possible’ and ‘not-*p* is possible’.

Could mental model theory account for our results? In the absence of a direct account of might-statements, we use what mental model theory says about ‘possible’ as a proxy, since traditionally the two have been assumed to be similar in meaning. As mental model theory is silent on issues pertaining to eavesdropper cases and the distinction between context of use and context of assessment, we will have to extrapolate to apply it to the Eavesdropper Task.

For this, the reader is encouraged to think back at the temporal structure in Table 4. When presented with the fact that not-*p* in this task, it necessarily follows that ‘it is not possible that *p*’ on mental model theory. Indeed, the two are equivalent on the revised theory, because they each only have one explicit model (‘$\Diamond\neg p$’).[119] Since, moreover, ‘it is possible that *p*’ presupposes both ‘*p* is possible’ and ‘not-*p* is possible’ on mental model theory, a presupposition of ‘it is possible that *p*’ is violated at t_2 when not-*p* has become known. Whether the violation of this presupposition merely renders the statement unassertable at t_2

[119] Due to the duality of ‘$\Diamond p$’ and ‘$\Box p$’ in modal logic, ‘it is not possible that *p*’ (‘$\neg\Diamond p$’) would normally be equivalent only to (‘$\Box\neg p$’), and not to the factual statement (‘$\neg p$’). The revised mental model theory here appears to have difficulties distinguishing between the explicit models of factual statements (‘*p*’) and of the much stronger statements of necessity (‘$\Box p$’). Perhaps for this reason, Johnson-Laird and Ragni (2019) offer an idiosyncratic reinterpretation according to which a single statement of a necessity states a necessary condition for another state of affairs (p. 11). But this account is liable to run into troubles when it comes to representing the strongest epistemic modals (e.g., ‘*p* is certain’) and it misrepresents the content of factual statements. The problem arises due to mental model theory’s modalization of descriptive sentences (see also Over, 2022).

(as all theories agree), a truth-value gap, or false (as Relativism and Objectivism would hold), mental model theory does not say.

To accord with the modal tendency of our results corresponding to the relativistic response pattern, mental model theory would have to hold that ‘it is possible that *p*’ as asserted in t_1 is made false by the disclosure of not-*p* at t_2. Accordingly, to capture the relativistic response, mental model theory would have to be extended by the auxiliary assumption that the violation of said presupposition results in a ‘false’ evaluation at t_2. Yet, mental model theory would still be challenged by the finding that there are other participants (classified as contextualists), who continue to treat the might-statement as true, despite the violated presupposition. Moreover, the difference between relativists and objectivists in whether the might-statement was considered true or false at t_1 is left unaccounted for.

To further apply mental model theory to account for all our results runs into the difficulty that mental model theory has not explicitly been developed to address the same questions as the semantic theories investigated in our experiments (i.e., eavesdropper cases, relativity of truth values to context of use and context of assessment, the distinction between truth and justification for might-statements, retractions, and argumentation with might-statements). As such, our results present an opportunity for both probabilistic approaches and mental model theory in psychology to expand their theories to cover new domains in competition with the cluster of theories we investigated. Since neither theory predicted our results, the challenge is whether the theories can be extended by suitable auxiliary hypotheses that can lead to novel findings of their own.

8.5 Conclusion

In this paper, we have investigated individual variation in the adherence to Relativism, Contextualism, and Objectivism about epistemic modals. Our experimental investigations were motivated by a re-analysis of existing empirical data (Knobe & Yalcin, 2014; Khoo & Phillips, 2018), which we showed contained a substantial degree of individual variation which is not captured by statistics for central tendencies at the aggregated group level.

As an alternative strategy, we followed the individual classification approach of Skovgaard-Olsen et al. (2019), which had previously only been applied to conditionals. On this approach, latent profiles of participants’ case judgments and reflective attitudes are established. It was then probed whether participants are internally consistent across multiple sessions when comparing their latent profile to their adherence to the consequential norms. In the first session of Experiment 1, it was found that participants’ truth value assignments in so-called eavesdropping cases could be used to identify three latent classes of participants of which Relativism was the largest. Through these classifications, we were able to obtain some of the first empirical evidence for shifts in truth values across different time points, in agreement with Relativism. This in turn provides an empirical validation for the theoretical notion of content with relative truth values used as a semantic device for modelling fragments of natural language in MacFarlane (2014).

In a second session, it was then tested whether these three semantic interpretations were consistent vis-à-vis their performance with the norm of retraction informing debates between Contextualism and Relativism. With the exception of participants who had received training in logic and committed to Objectivism, it was found that none of the groups showed a strong adherence to the norm of retraction for epistemic modals. Yet, this norm provides one of the most central motivations for Relativism in MacFarlane (2014). The results of Experiment 2 corroborated this conclusion of lack of general adherence to the norm of retraction for epistemic modals by applying Bicchieri's (2017) diagnostic criteria.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beddor, B. and A. Egan (2018). Might do better: Flexible relativism and the QUD. *Semantics and Pragmatics*, *11*(7).

Bicchieri, C. (2017). *Norms in the Wild. How to Diagnose, Measure, and Change Social Norms*. Oxford: Oxford University Press.

Bicchieri, C. and Chavez, A. K. (2013). Norm Manipulation, Norm Evasion: Experimental Evidence. *Economics and Philosophy*, *29*, 175-98.

Boghossian, P. A. (2007). *Fear of Knowledge: Against Relativism and Constructivism*. Oxford: Oxford University Press.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P., and Vuorre, M. (2018, February 28). Ordinal Regression Models in Psychological Research: A Tutorial. Retrieved from http://doi.org/10.17605/OSF.IO/X8SWP

Cantwell, J. (forthcoming). Objective epistemic modals. (*in review*)

Collins, P., and Hahn, U. (2020). We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(7), 1328–1348.

Cole, R. P., Barnet, R. C., and Miller, R. R. (1997). An Evaluation of Conditioned Inhibition as Defined by Rescorla's Two-Test Strategy. *Learning and Motivation, 28,* 323-341.

DeRose, K. (1991). Epistemic Possibilities. *Philosophical Review, 100*(4), 581–605.

Egan, A. (2007). Epistemic Modals, Relativism, and Assertion. *Philosophical Studies, 133*(1), 1–22.

Egan, A., J. Hawthorne, and Weatherson, B. (2005). Epistemic Modals in Context. In G. Preyer and P. Peter (Eds.), *Contextualism in Philosophy* (pp. 131–170). Oxford: Oxford University Press.

Egan, A. and Weatherson, B. (Eds.) (2011). *Epistemic Modality*. Oxford: Oxford University Press.

Elqayam, S. and Evans, J. St. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*, 233-90.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978-996.

Farell, S. and Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. New York: Cambridge University Press.

Hacking, I. (1967). Possibility. *Philosophical Review, 67*, 143–168.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Blackwell Publishing.

Hinterecker, T., Knauff, M., and Johnson-Laird, P. N. (2016). Modality, Probability, and Mental Models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1606-1620.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, 193, Article 103950.

Kaplan, D. (1989). Demonstratives. In Almog, J., Perry, J., and Wettstein, H. K. (Eds.). *Themes from Kaplan* (pp. 481-563). Oxford: Oxford University Press.

Katz, J. and Salerno, J. (2017). Epistemic Modal Disagreement. *Topoi, 36*, 141-153.

Kellen, D. and Klauer, K. C. (2018). Elementary signal detection and threshold theory. In

E. J. Wagenmakers (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive neuroscience (4th edition, vol. v)*. New York: Wiley.

Khoo, J. and Phillips, J. (2018). New horizons for a theory of epistemic modals. *Australian Journal of Philosophy*, *97*(2), 309-324.

Klauer, K. C., Beller, S., and Hütter, M. (2010). Conditional Reasoning in Context: A Dual-Source Model of Probabilistic Inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 298-323.

Kneer, M. (2021). Predicates of personal taste: empirical data. *Synthese*, *199*, 6455-6471.

Kneer, M. (forthcoming). Epistemic Modal Claims – Data and 'data'.

Knobe, J. and Yalcin, S. (2014). Epistemic modals and context: Experimental data. *Semantics & Pragmatics*, *7*(10), 1-21.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy, 1*(3), 337–356.

Kratzer, A. (2012). *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kölbel, M. (2003). Faultless Disagreement. *Proceedings of the Aristotelian Society,* 104, 53-73.

Kölbel, M. (2009). The Evidence for Relativism. *Synthese*, *166*(2), 375-95.

Kölbel, M. (2015a). Relativism 1: Representational Content. *Philosophy Compass 10*(1), 38-51.

Kölbel, M. (2015b). Relativism 2: Semantic Content. *Philosophy Compass 10*(1), 52-67.

Lasersohn, P. (2017). *Subjectivity and Perspective in Truth-Theoretic Semantics*. Oxford: Oxford University Press.

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives.* Oxford: Oxford University Press.

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell Publishing.

Li, Y., Lord-Bessen, J., Shiyko, M., and Loeb, R. (2018). Bayesian Latent Class Analysis Tutorial. *Multivariate Behav Res., 53*(3), 430-451.

Linzer, D. A., and Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, *42*(10), 1-29.

MacFarlane, J. (2011). Epistemic Modals Are Assessment-Sensitive. In B.Weatherson and A. Egan (Eds.), *Epistemic Modality* (pp. 144–178). Oxford University Press.

MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. 10.21105/joss.01541

Miller, R. R., Hallam, S. C., Hong, J. Y., & Dufore, D. S. (1991). Associative structure of differential inhibition: Implications for models of conditioned inhibition. *Journal of Experimental Psychology: Animal Behavior Processes*, *17***,** 141–150.

Mair, P. (2018). *Modern Psychometrics with R*. Cham: Springer.

Nozick, R. (2001). *Invariances: The Structure of the Objective World*. Cambridge, MA: Harvard University Press.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., Over D. E., & Cruz, N. (2019). Paradigms, possibilities, and probabilities: Comment on Hinterecker et al. (2016). J*ournal of Experimental Psychology: Learning Memory and Cognition*, *45*, 288-297.

Over, D. E. (2022). The new paradigm and massive modalization. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2021.2017346

Plummer, M. (2019). *rjags: Bayesian graphical models using MCMC*. R package version 4-10. URL = https://CRAN.R-project.org/package=rjags.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Skovgaard-Olsen, N. (2019). The Dialogical Entailment Task. *Cognition*, *193*.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., and Klauer, K. C. (2019). Norm Conflicts and Conditionals. *Psychological Review*, *126*(5), 611-633.

Teller, P. (1972). Epistemic possibility. *Philosophia*, *2*(4), 303–320.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 27(5), 1413-1432.

Vehati, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2019). Pareto smoothed importance sampling. https://doi.org/10.48550/arXiv.1507.02646 von Fintel, K. and A. Gillies (2008). CIA Leaks. *Philosophical Review*, *117*(1), 77–98.

von Fintel, K. and A. Gillies (2011). 'Might' Made Right. In A. Egan and B. Weatherson (Eds.), *Epistemic Modality* (pp. 108-130). Oxford: Oxford University Press.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev*, *25*, 35-57.

Weatherson, B. (2009). Conditionals and Indexical Relativism. *Synthese*, *166*, 333–357.

Wessells, M. G. (1973). Autoshaping, errorless discrimination, and conditioned inhibition. *Science*, *182*, 941–943.

Williamson, T. (2020). *Suppose and Tell. The Semantics and Heuristics of Conditionals. Oxford: Oxford University Press.*

Wright, C. (2008). Relativism about truth Itself: Haphazard thoughts about the very idea. In M. García-Carpintero & M. Kölbel (Eds.), *Relative Truth* (pp. 157-85). Oxford: Oxford University Press.

Appendix 1: Formal Details on Semantic Theories

At the core of a semantic theory lies a compositional (recursive) assignment of semantic values to syntactic expressions relative to some collection of parameters $\mathbf{p} = \langle p_1,\ldots,p_n\rangle$. We can let $[\![e]\!]^{\mathbf{p}}$ be the semantic value of the expression *e* at **p**. When sentences are involved the semantic value assigned is typically *true* or *false* (1 or 0). The parameters we use will depend on what kind of syntactic expressions one wishes to capture in the semantic theory. If, for instance, the language contains temporal operators, we need some parameter for time. Most semantic theories insist that there is some parameter denoting the state of the world—a representation of the 'factual' properties of the world—and the values for this parameter are usually referred to as *possible worlds*.

