# Arbitrating Norms for Reasoning Tasks[1]

Aliya R. Dewey

Department of Philosophy, University of Arizona

The psychology of reasoning uses norms to categorize responses to reasoning tasks as correct or incorrect in order to interpret the responses and compare them across reasoning tasks. This raises the *arbitration problem*: any number of norms can be used to evaluate the responses to any reasoning task and there doesn't seem to be a principled way to arbitrate among them. Elqayam and Evans have argued that this problem is insoluble, so they call for the psychology of reasoning to dispense with norms entirely. Alternatively, Stupple and Ball have argued that norms must be used, but the arbitration problem should be solved by favouring norms that are sensitive to the context, constraints, and goals of human reasoning. In this paper, I argue that the design of reasoning tasks requires the selection of norms that are indifferent to the factors that influence human responses to the tasks—which aren't knowable during the task design phase, before the task has been given to human subjects. Moreover, I argue that the arbitration problem is easily dissolved: any well-designed task will contain instructions that implicitly or explicitly specify a single determinate norm, which specifies what would count as a solution to the task—independently of the context, constraints, and goals of human reasoning. Finally, I argue that discouraging the use of these *a priori* task norms may impair the design of novel reasoning tasks.

**Keywords:** reasoning; rationality; rational analysis; norms; error; normativism; task design

In the psychology of reasoning, norms are used to categorize responses to tasks under two general normative categories (correct and incorrect) so that responses can be compared across different tasks (and/or different studies). But norms are cheap: there are arbitrarily many norms of logic, mathematics, and statistics, so there are any number of ways to evaluate the responses to any reasoning task. This creates the *arbitration problem*: there doesn't seem to be a principled way to arbitrate which norm should be used to categorize responses to any reasoning task (Cohen, 1981; Gigerenzer, 1991; Evans, 1993; Stanovich & West, 2000; Elqayam & Evans, 2011). This is a problem because success and failure call for different kinds of explanation (Anderson, 1990). So, if we can't arbitrate among norms, we won't be able to arbitrate among different explanations of reasoning on any given task.

Elqayam and Evans have argued that the arbitration problem is insoluble, so the psychology of reasoning should dispense with norms altogether (Evans, 2007; Elqayam & Evans, 2011; Elqayam, 2011, 2012; Elqayam & Over, 2016). They call for *descriptivism*, which rejects categorizing any response as incorrect. Their descriptivist framework uses a grounded conception of rationality, but this grounded rationality isn't genuinely normative: it rationalizes any actual response, such that

[1] Penultimate draft of article forthcoming in *Synthese*. Please cite the published article, not this draft.

there is no possibility for error. Elqayam and Evans take this to be a virtue of their account: they argue that it closes Hume's (1739) is-ought gap, and so ensures empiricism. However, many critics believe that descriptivism is too extreme: it's inconsistent with the fact that norms are productively used across the psychology of reasoning and other behavioural sciences (Stupple & Ball, 2014; Skovgaard-Olsen et al., 2019; see commentaries on Elqayam & Evans, 2011).

Many of these critics agree with Elqayam and Evans that we should reject the use of "hard norms", such as classical logic, elementary algebra, and probability theory, which are widely used in the heuristics-and-biases (H&B) program (Wason & Johnson-Laird, 1972; Tversky & Kahneman, 1974; Kahneman, 2011). *Hard norms* define correctness such that errors are common in actual human reasoning. Since there are an unlimited number of ways to define correctness such that correct and actual responses aren't very positively correlated, the arbitration problem is difficult for hard norms. But critics insist that the psychology of reasoning should continue to use "soft norms", such as the norms of instrumental rationality (Evans & Over, 1996), Bayesian rationality (Oaksford & Chater, 2007), and ecological rationality (Gigerenzer & Brighton, 2009), which are used outside the H&B program. Since there are fewer ways to define correctness such that correct and actual responses are very positively correlated, the arbitration problem is easier for softer norms. Instead of dispensing with all norms, then, *soft normativism* claims that we should dispense with hard norms and retain soft norms (Stupple & Ball, 2014; Skovgaard-Olsen et al., 2019).

Soft normativism faces a puzzling question, though: if we should prefer to use norms that define correctness such that correct and actual responses are *very positively* correlated, why not define correctness such that correct and actual responses are *perfectly* correlated? That is, if we should prefer soft norms to hard norms, why not go all the way to descriptivism and use *empty norms* (e.g., grounded rationality)? The answer has to be that weakening the correlation between correct and actual responses confers some advantage. If that's true, then wouldn't hard norms offer more of that advantage than soft norms? Wouldn't that favour hard norms, at least for some purposes? Or would that advantage be offset by the arbitration problem? I suspect that soft normativism lacks the resources to answer these questions. Instead, I propose that we consider *pluralist normativism*, which claims that soft and hard norms have different advantages, and the psychology of reasoning can acquire both advantages by carefully coordinating the use of both soft and hard norms.

In this paper, my goal is to argue that a successful psychology of reasoning requires the use of hard norms, to solve the arbitration problem for hard norms, and to show how to coordinate the use of hard and soft norms. In §1, I argue that designing reasoning tasks requires us to use hard (vs. soft) norms since they don't require us to speculate about the results of the experiment being designed. In §2, I develop a partial solution to the arbitration problem: I show how to arbitrate between two task norms for an easy instance of the arbitration problem. In §3, I develop a complete solution to the arbitration problem: I show how to arbitrate between two task norms for a difficult instance of the arbitration problem. I also defend my solution against an important objection. In §4, I explain how task norms guide the discovery of *cognitive norms*, which are the softer norms that human cognition tends to conform to. In §5, I argue that while task norms guide the discovery of cognitive norms, the initial choice of task norms is a matter of luck and hence, is immune to criticism. In §6, I conclude by emphasizing that novel research programs are especially dependent on hard norms and should not be dissuaded from doing so by descriptivism or soft normativism.

## §1. Task Norms

The psychology of reasoning requires norms to categorize and compare responses in two contexts: task design and data interpretation. The two contexts represent very different epistemic situations: the context, constraints, and goals that determine human reasoning in response to a task might be inferred during the data interpretation stage but are unknowable during the task design stage. Soft norms that are sensitive to the context, constraints, and goals that determine human reasoning may be required during the data interpretation stage, but it wouldn't be feasible to require them during the task design stage. In this section, I'll argue that task design instead requires hard norms, which aren't sensitive to the context, constraints, or goals that determine human reasoning.

