Norm Conflicts and Conditionals

Niels Skovgaard-Olsen

University of Göttingen

David Kellen

Syracuse University

Ulrike Hahn

Birkbeck University

Karl Christoph Klauer

Albert-Ludwigs-Universität Freiburg

Author Note

Niels Skovgaard-Olsen, Department of Psychology, Cognition and Decision Making, University of Göttingen, Göttingen, Germany. David Kellen, Department of Psychology, Syracuse University, USA. Ulrike Hahn, Department of Psychology, Birkbeck University, UK. Karl Christoph Klauer, Department of Psychology, Albert-Ludwigs-Universität Freiburg, Freiburg Germany.

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).

The supplemental materials including all data and analysis scripts are available at: https://osf.io/9fm45/.

Abstract

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.

Keywords: problem of arbitration, conditionals, and-to-if inferences, relevance, reflective attitudes, Bayesian mixture modeling

Norm Conflicts and Conditionals[1]

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

[1] 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.

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 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.[2] 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' 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.[3] 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

[2] 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.

[3] 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).

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 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.[4] 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).

---

[4] 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).

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 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:

$$P(\text{if A, then C}) = P(C|A),$$

which is referred to as the *conditional-probability hypothesis*.[5] 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).

---

[5] 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).

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 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(Conclusion) ≥ P(Premise)

This prediction was directly tested by Cruz *et al.* (2015), who found that participants conformed to this inequality at above-chance levels.[6]

[6] 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).

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 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 $\Delta P > 0$

A is *negatively relevant* for C (and a reason *against* C) iff $\Delta P < 0$

A is *irrelevant* for C iff $\Delta P = 0$

For $\Delta P = P(C|A) - P(C|\neg A)$

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, $\Delta P > 0$ 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):

$$W_C = \frac{\Delta P}{1 - P(E|\neg C)} \quad \textit{for } E = \textit{effect}, C = \textit{cause}$$

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.[7]

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] For a detailed discussion of how the Bayesian mixture model differs from previous regression-based approaches, see Online Supplementary Materials.

Experiments 1 and 2

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).

Method

Participants

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 Supplementary 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.

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.

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

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%.

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.

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).

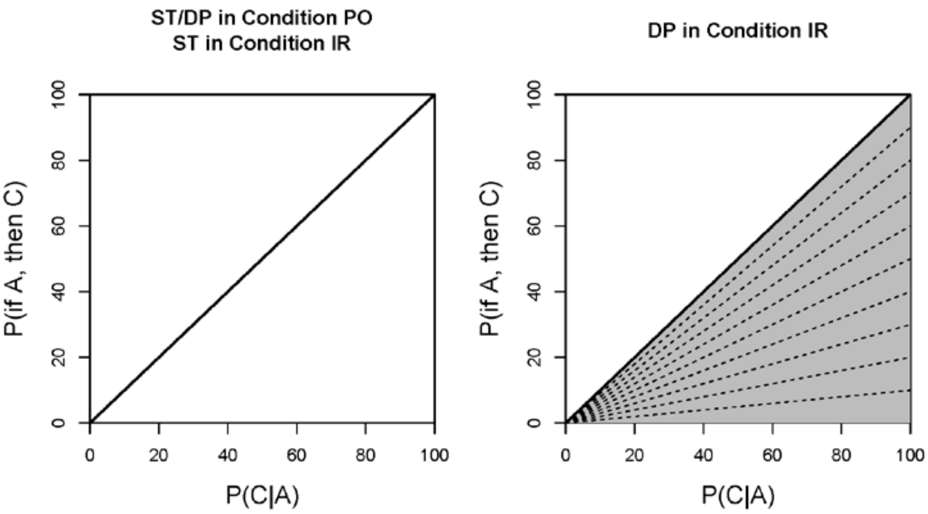


Figure 1. Predictions. The Suppositional Theory (ST) equates P(if A then C) and P(C|A). The Default-Penalty Hypothesis (DP) makes the same prediction only for positive relevance (PO). For irrelevance (IR), it expects a function that lies below the diagonal. Here we assume for our classificatory purposes 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.