Semantic theories for might-modalities typically assess a sentence of the form "It might be the case that *A*" (or, simply, "Might*A*") relative to a *set X* of possible worlds, with the idea that Might*A* is true if *A* is true at some world in *X*. Assuming that the only parameters used are a possible world *w* and a set of worlds *X*, the standard recursive semantic clause for the might modality then goes:

$$[\![\text{Might}A]\!]^{w,X} = 1 \text{ if and only if for some } w' \in \text{X: } [\![A]\!]^{w',X} = 1.$$

A basic compositional assignment of semantic values like the one sketched above in and by itself only provides the means to state *semantic* relationships between sentences. For instance, Might(*A*&*B*) logically entails Might*A* in virtue of the fact that in every model and for every *w* and *X*, if $[\![\text{Might(A\&B)}]\!]^{w,X} = 1$, then $[\![\text{MightA}]\!]^{w,X} = 1$. Importantly, the dividing lines between contextualist, relativist and objectivist analyses of might-modalities do not concern such semantic relationships. So, they can in principle agree on the same basic semantic theory. The differences show up when we try to link the basic semantics to norms and conditions of language *use*. This is sometimes referred to as the *postsemantics* and deals with a whole cluster of questions about how to employ the basic semantics when modelling conditions for assertion and assessment of assertions. Here are two such questions:

> **Assertability Conditions:** Under what conditions is a might-sentence properly assertable?
>
> **Assessment Conditions:** Under what conditions should the assertion of a might-sentence be properly assessed as true?

Consider the question of assertability first. It makes explicit reference to assertability and so implicitly presupposes a speaker, and a context of use c_u (including a potential audience) in which assertions are made. It is typically assumed that the collective or individual beliefs or knowledge—the *information state*—of the participants in c_u can be aggregated into an information state that can be represented as a set of possible worlds X_u. In the simplest case, this set represents what the speaker knows or believes. In more complex accounts, it is a combination of the beliefs and knowledge of the speaker and audience. Let us focus on the simplest case where $X_{c_u} = X_s$ = the set of worlds consistent with what the speaker s believes true. Given a set of possibilities X_s, most accounts would accept some variant of the following:

Might*A* is *reasonably* assertable for *s* if and only if for some $w \in X_S$: $[\![A]\!]^{w,\,X_s} = 1$.

That is, it is reasonable to assert Might*A* if *A* is consistent with one's information state. The key normative phrase here is *reasonably* assertable, which is here intended as being weaker than *properly* assertable. A competent speaker of English who falsely (but on reasonable good grounds) believes that it is raining in London can *reasonably* assert "It is raining in London", even though the assertion is false. Being reasonable does not free one of culpability. One can, for instance, be expected to apologize for a false assertion and retract it if one finds out that it is false. If a sentence is *properly* assertable, however, there can arise no issue of subsequent culpability or demands for retraction.

For might-sentences the contextualist and objectivist (for different reasons) agree on the stronger normative reading:

Might*A* is *properly* assertable for *s* if and only if for some $w \in X_S$: $[\![A]\!]^{w,\,X_s} = 1$.

That is, if *A* is consistent with the information state then there is nothing wrong *whatsoever* with asserting Might*A* (though there may of course be prudential reasons for not asserting Might*A*). In particular, there is no need to apologize or retract if it turns out that *A* is false. The relativist, however, does not agree: whether an assertion was *proper*—and so not susceptible to demands of retraction—depends on the information state of who is assessing the claim.

When we turn to conditions for proper assessment we need to take into consideration the information state of the *assessor a*, again represented as a set of possible worlds X_a. The relativist and the objectivist agree on the following:

An assertion of Might*A* is *properly* assessed as true if and only if $[\![\mathrm{Might}A]\!]^{w,X_a} = 1$.

On these accounts one *should* assess a might-claim based on the information state in the context of assessment. The contextualist, however, does not even agree on the normatively weaker claim:

> An assertion of Might*A* is *reasonably* assessed as true only if $[\![\text{Might}A]\!]^{w,X_a} = 1$.

That is, on the contextualist account, one does not even come across as *linguistically* competent if one assesses a might-claim relative to one's own information state.

The differing answers to the questions concerning assertability and assessibility is sufficient to differentiate contextualism, relativism and objectivism, as shown in Table A1.

Table A1. Post-Semantic Differences

	Contextualism	Relativism	Objectivism
Assertion of Might*A* assessed false if *A* known false.	No	Yes	Yes
Assertion of Might*A* assessed culpable if *A* is known false.	No	Yes	No

But how do the different accounts arrive at these different answers? Due to different answers to the following questions:

> **Content Conditions:** What does the propositional content of a might sentence depend on?
>
> **Truth Conditions:** Under what conditions is the proposition expressed by a might-sentence true?

According to the contextualist, the propositional content of a might-sentence depends on the information state of the speaker. If the relevant information state is what the speaker knows, then an assertion of Might*A* has the propositional content "For all I [the speaker] know, *A* is true".[120] The proposition expressed by an assertion of Might*A* in a context of utterance c_u, in symbols $|\text{Might}A|^{cu}$, can thus be represented as the set of worlds in which *A* is consistent with the speaker's beliefs. The proposition $|\text{Might}A|^{cu}$ is true in w_{c_u} (the world of the context of use), if w_{c_u} is an element of $|\text{Might}A|^{cu}$.

[120] Here a distinction is usually drawn between solipsistic Contextualism, where only the information state of the speaker matters, and the more widespread non-solipsistic version, where the information state of all members of the conversation is decisive (MacFarlane, 2014). On the latter non-solipsistic version of Contextualism, "might *p*" is closer in meaning to "for all we know, *p*".

Given this account of the propositional content of might-claims, this explains why, according to Contextualism, Might*A* is properly assertable if *A* is consistent with the speaker's beliefs. For $|\text{Might}A|^{cu}$ is true if *A* is consistent with the speaker's beliefs. It also explains why it is not even *reasonable* to assess a might-claim relative to the information state of the *assessor*: for the assessor would then be assessing the wrong proposition. In the context of assessment c_a, the proposition expressed by Might*A*, $|\text{Might}A|^{ca}$ = the set of worlds in which *A* is consistent with the *assessor's* beliefs. It would be like the following exchange. Jane asserts "My name is Jane" whereby Bill responds by "No you are wrong, my name is Bill".

The relativist, by contrast, gives a very different account of the propositional content of a might-claim. On this account the content expressed by an assertion of Might*A* does not depend on the information state and so does not differ from speaker to speaker or from speaker to assessor. On the relativist account, an assertion of Might*A* does not attribute any factual property to the world, it does not represent the state of the world in any objective sense. Accordingly, it's content cannot be represented as a set of possible worlds. The non-representational content of Might*A*, in symbols $|\text{Might}A|$, can instead be represented as the set of pairs of worlds and information states (w, X) such that $[\![\text{Might}A]\!]^{w,X} = 1$.

How can such a non-factual proposition be true or false? Here the context of assessment comes in: the truth value of $|\text{Might}A|$ depends on the context of assessment, specifically, on the information state of the context of assessment. $|\text{Might}A|$ is true in the context of assessment c_a if $|\text{Might}A|^{w_{c_a}, X_{c_a}} = 1$. This means that $|\text{Might}A|$ can be true relative to one context of assessment and false relative to another; even when they occur in the same world. This explains why, according to Relativism, Might*A* can *reasonably* be asserted if *A* is consistent with the speaker's beliefs. For if the context of assessment is the speaker's own context then $|\text{Might}A|$ is true if *A* is consistent with the speaker's beliefs. But it also explains why a speaker who asserts Might*A* can be held culpable—and so be called to retract the assertion—in a context of assessment, where it is known that *A* is false; for in such a context $|\text{Might}A|$ is *false*, and an assertion of a false proposition is grounds for culpability.

The objectivist, finally, can take the same view on the propositional content of a might-claim as the relativist: a might-claim does not attribute any property to the world. However, the objectivist deals with this nonrepresentational character of propositional content by shifting focus from the semantic status of the proposition to its epistemic and normative status. The core principle is that a speaker or assessor can *properly* assert or assess Might*A* as true if and only if *A* is consistent with the speaker's or assessor's information state. Whether an assertion or assessment is proper is thus an objective fact of the matter and does not depend on the information state of whoever assesses the propriety of having made the assertion. If one concedes as much then one must also concede that an assertion can be proper yet false. For if *A* is consistent with what a speaker S knows, then S's assertion of Might*A* is proper. An assessor who knows that *A* is false will *properly* judge S's assertion as false, but must still acknowledge that S's assertion was proper. With the idea that this combination of proper but false assertions is made possible by the non-representational character of modal content. No context (of use or assessment) provides any *semantically privileged* information state that determines whether the proposition expressed by the might-claim is true. The truth value of a might-claim is—in the sense in which truth-values form grounds for a normative assessment of an assertion—indeterminate. While one should judge a might-claim as true or false depending on one's information state, there is nothing so semantically special about one's own information state that one can use it as the basis for normative assessment. This means that the *normative* assessment of an assertion (was the assertion proper or should it be retracted as improper?) need not follow the *semantic* assessment of an assertion (was the assertion true or false?).

How one will assess the truth value of a modal assertion will depend on one's information state. While there is no *semantically* privileged information state there is an *epistemically* privileged information state: the state in which one knows all relevant facts. Assuming that it is in principle knowable whether *A* is true or not, |Might*A*| would be assessed true relative to the *best* information state (one in which one knows all relevant facts) if and only if |*A*| is true. If, as Peirce suggested, *truth lies at the end of enquiry*, this establishes a kind of objective notion of truth for epistemic modals. If *w* is the actual world then the ideal information state at the end of inquiry is the singleton set {*w*}. On this conception of objective truth for epistemic modals, |Might*A*| is objectively true in a world *w* iff (*w*, {*w*}) is an element of |Might*A*| if and only if *A* is true at *w*. Whether a modal assertion has the property of objective truth or falsity is an entirely objective matter. But this property of objective truth/falsity is to some extent inert: it has neither semantic significance (the conditions of objective truth do not determine semantic content), nor does it enter into conditions of proper assertion (one can *properly*, and so without fear of culpability, assert Might*A* even though the assertion is objectively false). Indeed the only reason for calling the property in question "objective *truth*", giving it a privileged status, is that a modal proposition will have this property if and only if anyone who knew all the facts would judge it true. Indeed, once one knows all relevant facts, including, say, that *A* is false, it makes no sense to say that |Might*A*| *was* true before we learned that *A*: once one knows all the facts one should judge that it was false all along.

As Table A2 shows, there are cases which regulate which assertions can be made, and which challenges are appropriate, on which interlocutors can agree, even if they differ on the underlying interpretation of epistemic modals. Thus, despite the uncertainty of whether one is communicating with a person who evaluates might-statements according to Contextualism, Relativism, or Objectivism, it is possible to follow these rules for arguing over might statements and coordinate which challenges need to be addressed by the speaker.

Table A2. Argumentation in a Mixture Distribution of Different Interpretations of Epistemic Modals without the Norm of Retraction

	Time	Information State	Assertability	Challenges
Case 1	t_1	$i \cap p \neq \emptyset$ *p* is not excluded by a shared information state.	Might-*p* is assertable.	1) The speaker is warranted in rejecting challenges to her assertion of might-*p*. 2) The interlocutor can use factual evidence to challenge the speaker to update the shared information state so that *p* is excluded and might-*p* is not assertable.
Case 2	t_2	$i \cap p = \emptyset$ The shared information state has become updated, *p* is now excluded.	Might-*p* is not assertable.	1) The interlocutor can warrantedly challenge assertions of might-*p*. 2) The interlocutor can use factual evidence to challenge the speaker to update the shared information state so that *p* is not excluded and might-*p* is assertable.
Case 3	t_2	$i \cap p = \emptyset$ The shared information state excludes *p* at t_2. At t_1 information accessible to the speaker only, but concealed by the speaker, excluded *p*.	Might-*p* is not assertable at t_2. Might-p was not assertable by the speaker at t_1.	1) The interlocutor can warrantedly challenge assertions of might-*p* at t_2. 2) The interlocutor can require that assertions of might-*p* at t_1 are retracted, because at t_1 might-*p* was not correctly assertable by the speaker; only the speaker's deception made it appear so.