To illustrate this point, let's suppose that we had to design a task for studying algebraic reasoning back in the 1950s, before the H&B program. Consider our epistemic position. On the one hand, we'd have known the formal rules of elementary algebra and could have solved any elementary algebra problem, at least in principle. So, we'd have been in an epistemic position to competently define the hard, *a priori*, and absolute norms of algebra. On the other hand, we'd have known little about the context, constraints, and goals that causally influence human reasoning about algebra problems. So, we wouldn't have been in an epistemic position to competently define the soft, *a posteriori*, and contextual norms of human algebraic reasoning.

Next, let's suppose that we designed a task with multiple components. Imagine that we reinvented Frederick's (2005) Cognitive Reflection Test (CRT) to study algebraic reasoning:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ____ cents
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ minutes
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ___ days

At this point, we wouldn't have known how the contexts, constraints, and goals of human subjects would causally influence their responses to these questions. Of course, we'd have our *suspicions*, though: we'd have noticed that our own reasoning about these questions felt counterintuitive to us, and we might suspect that they would elicit intuitive yet incorrect responses in our subjects.

Since the CRT has three task components, we need a way to compare responses across the three task components. A norm lets us do this: we can categorize the possible responses to each task component as either correct or incorrect. Then we can aggregate the correct and incorrect responses across the three questions. Like Frederick (2005), we could define a CRT score: we could assign "0" to incorrect solutions and "1" to correct solutions and then sum the scores. So, a subject who incorrectly answered all three questions would earn a score of 0 whereas a subject who correctly answered all of them would earn a score of 3. This score represents the frequency of normative response types across three questions (for each subject).

Which norm do we use to categorize the responses as correct or incorrect? The obvious answer is that we should use the hard, *a priori* norms of elementary algebra, as Frederick (2005) did. When we use these norms, we're evaluating the judgments in relation to the reasoning tasks themselves: whether they select answers that solve the tasks. This treats the selection process as a black box: it's indifferent to the factors that vary within the processes that could select responses to the CRT, from human reasoning to MATLAB. However, soft normativists might object that it would be appropriate to hold human reasoners to algebraic norms only if they have endorsed them. This objection suggests that it's best to evaluate judgments only in relation to the exercise of cognitive capacities: whether they select answers that achieve the function of the cognitive capacities that the subject deploys.[2] If the subject deploys a cognitive capacity for elementary algebra, then and only then should their judgments be evaluated by the norms of elementary algebra.

Once again, the problem is that we aren't in an epistemic position to evaluate judgments in relation to the exercise of cognitive capacities. I suppose that we could ask subjects to confirm that they will respond in accordance with the norms of elementary algebra, but that would prime them to use the norms of elementary algebra. To avoid priming, we'd have to infer whether they are following the norms of elementary algebra or some other set of norms. To this day, though, no one has identified soft norms that would categorize judgments that fail to solve the CRT questions as correct (rational) and judgments that successfully solve the CRT question as incorrect (irrational). Even now, then, we're only in an epistemic position to use the *a priori* norms of elementary algebra to evaluate judgments in relation to the algebraic tasks themselves. See §3 for a further discussion of task designs that recognize the norms that subjects endorse.

Why not compare responses across tasks by categorizing them under non-normative types? We'd need to find a set of types that all responses fall under. An example is parity: all numbers are even- or odd-valued (or zero). So, we could compare even- and odd-valued responses across the three questions. But this is absurd: the parity of an answer is *irrelevant* to the task. For example, there is no relevant relation between the odd parity of the response that "the ball is $0.05" and the bat-and-ball task. By comparison, the normative status of an answer is *relevant* to the task. In fact, the normative status of a response is grounded in its relation with the task: e.g., "the ball is $0.05" is correct in virtue of the fact that it is a solution to the bat-and-ball task that satisfies all the rules of elementary algebra. In general, the only relevant relation between a response and a task will be normative: whether the response counts as a solution to the task.

Suppose that we finally do run the CRT and observe how humans will respond to it. We'll find that they tend to give one of only two responses to each question: (a) 10 cents or 5 cents, (b) 100 minutes or 5 minutes, and (c) 24 days or 47 days. Moreover, the most common responses to each question are (a) 10 cents, (b) 100 minutes, and (c) 24 days. These results are unintelligible and incommensurable until we use algebraic norms to categorize and compare these responses. Then we'd find the unexpected but intelligible result that the most common responses to each question are categorized as *incorrect*—they fail to solve the CRT questions. Using these algebraic norms,

[2] Samuels et al. (2012) make a similar distinction between two things that can be evaluated: the exercises of cognitive capacities and the judgments that result from the exercises of the cognitive capacities. I've proposed here that we can evaluate judgments in relation to these things.

we'd also find a variety of other intelligible results: e.g., cognitive load increases the frequency of incorrect responses on the CRT (Johnson et al., 2016).

An unusual feature of the CRT is that it has multiple task components (3 questions). One might wonder whether we need *a priori* norms to design tasks with single task components. For example, would we need algebraic norms in a study where we only had to design the bat-and-ball task? This is a fair question: we wouldn't need to compare responses across task components, so we wouldn't need a norm to categorize them as correct or incorrect. But there are at least two reasons why we'd still need algebraic norms. First, they would be necessary to make our results intelligible: finding that "$0.10" and "$0.05" are the two most common responses to the bat-and-ball task is difficult to interpret until we categorize the former as incorrect and the latter as correct.

Second, we'd still need to make comparisons between studies, even if we don't make comparisons within studies. For example, suppose that we already had designed and implemented the bat-and-ball task to study additive reasoning and we were designing another task to study multiplicative reasoning. We might want to design this new task such that we could make direct comparisons of its results with our bat-and-ball task. For example, we might want to test whether cognitive load has a similar effect on responses to the new task vs. the bat-and-ball task. To do that, we'd have to design the new task such that it elicited possible responses that can be evaluated as correct and incorrect using algebraic norms. To do that, we could design the lily pads task. Algebraic norms can be used to evaluate both responses, so comparisons would be easily possible.