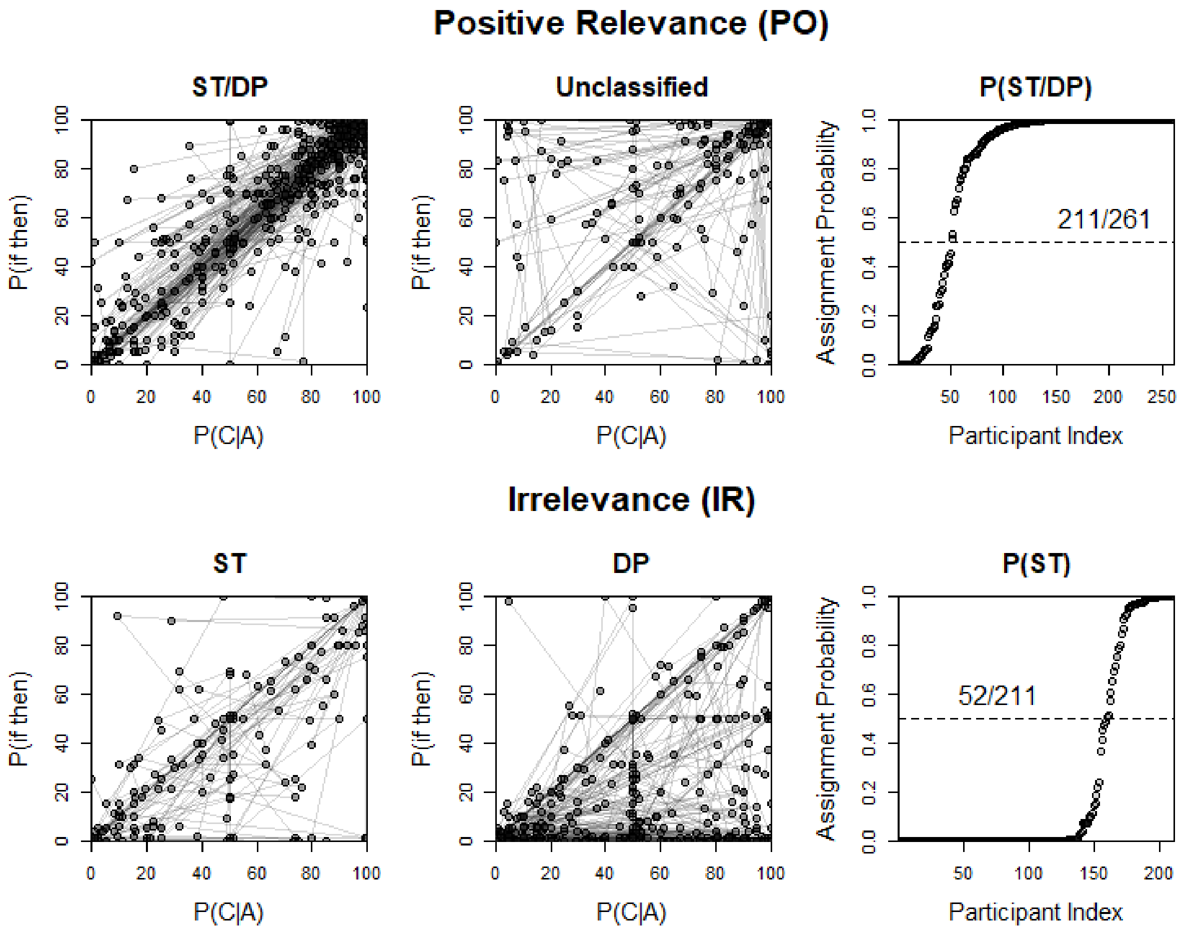


Figure 2. Left and Center Panels: Individual data associated to the phase 1 classifications in Experiment 1. Right Panels: Individuals' posterior classifications (note that in the irrelevance condition, only participants classified as ST/DP in the positive relevance condition were considered).