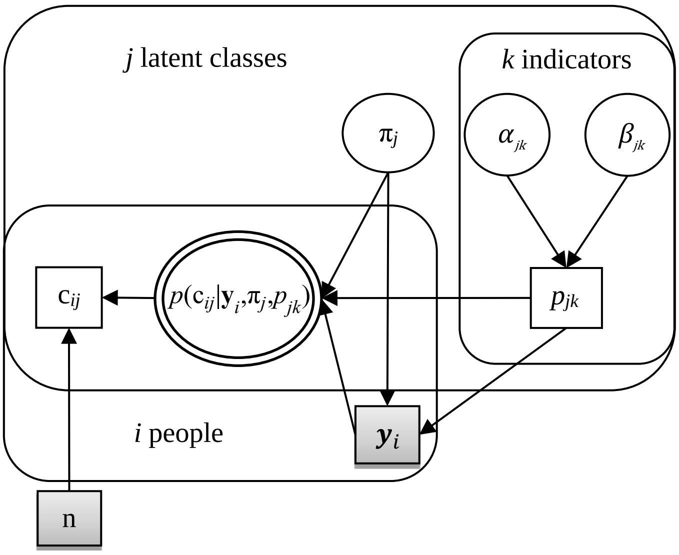
Note. Case 3 operates with a minimal notion of retraction on which Contextualism, Relativism, and Objectivism can agree even if the more controversial norm of retraction of MacFarlane (2014) does not apply. The difference is that on MacFarlane's norm of retraction, the truth value of might-assertions shifts with the context of evaluations and assertions that were correctly made at t_1 can acquire the status of being in need of retraction if evaluated by a context of assessment with a diverging information state. In contrast, Case 3 deals with the case in which the might statement was not even appropriately asserted at t_1 according to the speaker's information state, but the speaker concealed this fact to her interlocutors.

Notice moreover that these cases do not rely on the norm of retraction, proposed in MacFarlane (2014) to govern argumentation with epistemic modals. Thus, cases 1-3 in Table A1 illustrate how it is possible to account for argumentation with epistemic modals, even without the norm of retraction.

Appendix 2: Bayesian Latent Class Analysis

Table 2A specifies the Bayesian latent class model used for the phase 2 classification in the Scorekeeping Task. The model follows the tutorial in Li et al. (2018), which thoroughly explains the mathematical components of the model and its implementation in R.

Table 2A. Bayesian Latent Class Analysis



π_j ~ Dirichlet(1, 1, 1)

c_{ij} ~ Multinomial($p(c_{ij}|\mathbf{y}_i,\pi_j,p_{jk})$, n)

p_{jk} ~ beta($\alpha_{jk} = 1, \beta_{jk}=1$)

$$\mathbf{y}_i \sim \sum_{j=1}^{C} \pi_j \prod_{k=1}^{K} \text{Bernoulli}(p_{jk})$$

$$\mathbf{y}_i, c_{ij} \sim \prod_{j=1}^{C} \left[\pi_j\, p\left(\mathbf{y}_i|p_{jk}\right)\right]^{c_{ij}}$$

$$p(c_{ij}|\mathbf{y}_i, \pi_j, \mathrm{p}_{jk}) = \prod_{j=1}^{C} \left[\frac{\pi_j\, p\left(\mathbf{y}_i|\mathrm{p}_{jk}\right)}{\sum_{j=1}^{C} \pi_j\, p\left(\mathbf{y}_i|\mathrm{p}_{jk}\right)}\right]^{c_{ij}} \quad \text{(Eq.1)}$$

$$p(\pi_j,\mathrm{p}_{\mathrm{jk}}|\mathbf{y}_i,\mathbf{c}_i) \propto \prod_{i=1}^{N}\prod_{j=1}^{C} \left[\pi_j\, p\left(\mathbf{y}_i|p_{jk}\right)\right]^{c_{ij}} \prod_{j=1}^{C} \pi_j^{\,u_j\text{-}1} \prod_{k=1}^{K} p_{jk}^{\;\alpha_{jk}\text{-}1} \left(1 - p_{jk}\right)^{\beta_{jk}\text{-}1} \quad \text{(Eq.2)}$$

Note. $\mathbf{y}_\mathrm{i}$ is a data vector of the *k* indicator values for the *i*th participant. (Eq.1) calculates the conditional probability of each participant's latent class membership. (Eq. 2) states that the joint posterior distribution of class membership, π_j, and item response probability, p_{jk}, is proportional up to a normalizing constant to the product of the latent class weighted Bernoulli likelihood for the *k* indicator variables, the Dirichlet prior for latent class membership, and the beta prior for item response probability. $\mathbf{c}_i$ is a vector representing an assignment of a latent class to the *i*th participant, which may take the form of (1,0,0) for membership of the first class. In this case, the likelihood of observing $\mathbf{y}_\mathrm{i}$ from the first participant, who is assigned the first latent class, would be based solely on π_1 and p_{1k} (as the likelihood based on the second and third latent classes would yield a constant of 1 in virtue of being raised to a power of 0). Similarly, the likelihood of observing $\mathbf{y}_\mathrm{i}$ given membership to the other two latent classes is calculated by setting $\mathbf{c}_i$ to (0,1,0) and (0,0,1), respectively (where the likelihood based on the first latent class is raised to a power of zero). This relationship between $\mathbf{y}_\mathrm{i}$ and the latent class assignments, c_{ij}, is not captured by the plate diagram (*left side*). Yet, it is expressed in the box outlining the distributions (*right side*) by putting both $\mathbf{y}_\mathrm{i}$ and c_{ij} on the left-hand side in the likelihood function for $\mathbf{y}_\mathrm{i}$. An alternative way of stating the likelihood function, where the overall likelihood is summed over the *c* = 1,2, …, *C* latent classes, and is a function of just π_j and p_{jk}, is stated right above it. See Li et al. (2018) for an explanation of the further technical details of the model.

The model was fitted in a Bayesian framework through a Gibbs sampler, which estimates the posterior distributions of model parameters by means of Monte Carlo-Markov chains. A solution with three latent classes was fitted to the data and posterior item response probabilities, p_{jk}, were estimated along with the frequency of the classes in the population, π_j.

The latent classes are estimated based on commonalities in how the members of each latent class responded to the items in the Scorekeeping Task, which consist of *k* indicator variables that produce a data vector, $\boldsymbol{y}_i$, for each participant, *i*.

The likelihood of observing $\boldsymbol{y}_i$ is here modelled as a Bernoulli likelihood with item response probability p_{jk}, which is weighted by the distribution of latent class membership in the population, π_j, and made specific to an assignment of a latent class to participant *i* ($\mathbf{c}_i$). Accordingly, different probabilities of endorsing the *k* indicator variables will be estimated for members of the three latent classes. The likelihood of observing the responses, $\boldsymbol{y}_i$, of participant *i* can then be computed as a weighted average of the likelihood of observing these responses within each latent class, which is weighted by the estimated frequency of the latent classes in the population. Samples that assign participant *i* to a particular latent class (e.g., the first class in case of $\mathbf{c}_i = (1,0,0)$) are drawn from a multinomial distribution based on the conditional probability that participant *i* belongs to each of the three latent classes calculated via (Eq.1). By applying Bayes' Theorem, the joint posterior distribution of π_j and p_{jk} given $\boldsymbol{y}_i$ and $\boldsymbol{c}_i$, is estimated via (Eq.2).

Conjugate priors are specified for p_{jk} and π_{j} to make sure that the prior and posterior distributions are in the same distribution family. For p_{jk}, a beta distribution was used as a conjugate prior for proportions, which was made non-informative by setting the shape parameters (α_{jk}, β_{jk}) equal to 1. For π_j, a Dirichlet distribution was used as a conjugate prior for categorical distributions, which was made non-informative by setting the prior frequency of each of the three latent classes equal to 1. Via these conjugate priors, and the likelihood function of the model, the joint posterior distribution of π_j and p_{jk} given $\boldsymbol{y}_i$ and $\boldsymbol{c}_i$, can likewise be defined in terms of Dirichlet and Beta posterior distributions, as shown in Li et al. (2018).

Chapter 9:
Invariance Violations and the CNI Model of Moral Judgments[121]

Under review in

Personality and Social Psychology Bulletin

Niels Skovgaard-Olsen,
Karl Christoph Klauer

A number of papers have applied the CNI model of moral judgments to investigate deontological and consequentialist response tendencies (Gawronski et al., 2017). A controversy has emerged concerning the methodological assumptions of the CNI model (Baron & Goodwin, 2020, 2021; Gawronski et al. 2020). In this paper, we contribute to this debate by extending the CNI model with a skip option. This allows us to test an invariance assumption that the CNI model shares with prominent process-dissociation models in cognitive and social psychology (Klauer et al., 2015). Like for these process-dissociation models, the present experiment found violations of the invariance assumption for the CNI model. In addition, we show via structural equation modelling that previous findings for the relationship between gender and the CNI parameters are completely mediated by the association of gender with primary psychopathy. This analysis thereby extends several studies that have used the CNI model to investigate the relationship between psychopathy and moral judgments. Finally, we present recommendations for future use of the CNI model in light of our results.

[121] This chapter is under review:

Skovgaard-Olsen, N. and Klauer, K. C. (*Personality and Social Psychology Bulletin*). Invariance Violations and the CNI Model of Moral Judgment.

Authors' Note: Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).
Acknowledgement: We would like to thank Jonathan Baron for comments and discussion.

9.1 Introduction

Considerable parts of moral psychology have focused on the opposition between deontology and utilitarianism in sacrificial dilemmas featuring a run-away trolley (Waldmann et al., 2012). In these dilemmas, participants need to weigh the consequences (e.g., save 5 lives), which figure in utilitarian cost-benefit calculations, against the preservation of deontological moral principles prohibiting to kill other people intentionally.

Further methodological improvements lead to the development of two models of moral judgment based on the family of multinomial processing tree (MPT) models: the process-dissociation (PD) model (Conway & Gawronski, 2013) and the CNI model (Gawronski al., 2017). These models are applied to various real-world moral dilemmas, which introduce further controls for confounds than the run-away trolley scenario, as outlined below.

In this paper, we re-examine a recent controversy concerning the latest development of the CNI model, which unfolded between Baron and Goodwin (2020, 2021) and Gawronski et al. (2020). Our focus will be on the soundness of an invariance assumption concerning the estimation of MPT parameters (like in the PD model and the CNI model), which has proved to be problematic in applications of process-dissociation models in cognitive and social psychology (Klauer et al., 2015).

Through an experiment with a large sample size (N = 486), we test this invariance assumption as applied to the CNI model and evaluate an extension of the CNI model, which avoids making the invariance assumption. Based on this model comparison, we re-examine previously reported effects concerning psychopathy and the CNI parameters via structural equation modelling to assess how effects of gender on the CNI parameters are mediated. Finally, we make a recommendation for how the CNI model should be applied in future uses based on our results.

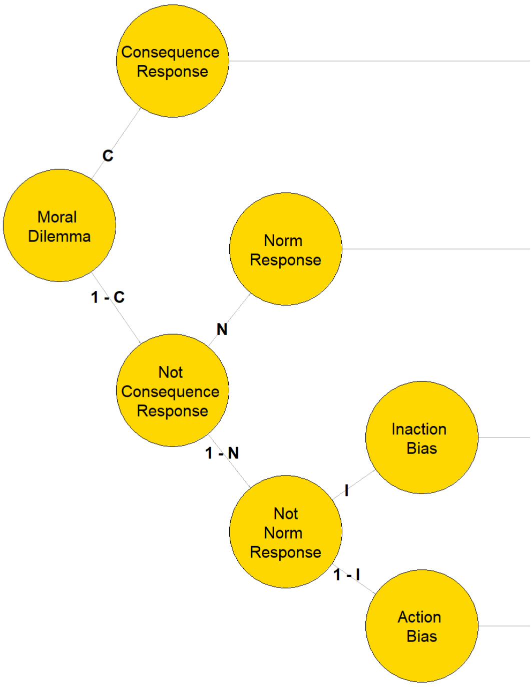
The CNI Model

To refine the classification of norm based and consequentialist moral judgments, a process dissociation model (Conway & Gawronski, 2013) and a multinominal processing tree (MPT) model (Gawronski et al., 2017) have been developed to disentangle factors that are not separated in the traditional, sacrificial dilemma. In Conway and Gawronski (2013), this is done by producing both congruent and incongruent conditions in which the benefits of action can be either smaller or greater than the cost of the outcome. In Gawronski et al. (2017), this takes the form of developing new stimulus materials that factorially combine action/inaction according to deontological norms and utilitarian consequences based on the insight that these two factors are confounded in the run-away trolley dilemma. Since a deontological response always requires inaction in standard sacrificial dilemmas, where victims are fixed to the tracks of a run-a-away trolley, it cannot be separated from a general response bias towards inaction. Moreover, since the utilitarian response always requires action in standard trolley dilemmas, it cannot be separated from asocial tendencies towards sacrifice.