Another unusual feature of the CRT is that it is what Elqayam & Evans (2011) call a *single norm paradigm*: the norms of elementary algebra are the only norms that have been proposed for the CRT. Till date, no one has developed an alternative norm—neither a hard norm that represents some alternate form of algebra nor a softer norm that's more sensitive to the context, constraints, and goals of human reasoning. One might object that this makes the CRT a misleading example. But I disagree: the CRT is a clear example of the fact that when we know relatively little about some form of reasoning (e.g., algebraic reasoning), we're only in an epistemic position to identify hard, *a priori* norms. When we know more about some form of reasoning (e.g., conditional reasoning), we're in an epistemic position to identify soft, *a posteriori* norms and then we are liable to forget the indispensable role of hard, *a priori* norms.

Henceforth, I'll use the term 'task norm' to refer to hard, *a priori* norms that evaluate judgments for whether they select solutions to the task (regardless of the nature of the process that does the selecting). I'll often refer to the achievement of task norms as *cognitive success*—to emphasize that we can use task norms to evaluate the products of cognition, even though they aren't sensitive to the nature of cognition.[3] Likewise, I'll use the term 'cognitive norm' to refer to soft, *a posteriori* norms that evaluate judgments for whether they are the result of the rational exercise of cognitive capacities in human reasoning. I'll often refer to the achievement of cognitive norms as *cognitive*

[3] My conception of cognitive success is completely different from Schurz & Hertwig's (2019). They argue that what is rational for the exercise of a cognitive capacity is whatever maximizes the likelihood of success. This is just a consequentialist conception of cognitive norms. My proposal is that task norms are distinct from cognitive norms and that achieving task norms counts as a success for cognition that is distinct from its counting as rational.

*rationality*, which is categorically different from cognitive success. We'll discuss the relations between task norms and cognitive norms further in §3.

## §2. Norm Arbitration

Once we distinguish task norms from cognitive norms, it's clear that the arbitration problem could be a problem of arbitrating between (a) task norms, (b) cognitive norms, or (c) task and cognitive norms. I suspect that descriptivists and soft normativists are worried about arbitrating between task norms. After all, they have responded to the arbitration problem by rejecting the hard, *a priori* norms that serve as task norms—not the soft, *a posteriori* norms that serve as cognitive norms. In this section, I'll consider an easy case of the arbitration problem and I'll show that it is possible (and relatively easy) to arbitrate between these norms. This will put us in a better position in the next section to consider a more difficult case of the arbitration problem.

Wason's (1968) selection task is an excellent example of a multiple norm paradigm (Elqayam & Evans, 2011). Here's one version of the task:

> **Wason Selection Task (WST):** in each round, participants are presented with four cards that each have a number on one side and a letter on the other side, but they can only see one side of each card: e.g., A, K, 2, and 7. Next, participants are asked to indicate which of these four cards they must turn over in order to determine whether various conditional sentences are true: e.g., "If there is an A on one side, then there is a 2 on the other side".

Unlike the CRT, there is only one task here, so we don't need to compare responses *across tasks* within a single study. But we still need to categorize responses using a norm to (a) interpret the responses and (b) compare them with responses to other tasks in other studies.

Traditionally, the norms of classical propositional logic have been used to categorize responses to this task.[4] According to these norms, the correct response is to turn over the A and 7 cards, because doing so enacts the true claim that turning over the A and 7 cards is the only way to determine whether the conditional sentence is true. Per *modus ponens*, the conditional sentence would be false if the A card did not have a 2 on its other side. Per *modus tollens*, it would also be false if the 7 card did have an A on its other side. After all, the conditional sentence is equivalent to the contrapositive sentence in classical propositional logic: "If there is a number different from 2 on one side, then there is a letter different from A on the other side". And according to these norms, every other response is incorrect: it can be plausibly interpreted as a deviation from the correct response. For example, turning only the A card is an incorrect response since it can be plausibly interpreted as failing to use *modus tollens*, despite that it succeeds at using *modus ponens*.

As Elqayam and Evans point out, though, the WST is a multiple norm paradigm—other norms can be used to categorize responses to it. One popular alternative is Oaksford & Chater's (1994, 2007) proposal that we can categorize responses as correct if they maximize the expected amount of information gained by turning the card and incorrect otherwise. I'll refer to this as the *Bayesian*

[4] Davies et al. (1995) argue that the WST is specifiable in predicate logic, not propositional logic, so the norms of classical predicate logic should be used to categorize responses instead of classical propositional logic. Johnson-Laird & Wason (1970) recognize the same problem but suspect that it doesn't make a significant empirical difference.

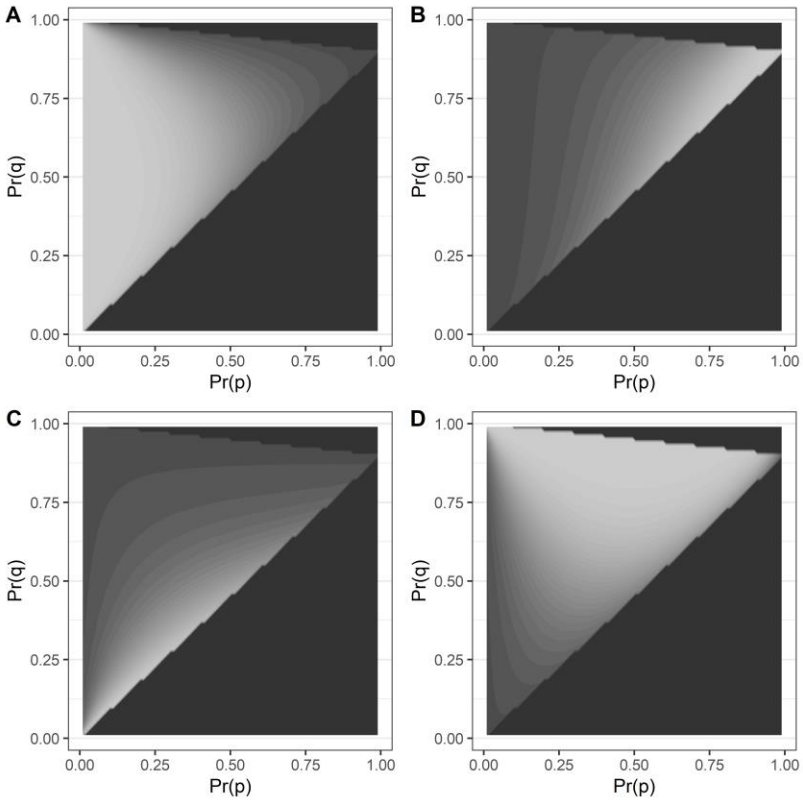


*Figure 1*. The probabilities with which a card should be selected for A: the *p*-card, B: the ¬*p*-card, C: the *q*-card, and D: the ¬*q*-card as a function of the probabilities of the antecedent [*P*(*p*), *x*-axes] and the consequent [*P*(*q*), *y*-axes] according to the scaled information gain model. The lighter the region, the greater the probability that a card should be selected. The prior probabilities [*P*(MI) and *P*(MD)] were set to .5, and *P*(*q*|*p*) to .9. The logistic selection tendency function parameters were as in Hattori (1999). Points in the lower right triangular region in black violate the requirement of the dependency model that $P(q) > P(p)P(q|p)$. (Please note that Oaksford and Chater provided this figure and its caption from the forthcoming latest edition of their 2007 book.)