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:

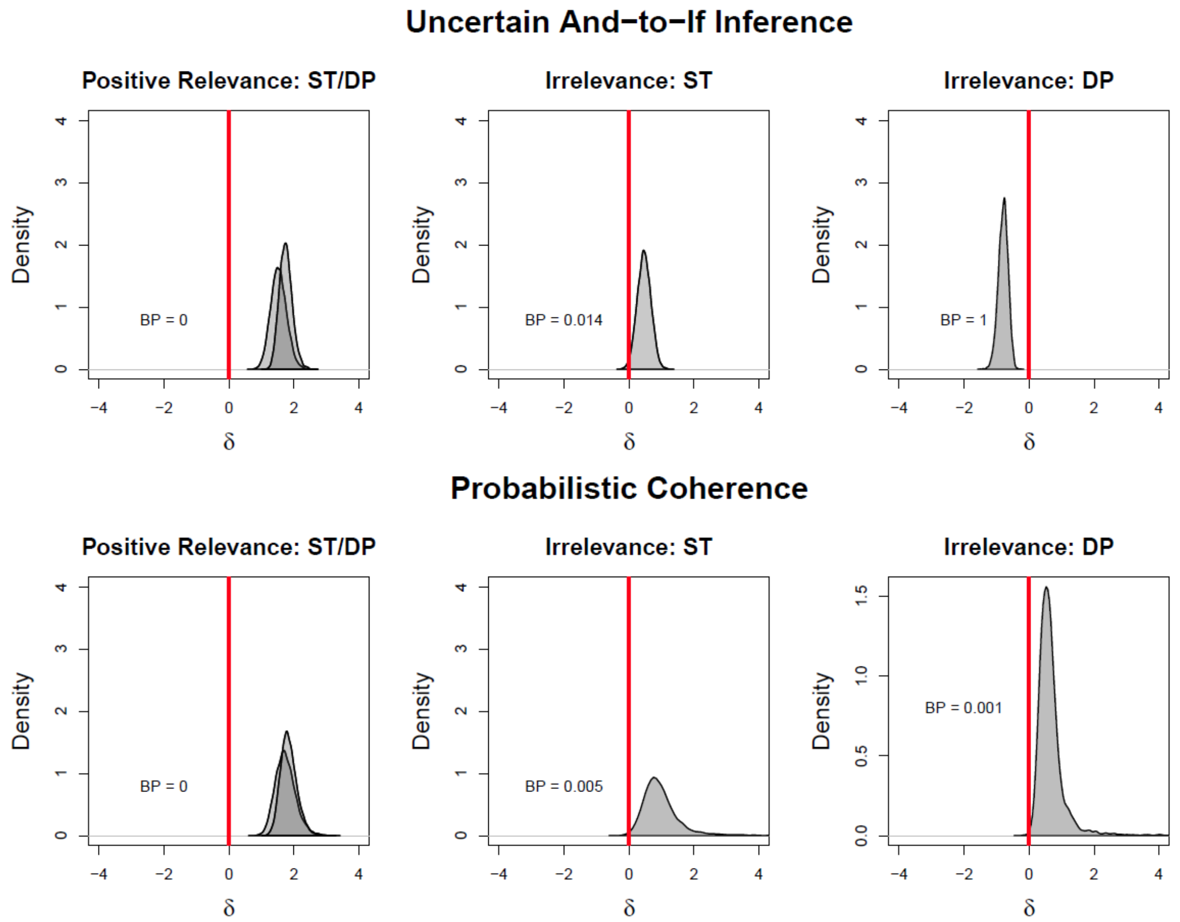


Figure 3. Posterior distributions of the deviations of the tested inequalities from chance-level occurrence (represented on an effect-size scale) in Experiment 1. The vertical lines indicate effect size 0 and BP corresponds to the probability of samples from the posterior distributions taking on values 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.

$$\text{P(Conclusion)} \geq \text{P(Premise)}$$

$$P(C|A) \geq P(A,C)$$

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.

Experiment 3

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):

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

Hence,

$$\text{A and C} \vDash \text{if A, then C} \quad \text{only if} \quad \text{P(if A, then C)} \geq \text{P(A and C)}$$

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[8]

$$\text{A} \vDash \text{A and C.}$$

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

[8] 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.

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 favorite color, 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 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.