In these improved scenarios, four conditions are created in which the cost of the outcome is either greater or smaller than the costs and the norms are manipulated to either prohibit an action to bring about the outcome (*proscriptive norm*) or to prescribe an action (*prescriptive norm*) in a situation in which some other agent plans to carry out a prohibited action. The scenarios describe realistic situations in which the sacrificial dilemma, for instance, arises in the context of a doctor treating patients. In this context, a norm-based response pattern (N) consists in a) selecting inaction, whenever actions are prohibited by deontological norms not to harm other people, and b) selecting action, whenever the action is to prevent another agent from carrying out the prohibited act. In contrast, the consequentialist response pattern (C) consists in selecting action if and only if the consequences are greater than the costs. Finally, a response bias towards inaction (I) consists in selecting inaction across all conditions without regard to variations in norms or the utility of the consequences.

The relative response frequencies are analyzed with a multinomial processing tree model (Batchelder & Riefer, 1999; Erdfelder et al., 2009) to characterize the processes underlying participants' selections of categorical outcomes. The processing tree contains three parameters (C, N, I) that represent the estimated probability that the observed response was based on the manipulated consequences, moral norms, or a general response bias for inaction, as illustrated in Table 1.

Table 1. CNI Model



	Proscriptive Norm		Prescriptive Norm	
	Benefits Greater	*Benefits Smaller*	*Benefits Greater*	*Benefits Smaller*
Consequence Response	Action	Inaction	Action	Inaction
Norm Response	Inaction	Inaction	Action	Action
Inaction Bias	Inaction	Inaction	Inaction	Inaction
Action Bias	Action	Action	Action	Action

Note. Illustration of the CNI model based on Gawronski et al. (2017).

Based on the tree structure in Table 1, equations for action and inaction responses are formulated for each of the four CNI conditions by multiplying the parameters along a path leading to a response and adding all paths leading to the same response. For instance, an action response in the prescriptive condition, where the benefits are greater than the costs, may either arise by reacting to the consequences [C], or by reacting to the norms given that the response is not produced by a reaction to the consequences [(1-C)×N], or by an action bias to always select action given that the response is neither produced by a reaction to the consequences nor to the norms [(1-C)×(1-N)×(1-I)]. Accordingly, p(action|prescriptive norm, benefits > costs) = C + [(1-C)×N] + [(1-C)×(1-N)×(1-I)].

Since action and inaction are complementary response options, Gawronski et al. (2017) formulate four non-redundant equations that quantify the probabilities of selecting an action and inaction, respectively, across the four CNI conditions. Modelling the responses as coming from a multi-nominal likelihood distribution with response probabilities given by the model equations, the three model parameters can be estimated via either maximum likelihoods methods or Bayesian statistics. It is then tested whether C and N parameters differ from 0 and whether the I parameter diverges from 0.5, via 95% confidence intervals or credible intervals, respectively.

While previous studies with the traditional sacrificial dilemma have indicated a positive correlation between psychopathic traits and utilitarian sacrifices (Marshall et al., 2018), one of the interesting findings of the CNI model is that its parameters tend to be negatively correlated with psychopathic traits (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021; Luke et al., 2021). While individuals high in psychopathy may have less restrictions towards sacrifice of human life, this result indicates that they also tend to be less influenced by the difference of whether sacrifice occurs when the benefits for the greater good are larger versus smaller than the costs of the outcome. More recently, the CNI model has been further extended to permit the study of individual differences by assessing its parameters at the individual level (e.g., Kroneisen & Heck, 2020; Körner et al. 2020).

The Invariance Assumption

Since only three MPT parameters are used to parameterize the multi-nominal likelihood distribution, one of the assumptions of the model is that the N parameter stays invariant across proscriptive and prescriptive norms, and that the C parameter stays invariant across the four CNI conditions. In other words, the model assumes that the strength of deontological norms is invariant to whether the norms forbid doing a questionable action (e.g., killing someone) or whether the norms prescribe interfering with the actions of someone else to prevent an action (e.g., preventing someone else in killing someone). Similarly, the model assumes that the probability of judging a questionable action with desirable consequences (e.g., saving lives) acceptable on utilitarian grounds is the same as judging the probability of the same action unacceptable on utilitarian grounds when its consequences are less desirable (e.g., averting only a minor damage).

The CNI model is not alone in making this type of invariance assumption. The type of process-dissociation models that is used in Conway and Gawronski (2013) as a predecessor to the CNI model similarly makes an invariance assumption in its model equations. More generally, process-dissociation models form a subset of the class of multinominal processing-tree models. In Klauer et al. (2015), it was tested empirically whether prominent instances of process-dissociation models from cognitive psychology (Stroop task, cued recall) and social psychology (racial bias in the weapon task) violated the invariance assumption. In several instances strong violations were found and it was conjectured that similar violations of the process-dissociation model of Conway and Gawronski (2013) would occur as well.

What are the consequences of violations of the invariance assumptions? As discussed by Klauer et al. (2015), such violations have the potential to compromise estimates of the model parameters and substantive conclusions drawn from them. Moreover, traditional analyses using the CNI model, that is premised on the invariance assumptions, unfortunately do not allow one to detect such violations, nor to assess the extent of distortions that may ensue from violations of invariance.

Via the addition of proscriptive and prescriptive norms, the CNI model improves upon the process-dissociation model of Conway and Gawronski (2013). Over and above the methodological issues surrounding the invariance assumptions, these assumptions are also at the root of a recent controversy surrounding the CNI model, however.

Controversy Surrounding the CNI Model

In a recent critical exchange between Baron and Goodwin (2020, 2021) and Gawronski et al. (2020), the CNI model was criticized on several counts. Some of these points could be addressed by the rebuttal in Gawronski et al. (2020) – in particular those concerning order-effects, the interpretation of the model and its parameters – but other points still stand.

Baron and Goodwin (2020, 2021) worry that the scenarios used to apply the CNI model leave interpretational ambiguities, which may help explain the high rates of so-called "perversive responses", where participants select responses that go against both deontological and utilitarian responses in congruent conditions, where both predict action (PreGreater) or inaction (ProSmaller). They argue that this makes the CNI scenarios unsuitable for studying inaction bias.

One of the other central arguments that Baron and Goodwin (2020, 2021) make is that the reason why deontological responses have previously been investigated mainly through inaction is that deontological norms prohibiting harmful action (e.g., "first, do no harm") are stronger than norms proscribing action to do good. Similarly, other researchers have studied asymmetries between a strict, duty-based system of proscriptive, moral regulation, which is focused on blame and identifying transgression versus a prescriptive regulatory system, which is focused on credit-worthy good deeds that is more desire-based and less strict (Janoff-Bulman et al. 2009). Janoff-Bulman et al. analyze the distinction between these two types of moral norms as being based on asymmetries between two motivational systems that are occupied with withdrawing from negative outcomes, or overcoming negative desires, and approach-behavior aiming at achieving positive end-states, respectively.

Yet, the distinction between proscriptive and prescriptive norms works differently in the context of the CNI dilemma, where there are always two bad outcomes and conflicting motivations for choosing between action/inaction. So, the idea of a motivational system aiming at positive end-states from Janoff-Bulman et al. (2009) does not directly carry over to the prescriptive norms in the context of the CNI scenarios, and it is thus slightly misleading when the two lines of research are compared without further qualification (e.g., in Henning & Hütter, 2020). At the root of this difference is the fact that the CNI model mainly introduces prescriptive norms to solve the methodological problem of avoiding a possible confound between an inaction chosen from deontological reasons and a general bias towards inaction, whereas Janoff-Bulman et al. (2009) develop a particular substantive interpretation of prescriptive norms.

In the context of the CNI implementation of prescriptive norms, a deontologically prohibited action is planned by another agent and the participant can choose to intervene to prevent this from taking place. In this way, the same action choice between two bad outcomes is presented in a configuration in which one of the actions has already been preselected by another agent. As Baron and Goodwin (2020) point out, this may lead to a weaker prescriptive norm, if this other person is a colleague or superior, as in some of the CNI scenarios, since it introduces further, unintended consequences such as the following:

> when the action is to contravene someone else’s action, it has additional consequences aside from preventing the consequences of that action. It may hurt the decision maker’s feelings, possibly leading him or her to take retaliatory action against the one who contravenes. It may also violate the lines of authority, thus weakening these lines for the future by discouraging those in command from taking their responsibility seriously (Baron, 1996). It may also be illegal or against the rules, and rule following likewise has a value as a precedent for future cases. (p. 424)

Accordingly, Baron and Goodwin (2020, 2021) argue that that the CNI scenarios do not succeed in keeping the relative strength of the deontological norms constant across the proscriptive and prescriptive conditions. Moreover, since unintended consequences are introduced by the way that prescriptive norms are manipulated, Baron and Goodwin (2020, 2021) also suggest that the consequences are not held constant across the two conditions. As they say (2021: 16): “the inferential problem results both from the difference in the norms between the two alternatives presented, as well as the difference in the consequence”. Both points lead to predictions of violations of the invariance assumption of the CNI model. Baron and Goodwin’s (2020, 2021) arguments most strongly suggest that the invariance assumption for parameter N should be violated so that $N_{Pro} > N_{pre}$. *A priori*, a case could, however, also be made for the converse violation with $N_{Pro} < N_{pre}$. For example, scenarios with prescriptive norms ask whether a proposed non-normative action should be thwarted and raising this very possibility of averting the action may in itself act as a clue for participants suggesting that the action is to be considered problematic and should indeed be refused. In an experiment, we set out to test the invariance assumption for parameters N and C.

9.2 Experiment: The Invariance Assumption

To investigate the invariance assumption of the CNI model, we conducted an experiment following the procedure of Klauer et al. (2015), which was used to test violations of the invariance assumption in process-dissociation models. To this end, the MPT equations of the CNI model are implemented in a Bayesian framework via hierarchical latent trait model proposed in Klauer (2010), which has also been applied to study individual variation in the context of the CNI model in Kroneisen and Heck (2020).

Since further degrees of freedom are needed to estimate separate parameters for N in the proscriptive and prescriptive condition and for C across all four CNI conditions, the model was extended via a skip option ("S"), whereby participants could opt out of selecting action/inaction in a given scenario. The MPT equations of this CNIS model are stated in Appendix A. The addition of this skip option was further motivated by reading participants' open-ended responses in Berentelg's (2020) replication study, where it was found that a sizable minority of participants complained about the exclusion of alternative courses of action in particular scenarios. Accordingly, if participants find the scenario ambiguous or the stipulation of the choice situation artificial (with neither C nor N favoring a unique choice, because information has been left out), they are permitted to skip the scenario via this extension of the CNI model.

The interpretation of the skip option is grounded in the logic of the CNI model which is our point of departure. According to that model, when consequences are activated (with probability C), the response is determined by consequences with probability 1, whether or not norms are activated and whether or not the dilemma is congruent or incongruent. When consequences are not activated (with probability 1-C), but norms are activated (with probability N), then the response is determined by norms with probability 1, whether or not the dilemma is congruent or incongruent. And thus, when consequences or norms are activated, the response is deterministically captured by all-or-none processes with consequences dominating norms.

Only when neither consequences nor norms are activated (with probability (1-C) × (1-N)) are responses not deterministic. In this state of uncertainty, participants, metaphorically speaking, throw a loaded dice which comes up with "inaction" with probability (I) and with "action" with probability (1-I). Given that participants were explicitly instructed to use the skip option in the case of uncertainty, this state of uncertainty, reached with probability (1-C) × (1-N), is the only place in the model in which skipping can come into play. Basically, the extension of the CNI model we present provides participants with a third face on their loaded dice, which now shows the faces "action", "inaction", and "skip". Because in the state of uncertainty, neither norms nor consequences are activated, it also makes sense that the I parameter and the S parameter do not depend upon type of dilemma, because the four CNI conditions are distinguished solely in terms of differences in norms and consequences.

Note that the CNI model does not have a mechanism by which participants are sensitive to the difference between congruent and incongruent scenarios. Conflicts between norms and consequences are simply pre-empted by the dominance of consequences over norms built into the model: Whenever norms and consequences are activated, consequences compel the response. In particular, the probability of reaching the state of uncertainty always equals (1-C) × (1-N), whether or not the dilemma is congruent or incongruent.

It could, however, be argued that contrary to the conflict insensitivity built into the CNI model, a state of uncertainty is reached more often for incongruent than congruent dilemmas and that the skip option provides a means of making such differences in response certainty visible. This is, however, just another way to argue for a violation of the invariance assumption. The product (1-C) × (1-N) for the probability of reaching the uncertainty state would then not be invariant across the four CNI conditions, implying that C and/or N cannot be invariant. In particular, if we were to find that C and/or N are depressed (and hence, (1-C) × (1-N) increased) for incongruent dilemmas relative to congruent dilemmas, this would suggest that an experiential state of perceived conflict enhances the probability of reaching the uncertainty state—a possibility that is hidden when the invariance assumption is imposed *a priori*, as in the CNI model.