*norm*. I won't consider the equations that they use to quantify the Bayesian norm because these equations are complicated and not specifically relevant to my response to the arbitration problem. It's unclear whether they take this norm to be a task norm (which defines solutions to the WST) or a cognitive norm (which defines cognitive rationality). I'll argue in this section that it cannot be a task norm—even though it's more plausible that this is a cognitive norm (see §4).

The first reason to reject that the Bayesian norm is a task norm for the WST is that it *contradicts* the task instructions. After all, the task instructions explicitly require each subject "to indicate which of these four cards they must turn over in order to determine whether various conditional sentences are true". Selecting the card that maximizes expected information gain wouldn't fulfill these task instructions. Moreover, this would be one of infinitely many arbitrary ways to respond to the WST by violating its instructions. Each of these ways is best interpreted as a deviation from the response that correctly followed these task instructions. Therefore, there is no non-arbitrary relation between the WST and a response that satisfies Oaksford & Chater's (2007) equations.

The second reason to reject that the Bayesian norm is a task norm is that the solution for the WST given that norm is *indeterminate*. Their equations specify the expected amount of information gain is some function of the following relata: (a) the probability of the antecedent (e.g., that the card

has an A face), (b) the probability of the consequent (e.g., that the card has a 2 face), (c) the prior probability that the conditional sentence is true, and (d) the error parameter. For example, Figure 1 from Oaksford & Chater (2007) illustrates the probability that each card should be turned over as a function of the independent probabilities of the antecedent and the consequent, given a prior probability of 0.5 that the conditional sentence is true and an error parameter of 0.1.

The critical point here is that *none* of these values are specified by the WST itself. As a result, there are infinitely many ways to specify the relata that are consistent with the WST. Even worse, there are infinitely many ways to specify the relata such that each of the four cards ends up being the card whose turning maximizes expected information gain. Just consider Figure 1: for each of the four cards, there are whole continuous regions in the probability space for the antecedent and the consequent such that turning the card is what maximizes expected information gain. Moreover, the task doesn't limit subjects to turning over only one card, so cost functions must further be specified to limit the number of cards that subjects turn over. These costs aren't determined by the task either, so there are infinitely many ways to do that too. Thus, there are infinitely many ways to justify each of the 16 possible responses to one instance of the WST.

Note that indeterminacy is a problem for task norms—not for cognitive norms. After all, suppose that we tried to show that Oaksford & Chater's (2007) equations identify a cognitive norm, which specifies the rational way for subjects to respond to the WST. The values in their equations that are undetermined by the WST would be free parameters in our cognitive model. But that wouldn't be a problem: we'd assign values to those free parameters such that they minimize the error by which the model predicts subjects' responses. We'll consider this process further in §4. But recall that task norms categorize responses in relation to the task only (not the cognition of any human reasoner): there is no way to determine a unique solution to the WST given the intrinsic properties of the task and the Bayesian norm. So, we'd be conflating task norms with cognitive norms if we used response data to specify values for these free parameters.

As a result, the Bayesian norms can't be task norms for the WST. To illustrate why, just consider how we'd need to modify the WST such that the Bayesian norms are task norms for the WST:

> **Bayesian Wason Selection Task (WST$_B$):** in each round, participants are presented with four cards that each have a number on one side and a letter on the other side, but they can only see one side of each card: e.g., A, K, 2, and 7. Next, participants are asked to indicate which of these four cards they should turn over in order to maximize the expected earnings of information gain minus the cost of flipping given:
> 1. The probability that a card has an A on one side is 0.2,
> 2. The probability that a card has a 2 on one side is 0.4,
> 3. The prior probability that "if there is an A on one side, then there is a 2 on the other side" is 0.5,
> 4. The error parameter is 0.1, and
> 5. Only one card can be flipped.

First, the task instructions clearly state that the solution to the task is the response that maximizes expected information gain, per the Bayesian norm.[5] Second, it specifies the values for the requisite variables such that there is a unique solution that maximizes expected information gain: flip the A card only. Thus, the Bayesian norm is the task norm for the $WST_B$. The differences between the WST and $WST_B$ clearly illustrate why the Bayesian norm can't be a task norm for WST.

Adjudicating between classical propositional logic and the maximization of expected information gain on the WST is easy because the former is consistent with the task instructions and the latter isn't and because the former has enough information in the task to define a solution and the latter doesn't. These illustrate two important standards for arbitrating task norms. The point should be obvious on reflection. But it's obscured by the fact that the psychology of reasoning tends to be more interested in cognitive norms than task norms and hence, tends to conflate task norms (and cognitive success) with cognitive norms (and cognitive rationality). Our example demonstrates how problematic this can be: treating task norms as cognitive norms may lead to contradiction and indeterminacies in task design.

## §3. Norm Indeterminacy

One might grant that we don't face a problem with arbitrating between classical logic and Oaksford & Chater's (2007) Bayesian norm for the WST. Still, the problem of arbitrating between norms won't always be so easy. Here's a harder case: classical logic and trivalent logic are both consistent with the instructions for the WST and provide determinate answers given the information provided by the WST. *Modus tollens* is valid in classical logic but invalid in certain forms of trivalent logic (henceforth, just "trivalent logic", for brevity) (Égré & Rott, 2021). If we interpret the conditional sentence as a sentence in classical logic, then the correct response is to turn both the A and 7 cards. But if we interpret the conditional sentence as a sentence in trivalent logic, the correct response is to turn the A card only. In this section, I'll propose a solution to this harder arbitration problem, and I'll defend it against an objection from subjectivism (a form of soft normativism).