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

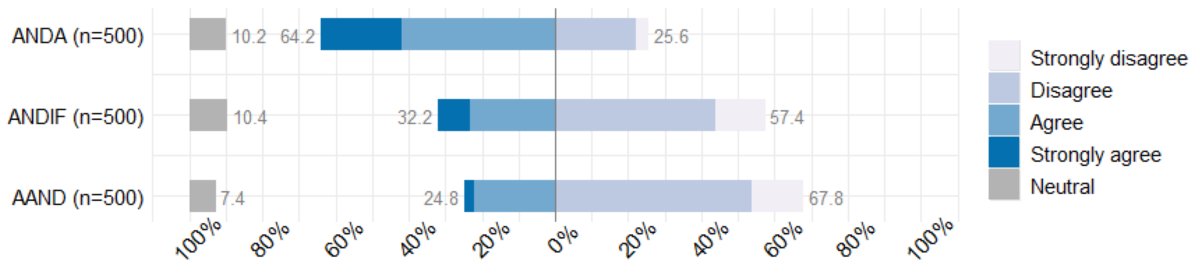
	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1	30307.93	10.04	2.15	30276.2	0.006
M2	30302.11	4.22	0.89	30270.3	0.108
M3	30297.89	0	--	30266.6	0.886
M4	4968.35	4.52	5.22	4964.8	0.095
M5	4963.84	0	--	4960.5	0.905
M6	5118.24	154.41	28.30	5113.7	0.000

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_{AAND_accept} = -0.57, 95%-CI [-0.658, -0.485], $BF_{H0H1} = -2.75 * 10^{-26} \approx 0$). 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_{ANDA_accept} = 0.09, 95%-CI [-0.001, 0.184], BF_{H0H1} = 17.17), which makes M5 the second most preferred model.

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.

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) ≥ P(A,C) at above-chance levels, but only one of the groups conforms to P(if A, then C) ≥ P(A,C) across relevance conditions. This behavior 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.

Implications for Rationality Research

Schurz & 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 fueled 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 neuro-biology 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).

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, Inc.

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.

Online Supplementary Materials:[9] Bayesian Mixture Model

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:

$$\mathrm{P(if\ A\ then\ C)} = \beta_0 + \beta_1 \mathrm{P(C|A)} + \varepsilon,$$

where β_0 and β_1 are the intercept and slope parameters respectively, and ε is the residual term.[10] 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 β_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,

[9] Further supplemental materials including all data and analysis scripts are available at: https://osf.io/9fm45/.

[10] 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.

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. 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:

P(Conclusion) ≥ P(Premise)

P(C|A) ≥ 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. $\Phi(K_{i,j}) = 1 - \text{P(Premise)}$; 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.