Through our experiment, we test whether such invariance violations occur through the addition of the S parameter to the CNI model (see Appendix A for further details).

9.2.1 Method

OSF link:

https://osf.io/569bv/?view_only=8b41d5a9229c4a31911f66a594d30a5a

Sampling Procedures

To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002). The following *a priori* exclusion criteria were used: not having English as native language, completing the task in less or more than the average response time ± 2 × SD, failing to answer at least one of two simple SAT comprehension questions correctly in a warm-up phase, and answering ‘not serious at all’ to the question 'how serious do you take your participation' at the beginning of the study.

Participants

The experiment was conducted over the Internet through the platform Mechanical Turk to obtain a large and demographically diverse sample. A total of 778 people finished the experiment. The participants were paid a small amount of money for their participation and sampled from USA, UK, Canada, and Australia. After applying our *a priori* exclusion criteria, the final sample consisted of 486 participants. Mean age was 39.10 years, ranging from 19 to 76.[122] 54.11% of participants identified as male; 45.24% identified as female; 5 participants preferred not to identify with either category. 84.06 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a within-participants design with the following factors varying within participant: Consequence (Smaller vs. Greater) and Norm (Proscriptive vs. Prescriptive). To allow for 10 trial replications for each of the four CNI conditions, each participant in total went through 40 within-subject conditions.

Materials and Procedures

Participants were presented with the four CNI conditions across 10 scenarios adopted from Gawronski et al. (2017) and Körner et al. (2020).[123] The scenarios were modified slightly so that the proposed action was the *intended* means needed to realize the described outcome uniformly across all scenarios. In contrast, in the existing CNI stimulus materials, the manipulated consequence sometimes constituted a foreseeable consequence.[124]

[122] We are here ignoring one occurrence of the value ‘5’ given that Mechanical Turk limits the participation to adults.

[123] The complete list of scenarios can be found at the osf project page cited above.

[124] This distinction matters for the Doctrine of Double-Effect in deontological ethics (McIntyre, 2019).

The order of the scenarios and the CNI conditions within scenarios were randomized for each participant anew.[125] Because different versions of the scenario look similar, the randomization was constrained so that different versions of the same scenario could not occur in immediate succession. Following Gawronski et al. (2017), participants were given the following instruction:

> On the following pages you will see 40 scenarios that people may come across in life. Please read them carefully. Even though some scenarios may seem similar, each scenario is different in important ways. After each scenario, you will be asked to make a judgment about whether you find the described action acceptable or inacceptable. Please note that some scenarios refer to things that may seem unpleasant to think about. This is because we are interested in people's thoughts about difficult, real-life issues.

In addition, participants were instructed that they could "skip" a moral decision for cases, where they were undecided about whether the described action was morally acceptable or inacceptable. They were also instructed that they should not make use of this option more than 10 times. One of the scenarios is illustrated in Table 2 below.

[125] Garownski et al. (2017) use a pseudo-random order, which is fixed to be the same for each participant. In pilot studies, we did not find differences between this procedure and the more rigorous randomized order and chose the latter instead.

Table 2. CNI Conditions

ProGreater	*ProSmaller*	*PreGreater*	*PreSmaller*
You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 500 lives by preventing spread of the virus.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 100 lives by preventing spread of the virus.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 500 lives by preventing spread of the virus. The plan is to use the vaccine, but you could veto this.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 100 lives by preventing spread of the virus. The plan is to use the vaccine, but you could veto this.
Is it acceptable in this case to infect 100 people?	Is it acceptable in this case to infect 100 people?	Is it acceptable in this case to veto the infection?	Is it acceptable in this case to veto the infection?
Yes, it is acceptable vs. No, it is not acceptable vs. Skip			

Note. Example of a CNI scenario.

Finally, some demographic questions were asked and participants' level of psychopathy was probed via Levenson et al.'s (1995) subscale for primary psychopathy in a noninstitutionalized population.

9.2.2 Results

For the analysis, we first fitted the original 4 MPT parameter version of the CNIS model (CNIS$_4$) to ensure construct validity after the addition of the skip option to the CNI model. In this analysis, we test whether the CNIS model is able to replicate the mean pattern observed for the model parameters as well as bivariate associations with external parameters (primary psychopathy, gender) reported in Gawronski et al. (2017).

Following this analysis, a 8 MPT parameter version of the CNIS model (CNIS$_8$) was fitted with 2 separate N parameters (N$_{pro}$, N$_{pre}$) and four separate C parameters (C$_{ProGreater}$, C$_{ProSmaller}$, C$_{PreGreater}$, C$_{PreSmaller}$). This allows us to test for violations of the invariance assumption. Finally, we extend these findings by fitting two structural equation models (SEM) to investigate whether the replicated gender effects are mediated through the association of gender and primary psychopathy.

For a Bayesian implementation of the MPT models, we followed the hierarchical extension of multinominal processing trees in Klauer (2010), which has also been implemented in the R package TreeBUGS (Heck et al., 2018). One of the benefits of the latent trait approach to MPT modelling proposed in Klauer (2010) is that its hierarchical structure makes it well-suited to estimate individual CNI parameters for each participant (Kroneisen & Heck, 2020). In addition, the individual MPT parameters are estimated through a multivariate normal distribution with a covariance structure that permits correlations among the individual MPT parameters, instead of stipulating *a priori* that they must be uncorrelated along the form of the Beta-MPT approach (Smith & Batchelder, 2010). We illustrate the hierarchical latent trait model of Klauer (2010) in Appendix A. The same appendix also states the MPT model equations for the extension of the CNI model with the skip parameter ("S") and explains how the invariance assumption distinguishes CNIS$_4$ from CNIS$_8$.

CNIS with 4 MPT Parameters

Figure 1 displays the distribution of the parameters estimated for each participant:

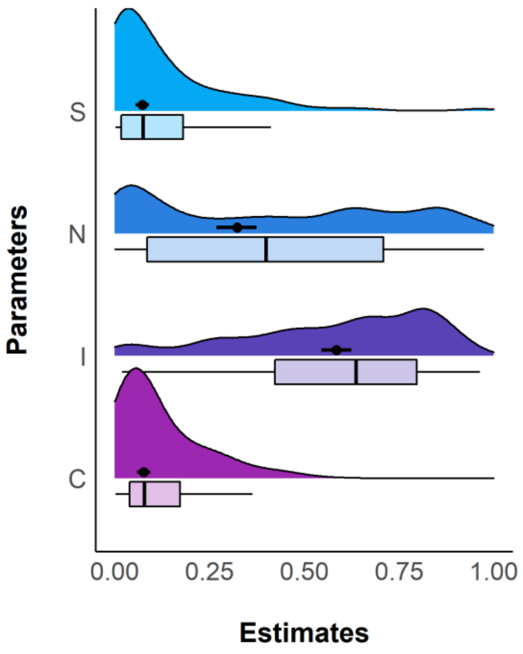


Figure 1. Distributions of the CNIS parameters estimated for each participant with boxplots indicating the quartiles of the individual estimates. The black points and lines indicate the group-level posterior medians and 95% HDI.

As Figure 1 shows, the 95 % HDI[126] for the posterior medians of the C and N parameters exclude zero, and a general bias towards inaction is found, since the 95 % HDI of the posterior median of the I parameter excludes .5.

Since published work reports bivariate correlations, and the first goal is to replicate previous results, we here plot bivariate correlations between the C, N, I parameters and primary psychopathy (P), self-reported gender (G) with ‘male’ encoded as 1, and total response time (T):

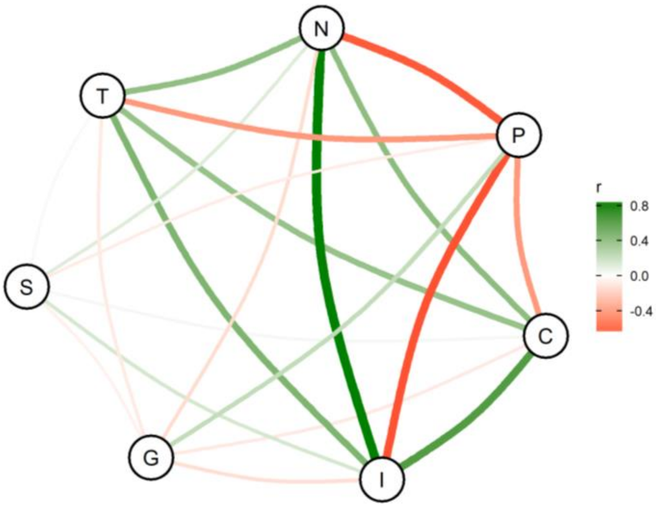


Figure 2. Bivariate associations between model parameters and external variables. ‘G’ = self-reported gender (excluding five participants, who preferred not to respond), ‘P’ = primary psychopathy, ‘T’ = total response time to complete the items, ‘S’ = skip.

As Figure 2 shows, negative bivariate associations between the CNI parameters and psychopathy were found (r_{PC} = -.43, 95% HDI [-.50, -.36]; r_{PI} = -.70, 95% HDI [-.74, -.65]; r_{PN} = -.67, 95% HDI [-.72, -.62]). In addition, Figure 2 shows a positive association between gender and psychopathy (r_{GP} = .23, 95% HDI [.14, .31]) and negative associations between gender and both N (r_{GN} = -.15, 95% HDI [-.23, -.06]) and I (r_{GI} = -.14, 95% HDI [-.23, -.06]).

These results replicate the findings in Gawronski et al. (2017) while adding the S parameter to the CNI model. Below we will use structural equational modelling (SEM) to further analyze mediation relationships in these results. But first we need to find out whether the invariance assumption is violated in the CNI model by contrasting the present model with a 8 parameter version.

[126] A HDI interval is an interval of the posterior distribution where all points within the interval have a higher probability density than points outside it.

CNIS with 8 MPT Parameters

Next, we tested the invariance assumption by fitting separate N and C parameters in a 8 MPT parameter version of the CNIS model. The parameter estimates are displayed in Figure 3 below.

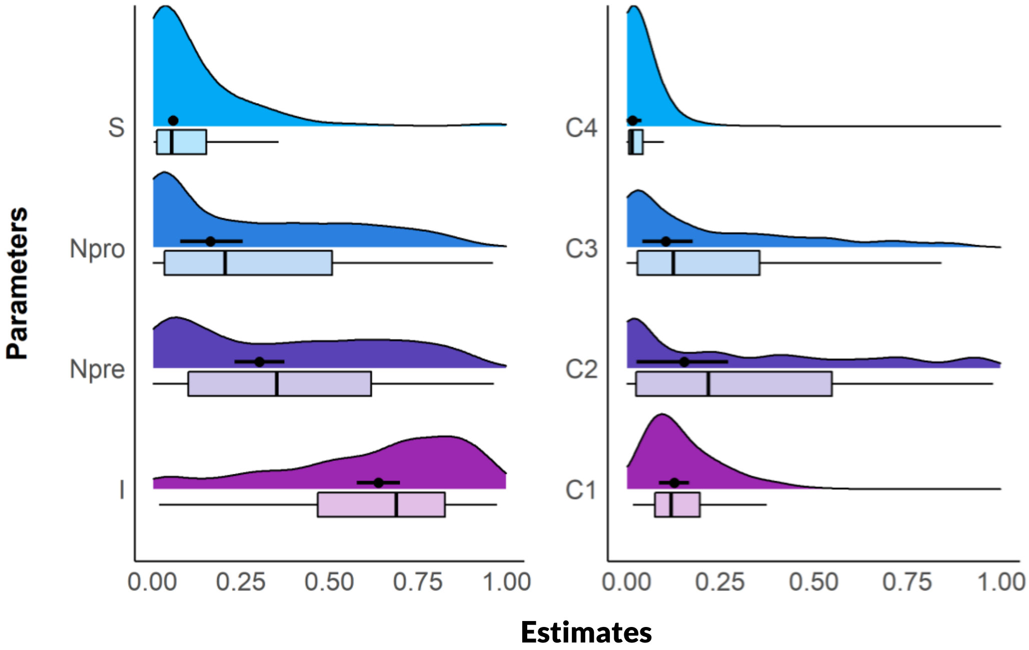


Figure 3. Distributions of the CNIS parameters estimated for each participant with boxplots indicating the quartiles of the individual estimates. The black points and lines indicate the group-level posterior medians and their 95% HDI. 'C1' = $C_{ProGreater}$, 'C2' = $C_{ProSmaller}$, 'C3' = $C_{PreGreater}$, 'C4' = $C_{PreSmaller}$.