Arbitrating between classical and trivalent logic is a serious problem. Recall that we can't solve it *a posteriori*—e.g., by observing which norm is preferred by more intelligent subjects (Stanovich & West, 2000). After all, it is a problem of adjudicating between task norms, which are intrinsic to the tasks and hence, indifferent to the externalities of human reasoning (and other processes of selecting responses). Instead, we have to solve it *a priori*. One way to do this would be to invoke some further standard to arbitrate between classical logic and trivalent logic. But classical and trivalent logics are formal norms: barring logical realism, they don't purport to be about anything, e.g., the logical structure of reality. So, there's no further standard that we could use to arbitrate between them. Hence, the arbitration problem seems to be insoluble (Elqayam & Evans, 2011).

I propose that the arbitration problem only seems insoluble because it's been misinterpreted as an arbitration problem. A better interpretation is that it's an indeterminacy problem: the instructions in the WST aren't sufficiently clear to specify whether classical logic or trivalent logic determines

[5] Oaksford & Wakefield (2003) design a modification of the WST that partly addresses the indeterminacy problem: it specifies the probabilities of the antecedent and consequent. However, their modified task fails to address the contradiction problem: their task instructions still ask subjects to evaluate the truth of the conditional, rather than asking them which card would maximize information gain.

what counts as a solution to the task—even though they are sufficiently clear to specify that the Bayesian norm doesn't. We don't have to uncover the logical structure of reality or find a further standard to arbitrate among logics. We can arbitrarily select one logic (e.g., classical propositional logic for Wason) and then clearly specify in our task's instructions which logic determines the solution to the task. More on the arbitrary selection of task norms in §5.

Clearly mentioning which norms determine the task's solutions will require care. Wason couldn't have explicitly stated that the norm in use is classical propositional logic, because most subjects wouldn't have known what 'classical propositional logic' refers to. Instead, he should have done this implicitly. For example, he could have provided examples of *modus tollens* and told them that these were examples of valid inferences.[6] This metalogical sentence would have confirmed that the task required a solution under classical logic, not trivalent logic. Of course, this might not make a difference to how subjects respond, since subjects might not be sensitive to the differences between classical and trivalent logic. Still, it would provide subjects with sufficient information to find a determinate solution to the task. It would ensure that classical logic is *the* task norm.

Till now, I've argued that experimenters can remove all norm indeterminacy by clearly specifying in the task instructions which norm determines what counts as a solution to the task. An important objection is that this disqualifies research that investigates how subjects use their own normative commitments to resolve indeterminacies in task norms. In this case, norm indeterminacies are a feature of the task—not a bug.[7] I agree that my approach should be rejected if it implied that no well-designed task could permit the norm indeterminacies required for these studies. However, I will argue here that it has a more nuanced implication for these tasks: only the highest-order norm in every well-defined task must be completely specified by the task to prevent a problematic kind of norm indeterminacy.

A recent study by Skovgaard-Olsen et al. (2019) is an excellent example of a well-defined task that asks how subjects resolve norm indeterminacies.[8] Their study aimed to study reasoning about conditionals while respecting and preserving individual differences in intuitive interpretations of conditionals. After all, there are at least two ways to interpret conditionals in natural language of the form 'If A, then C'. The *suppositional theory* (ST) claims that the probability of 'If A, then C' is equal to the extent to which A increases the conditional probability of C: P(if A, then C) = P(C | A). The *default and penalty theory* (DP) claims that ST is true only when A and C are relevant—when P(C | A) > P(C | ~A). Otherwise, if A and C aren't relevant (which violates the "default" expectation), then the probability of 'If A, then C' is equal to the extent to which A increases the conditional probability of C minus an irrelevance penalty ($p$): if P(C | A) ≤ P(C | ~A), then P(if A, then C) = P(C | A) – $p$.

Skovgaard-Olsen et al. (2019) designed their task such that the interpretation of conditionals was indeterminate: they never specified (implicitly or explicitly) whether the conditionals should be interpreted via ST or DP. Critically, though, they defined a further norm. First, they had subjects compare ST and DP justifications of conditionals written by fictional players and asked subjects

[6] To mitigate priming, he could provide them examples of *modus ponens* and other distractor inferences.
[7] I thank Reviewer 1 for raising this important challenge and for offering Skovgaard-Olsen et al. (2019) as an excellent example of such a task.
[8] For another example, see Elqayam's (2012) proposed subjectivist modification to Stanovich & West's (2008) task.

to penalize the fictional players whom they thought gave the inferior justification. Second, they asked subjects to evaluate inferences that were valid according to either ST or DP but not both. Then subjects were then evaluated as *correct* if they evaluated inferences consistent with the norm that they used to evaluate the justifications by fictional players and *incorrect* otherwise. This is a *no-hypocrisy norm*: it evaluates whether subjects used ST after they penalized DP, and vice versa.

In this design, there are three task norms that fall under two kinds. ST and DP are both *first-order norms*: they evaluate the validity of inferences over conditionals, not the use of further norms. Moreover, ST and DP are genuinely indeterminate in the task: evaluations of justifications and conditional inferences can be categorized as correct using either ST or DP using the information specified in the task without violating the task instructions (which are silent about how to interpret the conditionals). By comparison, the no-hypocrisy norm is a *second-order norm*: it evaluates whether these first-order norms (ST and DP) are used consistently across different phases of the experiment. Critically, Skovgaard-Olsen et al. (2019) design the task using a single second-order norm so that there was only one way to categorize responses as correct or incorrect. Moreover, they don't seek the endorsement of this norm from the subjects.

I propose that the task could be well-designed despite indeterminacy between first-order norms so long as the second-order norm is determinate. Is the second-order norm determinate? Suppose that someone challenged the no-hypocrisy norm. Imagine that certain subjects believed that ST was a more parsimonious norm and preferred it for theoretical use while simultaneously believing that DP was a more intuitive norm and preferred it for everyday use. Suppose also that they felt that they themselves should be held to the higher standard of theoretical use (ST) while it would only be fair to hold others to the everyday use (DP).[9] Skovgaard-Olsen et al. (2019) didn't check whether subjects endorsed the no-hypocrisy norm, so it's possible that subjects would have endorsed this alternative norm. If we found evidence for this, we could reject that inconsistencies in uses between phases of the experiment are best understood as hypocrisy errors. This would create another indeterminacy problem, and the task would no longer count as well-defined.