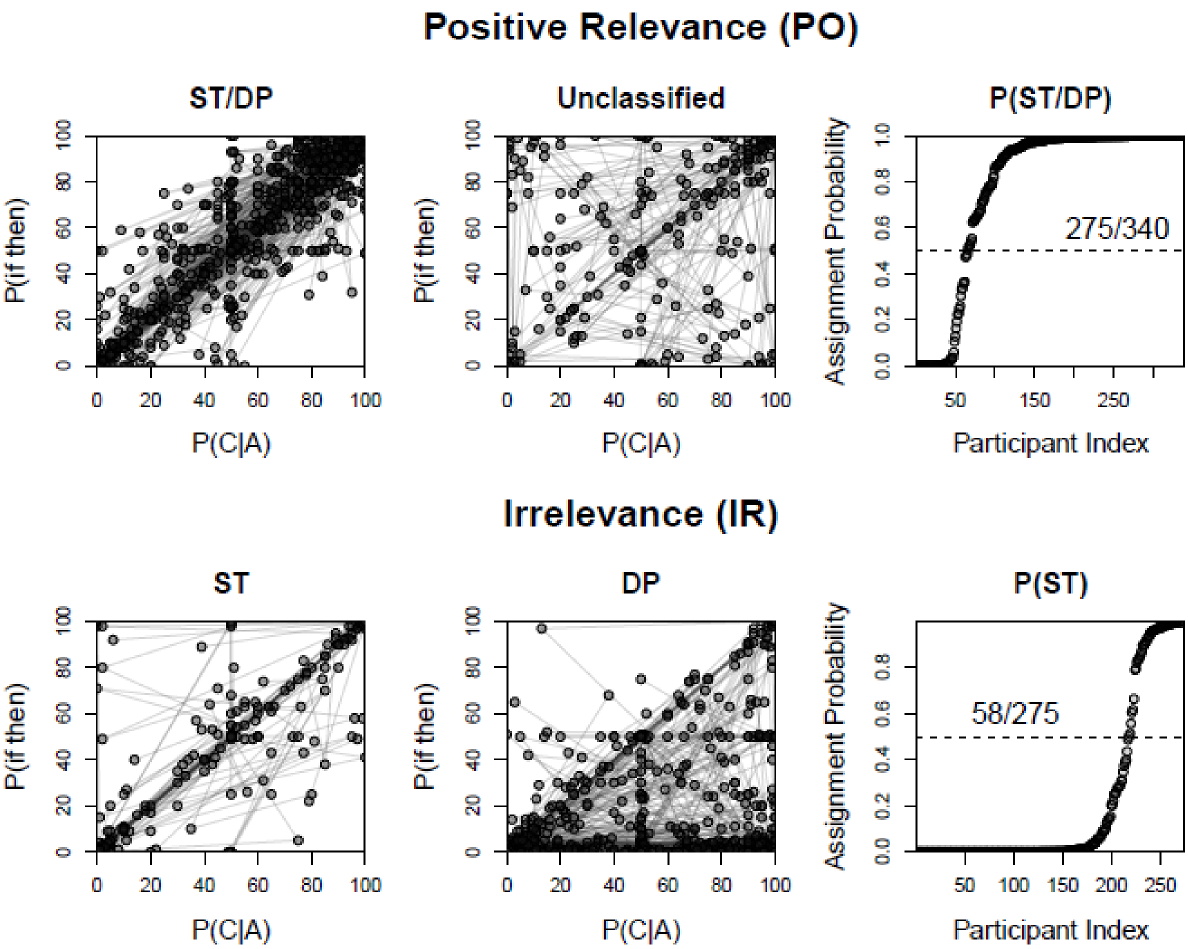


Figure A1. Left and Center Panels: Individual data associated to the phase 1 classifications in Experiment 2. Right Panels: Individuals' posterior classifications (note that in the irrelevance condition, only participants classified as ST/DP in the positive relevance condition were considered).

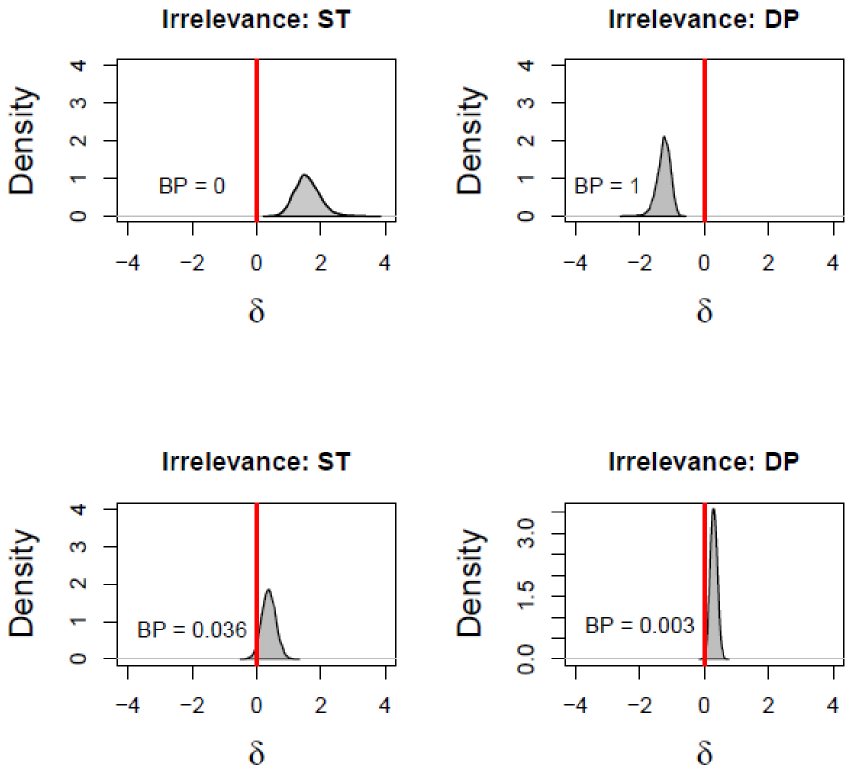


Figure A2. Posterior distributions of the deviations of the tested inequalities from chance-level occurrence (represented on an effect-size scale) in Experiment 2. The vertical lines indicate effect size 0 and BP corresponds to the probability of samples from the posterior distributions taking on values 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.