To test possible violations of the invariance assumption that Npro = Npre and that C1 = C2 = C3 = C4, we analyzed contrasts of pairs of parameters by identifying whether the 95% HDI intervals for the difference between the two parameters included zero. Credible differences were found for Npre > Npro ($b_{Npo\text{-}Npre}$ = -0.48, 95% HDI [-0.89, -0.06]), C1 > C4 ($b_{C1\text{-}C4}$ = 1.07, 95% HDI [0.45, 1.84]), C2 > C4 ($b_{C2\text{-}C4}$ = 1.15, 95% HDI [0.27, 2.05]), and C3 > C4 ($b_{C3\text{-}C4}$ = 0.95, 95% HDI [0.10, 1.93]).

Finally, we compared the two models in terms of their expected out-of-sample predictive accuracy via information criteria and found $CNIS_8$ to be the better fitting model, as shown in Table 3 below.

Table 3. Model Comparison

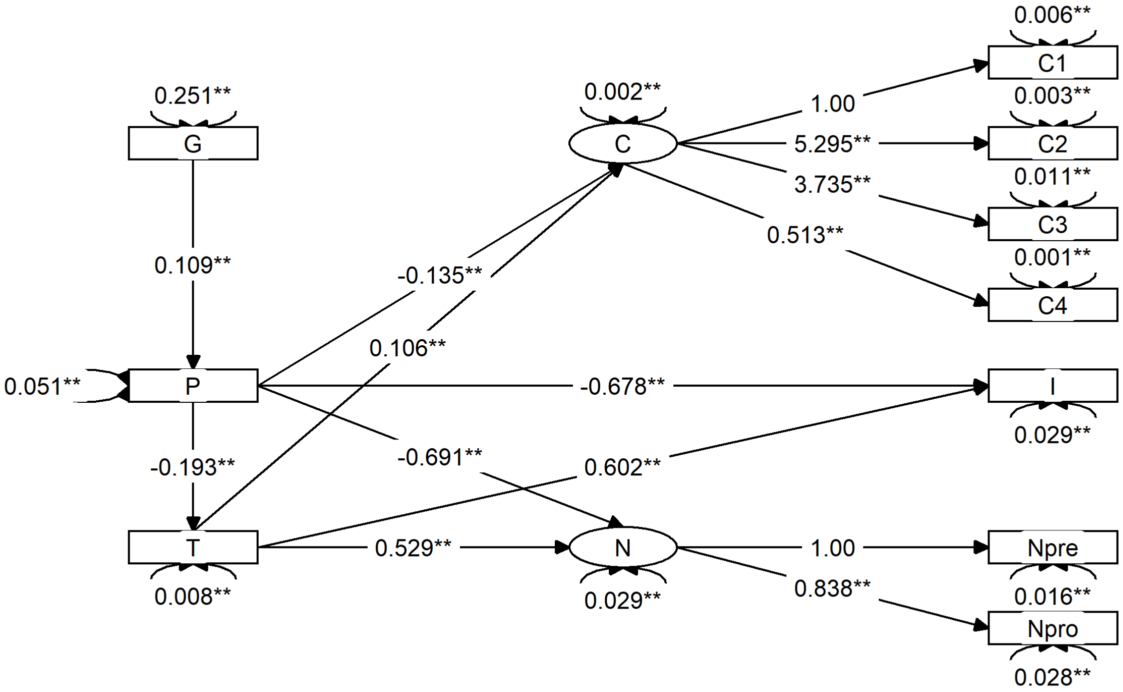
	WAIC	*LOOIC*	*Δelpd (SE)*	p_{T1}	p_{T2}
$CNIS_4$	25474.4	25527.0	-50.5 (12.1)	0.00	0.00
$CNIS_8$	25355.3	25426.0	--	0.04	0.02

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. The test statistics T_1 and T_2 represent Bayesian *p* values and are based on the posterior predictive model checks in Klauer (2010).

In addition, model fit was assessed with the posterior-predicted *p* values based on T_1 and T_2 posterior model checks proposed in Klauer (2010). T_1 measures the adequacy of the models in capturing the mean observed outcome frequencies (aggregated across persons). T_2 measures the adequacy of the models in capturing the variability (variances and covariances) among the observed response frequencies (computed across persons). The proportion with which T_i(observed) < T_i(predicted) are given by Bayesian *p* values. A small *p* value for these test statistics indicates that the posterior predictive distribution of the model fails to capture an aspect of the data. It was found for both the aggregate outcome frequencies and the variability across individuals that both models failed to capture aspects of the data. This is not unusual for large data sets such as the present, but the comparison also shows that $CNIS_8$ performed better than $CNIS_4$.

Structural Equational Modeling

Next, we fitted two SEM models with the R-package `blavaan` (Merkle & Rosseel, 2018) based on the winning $CNIS_8$ model.[127] The two SEM models differed on whether direct paths were included from gender to the CNI parameters (SEM_1) or whether the effect of gender was completely mediated through the effect of primary psychopathy on the CNI parameters (SEM_2), as displayed in Figure 4.



[127] We performed the same SEM analysis on $CNIS_4$, which produced the same qualitative results. Further details can be found in the supplementary materials or on the osf project page: https://osf.io/569bv/?view_only=8b41d5a9229c4a31911f66a594d30a5a

Figure 4. SEM$_2$ model. Path coefficients indicate posterior medians and are marked with '**', if the 95% HDI interval does not include 0. 'G' = self-reported gender (excluding five participants, who preferred not to respond), 'P' = Primary psychopathy, 'T' = total response time for all the items. Both 'P' and 'T' were scaled to take values between 0 and 1 before fitting the model to prevent large differences in the variances of the different parameters. Direct and indirect effects are encoded via arrows. The loops indicate variances. The covariances have been left out to simplify the graph. Latent variables are marked with circles. The scales of the latent variables were fixed by setting the first path coefficient equal to 1.0.

The C and N parameters in Figure 4 are estimated latent variables, which are measured via C1-C4 and Npre/Npro, respectively. The violations of the invariance assumption can be read off from the differences in the path coefficients from the latent C and N parameters to the C1-C4 and Npre/Npro parameters.

Table 4 shows a model comparison between SEM$_1$ and SEM$_2$ as well as a mediation analysis of SEM$_1$, which shows why including direct paths from gender to the CNI parameters does not improve the fit of the model. This in turn creates a slight preference for the SEM$_2$ model displayed in Figure 4.[128] A feature of SEM$_2$ is that its underlying directed acyclic graph (DAG) entails the following conditional independencies, which imply that the partial correlations between gender and the CNI parameters are zero, when controlling for the influence of primary psychopathy.

C _||_ G | P G _||_ I | P G _||_ N | P G _||_ T | P

In words: gender is independent of the C, N, and I parameters when conditioning on primary psychopathy.

Table 4. SEM Models

Model Comparison				
	WAIC	*LOOIC*	*Δelpd (SE)*	*R^2*
SEM$_1$	-7800.96	-7804.00	-1.49 (1.49)	C=.42, N=.56, I=.56
SEM$_2$	-7804.31	-7807.00	--	C=.42, N=.56, I=.56
Mediation Analysis based on SEM$_1$				
	Direct Path G → X	*Indirect Path* G → P → X	*Total Effect:* *Direct + Indirect*	*Proportion* *Mediated*
C	$\tilde{x}$ = -.00 [-.01, .01]	$\tilde{x}$ = -.02 [-.02, -.01]	$\tilde{x}$ = -.02 [-.03, .01]	0.981
N	$\tilde{x}$ = .01 [-.02, .05]	$\tilde{x}$ = -.08 [-.11, -.05]	$\tilde{x}$ = -.06 [-.11, -.02]	1.189
I	$\tilde{x}$ = .01 [-.02, .04]	$\tilde{x}$ = -.07 [-.10, -.05]	$\tilde{x}$ = -.07 [-.11, -.02]	1.140

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. The proportion mediated can take values larger than one in cases where the direct and indirect effects are of opposite signs, as here. The square brackets indicate 95% HDI.

[128] For the corresponding comparison between SEM$_1$ and SEM$_2$ based on CNIS$_4$ reported on the osf project page, the standard error of the elpd difference was lower and the preference for SEM$_2$ was stronger. URL = https://osf.io/569bv/?view_only=8b41d5a9229c4a31911f66a594d30a5a

The results in Table 4 show that the negative correlations between gender and N and I that are found in the bivariate correlations are completely mediated by the effect of gender on primary psychopathy.

A further advantage of structural equation modelling is that it gives a principled way of identifying the minimal adjustment set of covariates that need to be controlled for to avoid spurious correlations (Pearl, 2009; Kline, 2016). Applying the graphical criteria from Pearl (2009) on the underlying DAG in SEM_2, we thus find that associations between the CNI parameters and response time need to control for primary psychopathy to avoid spurious correlations, because P acts as a common cause on the T and CNI parameters in Figure 4.

9.2.3 Discussion

To test the construct validity of the CNIS model, a 4 MPT parameter version was first fitted to the data, and it was tested whether known patterns of means for the CNI parameters and known relations of the CNI parameters to primary psychopathy and gender could be replicated. Like previous work, we found evidence for parameters N and C to be substantively larger than zero, and for the inaction parameter to exhibit a credible bias towards inaction (I > .5). In addition, it was found that the C and N parameters were negatively associated with primary psychopathy and that negative associations between gender and the N and I parameters could be found. This result replicates previous work reporting similar gender effects and negative associations between the model parameters and primary psychopathy with mixed findings concerning a negative association with the C parameter across studies (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021; Luke et al., 2021).

In a second analysis, a 8 MPT parameter version of the CNIS model was fitted to the data and it was found that credible differences between the N and C parameters emerged. This indicates a violation of the invariance assumption. In a further exploratory analysis, we investigated whether the invariance assumption was violated within each item. The item specific estimates of the 8 MPT parameters are reported in Appendix B, and it is found that violations of the invariance assumption occur almost within every scenario tested.

As explained above (Section “The Invariance Assumption”), violations of invariance have the potential to compromise estimates of the model parameters by introducing systematic bias and to invalidate substantial conclusions drawn from them. Comparing the parameter estimates and results pattern for the 4 and 8 MPT parameter versions of the CNIS model suggests that the consequences in terms of substantive conclusions were relatively minor for the present data: Both models yielded roughly similar overall patterns of mean parameter estimates and correlational results. This need of course not be the case for other data sets and situations; there is simply no way to tell unless the model is extended as exemplified here to allow one to estimate separate N and C parameters.

9.3 General Discussion

The CNI model (Gawronski et al., 2017) has advanced the computational modelling of moral judgments by systematically pairing factors that are normally confounded in traditional research on moral judgment via Trolley dilemma. Using multinominal processing trees and scenarios with four contrast cases, the CNI-model attempts to dissociate adherence to utilitarianism and deontology in participants' case judgments.

At the same time, the model is surrounded by controversy concerning its underlying assumptions and their implications for moral psychology (see e.g., Baron & Goodwin, 2020, 2021; Gawronski et al., 2020). Part of the latter controversy implicitly concerns an invariance assumption made by process-dissociation models and MPT models alike, which has been found to be problematic in other domains of psychology in Klauer et al. (2015). For estimating adherence to Utilitarianism and Deontology, the CNI-model assumes that the probability of judging a questionable action with desirable consequences (e.g., saving lives) acceptable on utilitarian grounds is the same as judging the probability of the same action unacceptable on utilitarian grounds when its consequences are less desirable (e.g., averting only a minor damage). Similarly, the model assumes that the strength of deontological norms is invariant to whether the norms forbid doing a questionable action (e.g., killing someone) and whether the norms prescribe interfering with the actions of someone else to prevent an action (e.g., preventing someone else in killing someone).

To investigate this invariance assumption, we compared two hierarchical Bayesian implementations of the CNI model, which differ in whether they assume different or the same parameters for utilitarian and deontological judgments in these contrast cases.

What enabled the estimation of the parameters of the CNI model without making the invariance assumption was extending the CNI paradigm with a skip option and the CNI model by a S parameter ("skip"). It was found through a model comparison that the extended 8 parameter version of the CNIS model, which does not make the invariance assumption, outperformed the 4 parameter version, which differs from it solely by making the invariance assumption (Table 3).