There are two ways that we could modify the task in response so that it is well-defined. First, we could implicitly or explicitly specify in the task instructions that the same norm should be used across all phases of the experiment. Then subjects who would have held themselves to "higher" standards than their peers would know that this wouldn't count as solving the task. Second, we could define a third-order norm that prohibited holding others to higher standards than oneself but permitted holding others to the same or lower standards than oneself. Again, though, the third-order standard may face challenges from other third-order norms. Eventually, then, a well-defined task will have to end the regress by specifying in the task instructions a single highest-order norm that ultimately categorizes whether responses are correct or incorrect.

This is an *objectivist* solution: whether a norm applies to the subjects when they respond to any well-defined task depends on whether it is specified by the task, not on whether subjects have avowed that norm. This contradicts the notion of *subjectivism* (or relativism) in descriptivism (Elqayam & Evans, 2011) and soft normativism (Stupple & Ball, 2014): whether a norm applies to subjects when they respond to any well-defined task is dependent on whether they've implicitly or explicitly avowed that norm. Avowals are insufficient to determine norms, because avowals are

[9] I expect that many philosophers experience the use of such double standards.

fallible: they may be inconsistent, so higher-order norms are needed to evaluate consistency among avowals. This generates a regress that can only stop when we resort to using norms that subjects haven't avowed and building those norms into our tasks (such that they aren't fallible).

This creates a trilemma for subjectivists. First, they can give up on using norms altogether: they can merely describe avowals without evaluating whether they are consistent or correct. The result is descriptivism. Second, they can insist on using norms but limit themselves to using norms that subjects have avowed. The result is soft normativism. But then they face an infinite regress of avowals from subjects on higher- and higher-order norms. This is untenable. Third, they can give up subjectivism and allow themselves to use norms that subjects haven't avowed. The result is hard normativism. Then they can easily end the regress by designing their tasks so that they specify a single determinate norm that subjects haven't avowed. This norm may be first-order or any other order. Despite their subjectivist ambitions, Skovgaard-Olsen et al. (2019) ultimately designed a task with a hard, second-order norm, consistent with hard (and pluralist) normativism.

## §4. Explaining Success & Failure

I've argued that any well-designed task has instructions that specify the highest-order task norm. Ultimately, though, the psychology of reasoning isn't interested in task norms. That's the domain of normative theory, like classical logic, elementary algebra, probability theory, etc. Rather, it's interested in cognitive norms. This raises an important question: how does the correct use of *a priori* task norms during task design facilitate the identification of *a posteriori* cognitive norms during data interpretation? How would indeterminacies in task norms prevent us from identifying the rational standards for the exercise of cognitive capacities in human reasoning? In this section, I'll argue that different kinds of cognitive explanations are warranted for successes, fragile failures, and robust failures, and task norms are necessary to sort responses under these categories.

First, success indicates the *rational* exercise of *relevant* cognitive capacities. As a result, cognitive explanations of successful responses are the simplest kind. We begin by identifying the simplest algorithm that explains why the response is categorized as correct for the task norm. For the CRT, we identify the algorithms that progressively simplify the task equations to their unique solutions using the norms of elementary algebra. For the WST, we identify the algorithms that apply *modus ponens* and *modus tollens* to identify which card could falsify the conditional. For the $WST_B$, we identify the algorithms that progressively simplify the task equations to their unique solutions using the norms of Bayesian inference. Next, we attribute those algorithms to cognition: we infer that the best explanation of success is that cognition must be implementing those norms.

That might not be the true explanation, of course, but it is the most likely given the evidence of success. After all, suppose that we did see subjects achieving high rates of success on the CRT. The best explanation does seem to be that they acquired the capacity for algebraic reasoning in their secondary education and the rational exercise of that capacity amounts to correctly following the taught rules of algebra as they apply to this task. In a sense, then, the structure of the cognitive response is just an artefact of the structure of the task. As such, this is the least interesting outcome: we confirm that human cognition is following rules that we already understand and in fact, that we've taught to it. We gain relatively little information in this outcome: the experimental output is little more than an artefact of our experimental inputs.

Second, fragile failure indicates the *irrational* exercise of *relevant* cognitive capacities. Fragile failure is easily disrupted: minor modifications to the task restore success. Cognitive explanations of fragile failure aim to "rationalize" the irrational exercise of relevant cognitive capacities. First, we again identify the simplest algorithm that explains why the response is categorized as correct for the task norm. Then we assign costs and constraints to each step of this algorithm such that the actual incorrect response is the response that maximizes the benefit-to-cost ratio under the constraints. Finally, we attribute those algorithms to cognition: we infer that the best explanation of fragile failure is that cognition must be failing to exercise its capacity for algebra because it is conserving processing costs under its constraints.[10]

The CRT is an excellent example: minor modifications to the CRT are known to restore success.[11] This suggests that subjects do have a cognitive capacity for algebra but they're simply failing to correctly exercise it for the CRT. The most popular explanation of failure on the CRT is that subjects use *substitution heuristics*: they replace the task with a simpler task that is easier to solve using the norms of elementary algebra (Kahneman & Frederick, 2005). This can be rationalized: it would be too costly for reasoning to correctly implement each step of the algorithm that solves the task while following the norms of elementary algebra. Instead, it implements the most similar algorithm that solves the task within its budget: an algorithm that substitutes the task with a similar task and then solves the substitute task while following the norms of elementary algebra.

Fragile failure is the most informative outcome for studying cognitive control. After all, cognitive control has the function to correct errors in the exercise of cognitive capacities. If cognitive failure is the result of irrationality, then task norms and cognitive norms are aligned for the task. And if that's so, then cognitive control should have recognized the task errors as cognitive errors but failed to do so. Hence, fragile failures tend to involve failures in cognitive control. This explains why the CRT has been so popular in research on cognitive control (see De Neys, 2012, 2014 for reviews). For example, I've argued that recent studies using the CRT (e.g., Hoover & Healy, 2017, 2019, 2021) provide the most effective ways to isolate errors in metacognitive control, which functions to modulate the exercise of the capacity for algebraic reasoning (Dewey, 2022).

Moreover, cognitive explanations of fragile cognitive failure will refine cognitive explanations of success. A lot of evidence suggests that *all* subjects detect conflict as they respond to the CRT, regardless of whether they gave correct or incorrect answers (De Neys, 2012, 2014). This indicates that all subjects initially use the substitution heuristic, notice it, and detect the conflict between their accurate and inaccurate representations of the CRT (Hoover & Healy, 2017). But the minority of subjects who correctly solve tasks like the CRT show significantly more cognitive control (e.g., De Neys et al., 2008; Simon et al., 2015). This indicates that only a small minority of subjects are able to selectively respond to the accurate representation of the CRT after they detect conflict between the two representations of the CRT (Thompson, 2009; De Neys, 2012, 2014; Thompson et al., 2011, 2013; Pennycook et al., 2015).

[10] Different forms of rational analysis offer different formalizations of these inferences (Anderson, 1990; Russell, 1997; Oaksford & Chater, 2007; Howes et al., 2009, 2016; Lewis et al., 2014; Griffiths et al., 2015; Lieder & Griffiths, 2021).

[11] These slightly modified tasks are often used as controls for the CRT (see De Neys, 2012, 2014 for reviews).

Third, robust failure indicates the *rational* exercise of *irrelevant* cognitive capacities. Robust failure isn't easily disrupted: major task modifications preserve failure. Cognitive explanations of robust failure aim to "rationalize" the exercise of irrelevant cognitive capacities. First, they infer that the subject lacks the relevant cognitive capacity or fails to recognize that it is relevant to the task. Second, they identify cognitive capacities that the agent does possess and that might seem relevant for the task. Third, they reinterpret the task such that it would be relevant to exercise that cognitive capacity. Finally, fourth, they attribute the exercise of a cognitive capacity to a subject only if the observed responses would be rational if they were responding to the reinterpreted task.

The WST is a good example of a task that elicits robust failure: subjects continue to fail at high rates across several (not all) modifications.[12] [13] Oaksford & Chater (2007) develop a compelling (albeit disputed: see Klauer et al., 2007; Ragni et al., 2018) explanation for this. They argue that *modus tollens* is a kind of falsification and that the history and philosophy of science reveal that falsification isn't a useful strategy for testing claims: false predictions suggest that at least one premise (antecedent) is false, but they never indicate which.[14] Instead, it's better to investigate options that maximize expected information gains when testing claims in our natural environment with our limited computational resources. If that's true, then subjects who are presented with the WST will register the conditional as requiring testing and they will fulfill that need by deploying and exercising their useful capacity for maximizing expected information gain—not a capacity for *modus tollens*.

To test this hypothesis, Oaksford & Chater (1994, 2007) build a formal model of the hypothetical capacity for maximizing expected information gains. They recognize that the model calls for more information than is specified in the WST, but they reasonably infer that the subject will draw on their own experience (outside the task) to specify the underdetermined information themselves. Next, they reverse engineer the way that subjects specify this underdetermined information by trying to match their behaviour. For example, they showed that the rational exercise of this capacity would cause subjects to turn over the A card at a frequency of about 0.9 and the 2 card at a frequency of about 0.6—given that subjects assume that (a) the probability that a card has an A on one side is sufficiently low (0.22), (b) the probability that a card has a 2 on one side is sufficiently low (0.27), (c) the prior probability of the conditional sentence is 0.5 (in accordance with the principle of indifference), and (d) the frequency of exceptions to the rule is 0.1. Their model achieved a high fit with data across 34 studies.

[12] Subjects still fail at high rates if the antecedent and consequent propositions attribute arbitrary properties to the rule and cards (alphanumeric, geometric, color, etc.) (Wason, 1969) or if they attribute meaningful properties (Manktelow & Evans, 1979). But cognitive success can be partially restored if the antecedent and consequent propositions describe social rules in the context of enforcement (Cox & Griggs, 1982; Pollard & Evans, 1987).

[13] In principle, we could say that subjects are deploying their relevant capacity for modus tollens but are *somehow* robustly failing to rationally exercise it (Wason & Johnson-Laird, 1972). But this seems implausible and uncharitable (Cohen, 1981; Stich, 1990; Schurz & Hertwig, 2019).

[14] For example, they cite Putnam's (1974) example that perturbations in the orbit of Uranus were known to contradict the prevailing Newtonian model of the solar system, but it wasn't clear which part of the model was false until Neptune was found. By comparison, perturbations in the orbit of Mercury were also known to contradict the prevailing Newtonian model of the solar system, and it wasn't clear which part of the model was false until Newton's theory of gravity was supplanted by Einstein's theory of relativity (vs. the discovery of the hypothetical planet Vulcan). So, they conclude that it wouldn't be useful for us to acquire the capacity for falsification or *modus tollens*.

Oaksford & Chater's (2007) rational analysis of human performance on the WST demonstrates that robust failure is the most informative outcome for studying cognitive capacities. After all, it's an outcome where subjects exercise a new cognitive capacity that operates according to new rules that are independent from the rules that we used to design the task. These are rules that we haven't taught to subjects—rules that we don't already understand.[15] Critically, though, our discovery of these new cognitive norms is significantly facilitated by our choice of task norms: the cognitive capacities that are exercised are similar to, but different from, the cognitive capacities that would have been relevant to the task. Therefore, task norms constrain the space for unknown cognitive capacities, making it easier to search for and hence, discover related cognitive capacities. Without task norms, we'd have to reverse engineer cognitive norms *ex nihilo*.

## §5. Search for Cognitive Norms

I've argued that the choice of task norms guides and facilitates the discovery of cognitive norms. This might seem to raise a new arbitration problem: we must arbitrate among possible task norms, select one, and then design a task with that norm.[16] For example, Wason had to implicitly decide whether to use the WST and its norm of classical propositional logic or the $WST_B$ and Oaksford & Chater's (2007) norm of maximizing expected information gain. In this section, I'll argue that there is no problem with arbitrating among possible task norms. After all, the choice of task norms during the task design phase can't be informed by the discovery of cognitive norms. Moreover, our explanatory responses to success, fragile failure, and robust failure bring about convergence between task norms and cognitive norms regardless of how different they are at first.

Let's consider the decision that Wason (1968) faced when he aimed to study how human subjects test conditional claims. Before he could design a task for testing conditional claims, he would have to choose a norm that specifies what counts as a solution to the task. There were arbitrarily many norms for him to choose from, but let's imagine that he only had to arbitrate over the three norms that we've considered in previous sections: classical logic, maximizing expected information gain, and trivalent logic. Which norm should he have chosen? This is a different kind of arbitration problem from the one that Elqayam & Evans (2011) raise: it's a question about which norm should be used to design a task, not a question about which norm should be used to evaluate responses to a task that's already been designed.