While previous controversy surrounding the CNI model suggests that the invariance assumption would be violated, Baron and Goodwin (2020, 2021) strongly predict that such violations would take the form of $N_{Pro} > N_{pre}$. In contrast, the data show that the violations go in the opposite direction: $N_{Pro} < N_{pre}$. We offered a speculative account for why N_{pre} might be larger than N_{Pro} above based on the idea that presenting the possibility of overwriting the action of another agent pragmatically implicates that the action is to be considered problematic and should indeed be refused. Further violations of the invariance assumption occurred with respect to the C parameter, where it was found that the posterior median of $C_{PreSmaller}$ approaches zero and is reliably smaller than the posterior median of the C parameters in all other conditions. For PreSmaller scenarios, consequentialist choices imply judging refusals to act inacceptable. We suspect that the double negation uniquely implied in understanding and making this particular choice leads to it being adopted only infrequently and to the depressed $C_{PreSmaller}$ parameter. Taken together, the extended CNIS model provides (a) a methodological tool for estimating the CNI parameters without the need for the problematic invariance assumption and (b) suggests interesting new hypotheses (e.g., possible roles for pragmatic implicatures and double negation) and thereby opens avenues for future research when violations of invariance are found.

The distributions of the individual C parameters in Figure 3 moreover indicate that the variance for the C parameter in the congruent conditions ($C_{ProSmaller}$, $C_{PreGreater}$) is larger than the variance for the C parameter in the incongruent conditions ($C_{ProGreater}$, $C_{PreSmaller}$). We refrain from interpreting this finding substantively, however, because it may have to do with the amount of statistical information that is available for estimating the different parameters, and hence the estimation uncertainty expressed in the variances, that may differ between the conflict scenarios and the congruent scenarios.

Finally, based on a structural equation analysis of $CNIS_8$, we were able to show that the previously reported gender effects on the CNI parameters (e.g., Gawronski et al. 2017) were completely mediated by the association of gender with primary psychopathy (see Table 4 and Figure 4).

An additional finding in Figure 4 is that longer total response time is positively associated with the CNI parameters. Accordingly, a consequentialist response pattern, sensitivity to norms, and an inaction bias have a higher probability for participants who spend more time on the task. In contrast, primary psychopathy is found to be negatively associated with both total response time and the I parameter. This indicates that participants, who score higher on primary psychopathy, have a higher probability of spending less time on the task and having an action bias. In contrast, the negative associations between primary psychopathy and the C and N parameters indicate that participants, who score higher on primary psychopathy are less sensitive to the effect of norms (proscriptive vs. prescriptive) and whether the outcomes benefit the greater good (greater benefit vs. smaller benefit), in line with previous results (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021).

That the total response time was positively associated with all of the C, N, and I parameters indicates that in the context of the CNI scenarios, neither a norm-based response pattern nor a bias towards inaction is the result of a rapid, automatic response. In contrast, previous work has assumed that utilitarian judgments were produced by controlled cognitive comparisons of costs and benefits while deontological responses in sacrificial dilemmas were based on automatic, emotional responses (Greene et al., 2001, 2004). Other authors have argued that deontological judgments are the result of participants' efforts to arrive at coherence by satisfying often conflicting constraints concerning rights and duties in their common-sense moral reasoning (Holyoak & Powell, 2016). While the former view would have predicted a negative association of the N parameter with total response time, the latter view is consistent with our finding of a positive association.

Figure 4 shows that the C1-C4 and Npre/Npro parameters of the CNIS$_8$ can be used as a measurement model for two latent C and N parameters. From this structural equation model, the violations of the invariance assumption can be read off from the differences in the path coefficients from the latent C and N parameters to the parameters of the CNIS$_8$ model that they are measured by. This in turn shows that the violation of the invariance assumption does not only take the form of an additive shift to the means but can also be found in noticeably different path coefficients and thereby in the correlations between the different measures. The possibility of fitting a structural equation model with C and N parameters as latent variables, which takes the violations of the invariance assumption into account, shows that it possible to specify a model that fits the CNI model's intended use while addressing the methodological skepticism raised by Baron and Goodwin (2020, 2021) and others.

9.4 Conclusion

Implicit in a recent controversy concerning the CNI model of moral judgment (Gawronski et al., 2017) lies a problematic invariance assumption that process-dissociation and multinomial processing models make in their applications in different areas of psychology (Klauer et al., 2015). By extending the CNI model with a skip option, we implemented a version of the CNI model, which avoided making the invariance assumption. This allowed us to test the invariance assumptions built into the CNI model, which were found to be violated both for the C and the N parameters.

In light of these results, we recommend that future use of the CNI model adds this further S parameter and follows the 8-parameter version that we presented (Appendix A). Through structural equation modelling, we further analyzed mediation effects on the role of psychopathy on the CNI parameters and extended previous findings which were primarily based on bi-variate correlations. It was found that the previously reported effect of gender on the CNI parameters is completely mediated by the association of gender with primary psychopathy.

References

Baron, J. (1996). Do no harm. In D. M. Messick & A. E. Tenbrunsel (Eds.), *Codes of conduct: Behavioral research into business ethics* (pp. 197–213). New York: Russell Sage Foundation.

Baron, J. & Goodwin, G. P. (2020). Consequences, norms, and inaction: A comment. *Judgment and Decision Making, 15*(3), 421–442.

Baron, J. & Goodwin, G. J (2021). Consequences, norms, and inaction: Response to Gawronski et al.. *Judgment and Decision Making, 16*(2), 566-595.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86.

Berentelg, M. (2020). Multinomial Modeling of Moral Dilemma Judgment: A Replication Study. Retrieved in November 2021 from https://osf.io/mb32t/.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235.

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift fur Psychologie / Journal of Psychology, 217*, 108–124.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hutter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113*, 343–376.

Gawronski, B., Conway, P., Hütter, M., Luke, D.M., Armstrong, J., & Friesdorf, R. (2020). On the validity of the CNI model of moral decision-making: Reply to Baron and Goodwin (2020). *Judgment and Decision Making, 15*(6), 1054-1072.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Heck, D.W., Arnold, N.R. & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50***,** 264–284.

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology, 118*, 22–56.

Holyoak, K. J. & Powell, D. (2016). *Deontological coherence: A framework for commonsense moral reasoning*. *Psychol Bull., 142*(11), 1179-1203

Janoff-Bulman, R., Sheikh, S. and Hepp, S. (2009). Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation. *Journal of Personality and Social Psychology, 96*(3), 521-537.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*, 131–164.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). “Utilitarian” judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Klauer, K. C. (2010). Hierarchical Multinominal Processing Tree Models: A Latent-Trait Approach. *Psychometrika, 75*(1), 70-98.

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General, 144*(1), 198–221.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Edition). New York: The Guildford Press.

Körner, A., Deutsch, R., and Gawronski, B. (2020). Using the CNI Model to Investigate Individual Differences in Moral Judgments. *Personality and Social Psychology Bulletin*, 1-16.

Kroneisen, M., & Heck, D. W. (2020). Interindividual Differences in the Sensitivity for Consequences, Moral Norms, and Preferences for Inaction: Relating Basic Personality Traits to the CNI Model. *Pers Soc Psychol Bull., 46*(7), 1013-1026

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Levenson, M. R., Kiehl, K. A., Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population, *Journal of personality and Social Psychology, 68,* 151-158.

Luke, D. M. & Gawronski, B. (2021). Psychopathy and Moral Dilemma Judgments: A CNI Model Analysis of Personal and Perceived Societal Standards. *Social Cognition, 39*(1), 41-58.

Luke, D. M., Neumann, C.S., Gawronski, B. (2021). Psychopathy and Moral-Dilemma Judgment: An Analysis Using the Four-Factor Model of Psychopathy and the CNI Model of Moral Decision-Making. *Clinical Psychological Science*, 1-17. doi:10.1177/21677026211043862

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2013). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

McIntyre, A. (2019). Doctrine of Double Effect. In: The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.). Retrieved in November 2021 from https://plato.stanford.edu/archives/spr2019/entries/double-effect/.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(4), 1–30.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2th Ed.). Cambridge: Cambridge University Press.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*, 167–183.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford University Press.

Appendix A: The CNIS Model

Model Equations of the CNIS Model

The model equations for the 8-parameter version of the CNIS model (CNIS$_8$) are:

$P(\text{action}|\text{ProGreater}) = C_1 + (1\text{-}C_1) \times (1\text{-}N_{pro}) \times (1\text{-}S) \times (1\text{-}I)$

$P(\text{inaction}|\text{ProGreater}) = (1\text{-}C_1) \times N + (1\text{-}C_1) \times (1\text{-}N_{pro}) \times (1\text{-}S) \times I$

$P(\text{skip}|\text{ProGreater}) = (1\text{-}C_1) \times (1\text{-}N_{pro}) \times S$

$P(\text{action}|\text{ProSmaller}) = (1\text{-}C_2) \times (1\text{-}N_{pro}) \times (1\text{-}S) \times (1\text{-}I)$

$P(\text{inaction}|\text{ProSmaller}) = C_2 + (1\text{-}C_2) \times N_{pro} + (1\text{-}C_2) \times (1\text{-}N_{pro}) \times (1\text{-}S) \times I$

$P(\text{skip}|\text{ProSmaller}) = (1\text{-}C_2) \times (1\text{-}N_{pro}) \times S$

$P(\text{action}|\text{PreGreater}) = C_3 + (1\text{-}C_3) \times N_{pre} + (1\text{-}C_3) \times (1\text{-}N_{pre}) \times (1\text{-}S) \times (1\text{-}I)$

$P(\text{inaction}|\text{PreGreater}) = (1\text{-}C_3) \times (1\text{-}N_{pre}) \times (1\text{-}S) \times I$

$P(\text{skip}|\text{PreGreater}) = (1\text{-}C_3) \times (1\text{-}N_{pre}) \times S$

$P(\text{action}|\text{PreSmaller}) = (1\text{-}C_4) \times N_{pre} + (1\text{-}C_4) \times (1\text{-}N_{pre}) \times (1\text{-}S) \times (1\text{-}I)$

$P(\text{inaction}|\text{PreSmaller}) = C_4 + (1\text{-}C_4) \times (1\text{-}N_{pre}) \times (1\text{-}S) \times I$

$P(\text{skip}|\text{PreSmaller}) = (1\text{-}C_4) \times (1\text{-}N_{pre}) \times S$

For the *i*th participant, a data vector, $\boldsymbol{y}_i$, consisting of counts of each of these three response categories (action, inaction, skip) across the four CNI conditions (ProGreater, ProSmaller, PreGreater, PreSmaller) is formed. Via the CNIS model equations, these counts are modelled through a vector of 8 theta parameters for each participant, $\boldsymbol{\theta}_i$. In the four-parameter version, the invariance assumption is made, whereby $N_{pre} = N_{pro} = N$ and $C_1 = C_2 = C_3 = C_4 = C$, resulting in a vector of 4 theta parameters for each participant, $\boldsymbol{\theta}_i$.

In the standard CNI model, the inaction bias corresponding to the I parameter governs responses when neither moral cue (norms or consequences) compels a response. Similarly, in the extended CNIS model, the skip option comes into play, if participants have no guidance as to their response from norms and consequences and thus, in the $(1\text{-}C_j) \times (1\text{-}N_k)$ cases. This dovetails with the instruction to be permitted to skip in case participants are undecided about whether the described action is morally acceptable or inacceptable. In the original model, participants have the choice between action and inaction in this state of uncertainty (cases with $(1\text{-}C) \times (1\text{-}N)$) with preferences governed by parameter I. One consequence is that although the skip parameter S is constant, the actual frequency of the use of the skip option can differ between the four types of dilemmas to the extent that C and N differ between them.

In the extended CNIS model, we offer participants three choices instead of only two in the case of reaching the uncertainty state with probability (1-C_j) × (1-N_k): They can then skip, choose action, or choose inaction with probabilities S, (1-S) × (1-I) and (1-S) × I. Both the S and I parameters can also vary between persons, and in the model with random effects by scenario as a function of scenario (see Appendix B). Yet, both the S and I parameters remain invariant across the four CNI conditions within every scenario.

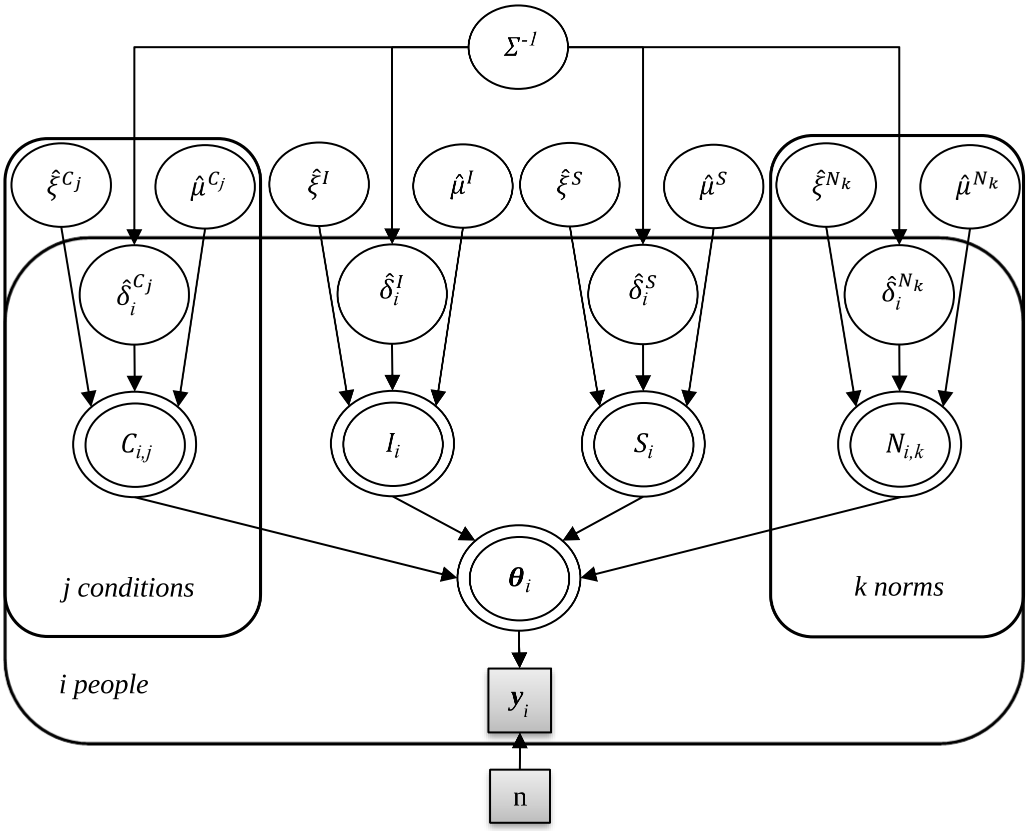
Bayesian Hierarchical Implementation

To estimate the MPT parameters of the CNIS model for each participant separately, we here follow the hierarchical latent trait model of Klauer (2010), which has also been implemented in the `TreeBUGS` R-package by Heck et al. (2018).

In this approach, a probit link function is used to transform MPT parameters (representing probabilities between 0 and 1) to the real line, $\Phi^{-1}(\theta)$. The transformed parameters are then modelled via a multivariate normal distribution while estimating mean, μ, and covariance matrix, Σ, from the data. The advantage of this approach is that heterogeneity in parameter estimates across participants and correlations among MPT parameters can be accommodated while allowing for partial aggregation of statistical information across participants in the posterior parameters of the multivariate normal distribution (Klauer, 2010). Accordingly, for each participant, i, the probit-transformed parameters are additively decomposed into a group mean, μ, and a random effect, $\Phi^{-1}(\theta) = \mu + \delta_i$.

We contrasted two hierarchical multinominal models following this approach with different numbers of MPT parameters (4 vs. 8). Table 1A illustrates CNIS_8, whereby a distinct C parameter is estimated for each of the j = 1, …, 4 CNI conditions, and a distinct N parameter is estimated for each of the k = 1, 2 types of norms. For CNIS_4, one shared C parameter is estimated (j = 1) together with one shared N (k = 1) parameter.

Table 1A. Hierarchical Latent Trait MPT Model



$$\hat{\mu}^{C_j}, \hat{\mu}^{N_k}, \hat{\mu}^{I}, \hat{\mu}^{S} \sim \text{Gaussian}(0,1)$$

$$\hat{\xi}^{C_j}, \hat{\xi}^{N_k}, \hat{\xi}^{I}, \hat{\xi}^{S} \sim \text{Uniform}(0,10)$$

$$\Sigma^{-1} \sim \text{Wishart}(\mathbf{I}, df)$$

$$\left(\hat{\delta}_i^{C_j}, \hat{\delta}_i^{N_k}, \hat{\delta}_i^{I}, \hat{\delta}_i^{S}\right) \sim \text{MvGaussian}(\mathbf{0}, \Sigma^{-1})$$

$$C_{i,j} \leftarrow \Phi\left(\hat{\mu}^{C_j} + \hat{\xi}^{C_j}\hat{\delta}_i^{C_j}\right)$$

$$N_{i,k} \leftarrow \Phi\left(\hat{\mu}^{N_k} + \hat{\xi}^{N_k}\hat{\delta}_i^{N_k}\right)$$

$$I_i \leftarrow \Phi\left(\hat{\mu}^{I} + \hat{\xi}^{I}\hat{\delta}_i^{I}\right)$$

$$S_i \leftarrow \Phi\left(\hat{\mu}^{S} + \hat{\xi}^{S}\hat{\delta}_i^{S}\right)$$

$$\boldsymbol{y}_i \sim \text{Multinomial}(\boldsymbol{\theta}_i, n)$$

Note. There are four CNI conditions with three categorical responses (action, inaction, skip). Via the CNIS model equations displayed above, the outcome probabilities of the responses in the data vector, $\boldsymbol{y}_i$, are represented by 8 theta parameters. For each participant, a vector of 8 theta parameters, $\boldsymbol{\theta}_i$, is estimated. The inverse Wishart distribution has 8+1 degrees of freedom, *df*, and a 8×8 identity matrix, **I**.

The models were fitted in a Bayesian framework through a Gibbs sampler, which estimates the posterior distributions of model parameters by means of Monte Carlo-Markov chains.

Appendix B: Item Effects

Baron and Goodwin (2020) suggest that both item and participants effects should be estimated for the CNI parameters. Above, we have already estimated the CNI parameters for each participant to test individual variation. In an exploratory analysis, we also estimated item effects in a model with crossed random effects for participants and scenarios (Matzke et al., 2013) to test whether the invariance assumption would be violated within the individual scenarios used in the experiment. For an analysis with less uncertainty in the estimates, a larger sample size would be required. However, the exploratory analysis displayed in Table 1B below already suggests violations of the invariance assumption almost in every scenario investigated.

Table 1B. Item Effects and the Invariance Assumption

Scenario	$C_{1,\text{ ProGreater}}$	$C_{2,\text{ ProSmaller}}$	$C_{3,\text{ PreGreater}}$	$C_{4,\text{ PreSmaller}}$	I	N_{pre}	N_{pro}	S
Assisted-suicide	$\tilde{x}$ = .34, 95% [.23, .44]	$\tilde{x}$ = .01, 95% [$3.0 \cdot 10^{-9}$, 0.05]	$\tilde{x}$ = .05, 95% [.004, 0.12]	$\tilde{x}$ = .05, 95% [.001, .11]	$\tilde{x}$ = .58, 95% [.52, .64]	$\tilde{x}$ = .08, 95% [.03, .15]	$\tilde{x}$ =.18, 95% [.05, .33]	$\tilde{x}$ = .12, 95% [.09, .15]
Bishop	$\tilde{x}$ = .0003, 95% [$6.7 \cdot 10^{-11}$, .002]	$\tilde{x}$ = .49, 95% [.26, .70]	$\tilde{x}$ = .25, 95% [.07, .43]	$\tilde{x}$ = .0002, 95% [$5.0 \cdot 10^{-16}$, .002]	$\tilde{x}$ = .62, 95% [.54, .69]	$\tilde{x}$ = .31, 95% [.19, .42]	$\tilde{x}$ = .16, 95% [$3.8 \cdot 10^{-16}$, 0.63]	$\tilde{x}$ = .15, 95% [.04, .36]
Construc-tionsite	$\tilde{x}$ = .009, 95% [.0006, .02]	$\tilde{x}$ = .14, 95% [.004, .33]	$\tilde{x}$ = .02, 95% [$1.4 \cdot 10^{-7}$, .09]	$\tilde{x}$ = .006, 95% [$7.7 \cdot 10^{-5}$, .02]	$\tilde{x}$ = .59, 95% [.52, .66]	$\tilde{x}$ = .43, 95% [.31, .54]	$\tilde{x}$ = .37, 95% [.20, .53]	$\tilde{x}$ = .04, 95% [.02, .06]
Dialysis	$\tilde{x}$ = .12, 95% [.06, .18]	$\tilde{x}$ = .31, 95% [.15, .48]	$\tilde{x}$ = .43, 95% [.29, .57]	$\tilde{x}$ = .03, 95% [$1.2 \cdot 10^{-5}$, .08]	$\tilde{x}$ = .59, 95% [.53, .66]	$\tilde{x}$ = .03, 95% [.005, .07]	$\tilde{x}$ = $5.77 \cdot 10^{-5}$, 95% [$5.5 \cdot 10^{-18}$, .003]	$\tilde{x}$ = .08, 95% [.06, .11]
Immune-deficiency	$\tilde{x}$ = .0002, 95% [$1.5 \cdot 10^{-14}$, .003]	$\tilde{x}$ = .30, 95% [.09, .52]	$\tilde{x}$ = .17, 95% [.02, .35]	$\tilde{x}$ = $4.3 \cdot 10^{-5}$, 95% [$9.3 \cdot 10^{-13}$, .0007]	$\tilde{x}$ = .66, 95% [.58, .73]	$\tilde{x}$ = .62, 95% [.51, .72]	$\tilde{x}$ = .26, 95% [.08, .45]	$\tilde{x}$ = .06, 95% [.04, .08]
Mother	$\tilde{x}$ = .10, 95% [.05, .16]	$\tilde{x}$ = .45, 95% [.21, .69]	$\tilde{x}$ = .29, 95% [.12, .46]	$\tilde{x}$ = .01, 95% [$4.6 \cdot 10^{-5}$, .03]	$\tilde{x}$ = .61, 95% [.55, .68]	$\tilde{x}$ = .46, 95% [.34, .57]	$\tilde{x}$ = .34, 95% [.08, .41]	$\tilde{x}$ = .13, 95% [.09, .17]
Peanuts	$\tilde{x}$ = .08, 95% [.03, .13]	$\tilde{x}$ = .03, 95% [$9.7 \cdot 10^{-10}$, 0.16]	$\tilde{x}$ = .004, 95% [$1.1 \cdot 10^{-7}$, .03]	$\tilde{x}$ = .009, 95% [.0001, .03]	$\tilde{x}$ = .42, 95% [.13, .74]	$\tilde{x}$ = .30, 95% [.20, .39]	$\tilde{x}$ = .23, 95% [.08, .40]	$\tilde{x}$ = .05, 95% [.03, .06]
Torture	$\tilde{x}$ = .38, 95% [.27, .48]	$\tilde{x}$ = .56, 95% [.34, .76]	$\tilde{x}$ = .26, 95% [.11, .42]	$\tilde{x}$ = .09, 95% [.02, .17]	$\tilde{x}$ = .59, 95% [.52, .67]	$\tilde{x}$ = .21, 95% [.10, .34]	$\tilde{x}$ = .03, 95% [$2.2 \cdot 10^{-6}$, .14]	$\tilde{x}$ = .06, 95% [.04, .09]
Transplant	$\tilde{x}$ = .002, 95% [$1.5 \cdot 10^{-17}$, .01]	$\tilde{x}$ = .0006, 95% [$2.4 \cdot 10^{-21}$, 0.01]	$\tilde{x}$ = .0002, 95% [$1.3 \cdot 10^{-21}$, .006]	$\tilde{x}$ = $2.6 \cdot 10^{-5}$, 95% [$8.4 \cdot 10^{-20}$, .002]	$\tilde{x}$ = .55, 95% [.48, .61]	$\tilde{x}$ = .11, 95% [.05, .17]	$\tilde{x}$ = .14, 95% [.04, .26]	$\tilde{x}$ = .05, 95% [.03, .06]
Vaccine	$\tilde{x}$ = $5.5 \cdot 10^{-5}$ 95% [$1.3 \cdot 10^{-23}$, .002]	$\tilde{x}$ = .24, 95% [.05, .44]	$\tilde{x}$ = .009, 95% [$2.7 \cdot 10^{-7}$, .04]	$\tilde{x}$ = $1.9 \cdot 10^{-6}$, 95% [$2.1 \cdot 10^{-18}$, .0001]	$\tilde{x}$ = .60, 95% [.52 .66]	$\tilde{x}$ = .27, 95% [.18, .36]	$\tilde{x}$ = .15, 95% [.03, .30]	$\tilde{x}$ = .06, 95% [.04, .08]

Note. The square brackets indicate 95% highest density intervals (HDI).