First, let's consider the merits of Wason's actual choice to design and implement the WST using the norms of classical propositional logic. He found that robust failure was the most frequent outcome. Recall from §4 that this was the most informative outcome: it suggested that human subjects responded to the WST by exercising an unknown cognitive capacity. In effect, they used a *substitution heuristic*: they translated the WST into some modification of the WST and then successfully solved that modified WST. Figuring out how subjects reinterpret the WST is a

[15] This explains why robust failures aren't useful for studying cognitive control: cognitive control has the function to correct errors in the exercise of cognitive capacities. But robust failure is the result of the rational exercise of irrelevant cognitive capacities, so cognitive control shouldn't recognize the task errors as cognitive errors, and so shouldn't intervene. Since cognitive failures aren't failures in cognitive control, there is little information that we can gain from them about cognitive control.

[16] This is an *across-task* arbitration problem: we have to arbitrate between well-defined tasks, each of which has a single determinate norm. This is distinct from the problem that Elqayam & Evans (2011) raise, which is a *within-task* arbitration problem: we supposedly have to arbitrate across multiple norms for a single task.

difficult task. Oaksford & Chater (1994) argue that subjects translate the WST into something like the $WST_B$ and then rationally exercise their capacity for maximizing expected information gains. By comparison, Klauer et al. (2007) use multinomial processing tree analysis to argue that different subjects respond to the WST with different strategies—effectively translating the WST into a variety of other problems.[17]

Despite its difficulty, this experimental outcome requires learning by researchers: designing a task using task norms that don't correspond to any cognitive norms initiates a difficult learning process that eventually converges on cognitive norms that are relatively close (i.e., similar) to the chosen task norms. The distance between the chosen task norms and the nearest cognitive norms determines the difficulty of the learning process. Suppose for the sake of argument that Oaksford & Chater (2007) are right (contra Klauer et al., 2007). Then the norms of classical propositional logic were quite far from the Bayesian norms for maximizing expected information gain, but they were still both norms for testing conditionals. Now, Wason couldn't have known that this outcome would obtain when he chose the norms of classical propositional logic to design the WST. Even with the benefit of hindsight, though, this was a good outcome: it initiated a difficult learning process that resulted in the most informative kind of cognitive explanation.

Second, let's imagine that Wason had designed a well-defined task that fully specified the norm for maximizing expected information gain, such as $WST_B$. He might have found that humans achieve a high rate of success on the $WST_B$. After all, human responses seem to conform to the Bayesian norm for maximizing expected information gain on the WST (Oaksford & Chater, 2007; cf., Klauer et al., 2007; Ragni et al., 2018) and on a modified WST that specifies information about the probabilities of the antecedent and consequent (Oaksford & Wakefield, 2003; cf., Oberauer et al., 1999). This would indicate that humans have a cognitive capacity for maximizing expected information gain and hence, can be expected to deploy that capacity and succeed on a task like the $WST_B$. If that's so, then Wason would have had to explain that human subjects respond to the $WST_B$ by identifying and selecting the response that maximizes expected information gain.

Still supposing that Oaksford & Chater (2007) are right for the sake of argument (contra Klauer et al., 2007), this outcome would be least informative because Wason would have *correctly* guessed that humans test conditionals by exercising a capacity for maximizing expected information gain and then simply confirmed his guess. Afterwards, a follow-up study using the WST would have been required to show that human subjects (a) lack the capacity for *modus tollens* and (b) exercise their capacity for maximizing expected information gain instead. That process would have been much easier: it would be obvious how to rationalize robust failure on the WST if we'd already observed robust success on the $WST_B$. But this ease would be due to Wason's lucky guess: during task design, he couldn't have known that human reasoning would conform to Bayesian norms for maximizing expected information gain. Even with the benefit of hindsight, though, it's not clear

[17] Klauer et al. (2007) and Ragni et al. (2018) both show that Oaksford & Chater's (1994) model fails to sufficiently explain all 16 possible types of responses to the WST. In a sense, this isn't surprising: Oaksford & Chater attribute the same reasoning response to every subject, which is a very strong assumption. By comparison, Klauer et al. and Ragni et al. attribute different reasoning responses to different subjects using multinomial processing tree models. This is a much weaker assumption, which significantly increases the flexibility of their models. Still, Oaksford & Chater's model is simpler (it uses 4 parameters rather than 10), so it will be easier for me to discuss its advantages vis-à-vis the classical logic model. Hence, I'll focus on their model, even though my point would be the same if I had focused on Klauer et al. and Ragni et al.'s model instead. I thank Reviewer 1 for pressing this point.

that choosing Bayesian norms would have been better: it would be a lucky guess that found a correct explanation, albeit the least informative kind of cognitive explanation.

Third, let's imagine that Wason had designed a well-defined task that fully specified the norms of trivalent logic. Interestingly, he would have found a high rate of success: *modus ponens* is valid and *modus tollens* is invalid in trivalent logic and most subjects only reason in accordance with *modus ponens*. So, Wason should have concluded that subjects respond to the WST by computing derivations in trivalent logic. Although that explanation would be the most likely given the data, it would be false. We know this because it wouldn't explain why responses are sensitive to, e.g., the probabilities of the antecedent and consequent (Oaksford & Chater, 2007) or the direction of the conditional (i.e., forward, as in "if *p*, then *q*", vs. backward, as in "*q*, only if *p*") (Klauer et al., 2007). So, this outcome would not only have been least informative, but it would have been misleading. However, this wouldn't have been Wason's fault: he couldn't have known in advance that the norms of trivalent logic would be misleading—it would just be bad luck.

Fortunately, though, it would be possible to correct this unlucky error. Wason would have had to look for modifications to the WST that disrupted the high rate of success, given trivalent logic. With trial and error, he might have discovered that he could elicit robust failure by designing variants of the WST that specified and varied the probabilities of the antecedent and consequent (Oaksford & Wakefield, 2003). With more trial and error, he might have found that subjects actually conform to the Bayesian norms of maximizing expected information gain. The entire process would be even more difficult than if he had used the WST variant that specified the norms of classical logic. But difficulty corresponds to learning: Wason would have been less lucky to start with trivalent logic than classical logic, so he would have had to learn more to finally conclude that subjects respond by maximizing expected information gain.

These three cases demonstrate that each choice of task norm is compatible with learning that the cognitive norm for human reasoning about testing conditionals conforms to the Bayesian norm for maximizing expected information (again, assuming that Oaksford & Chater's model is correct). Of course, some searches will be more efficient than others, but that's trivially true: if we start out with hypotheses that better approximate the truth, we'll get at the truth faster than if we had started out with hypotheses that approximated the truth worse. At the outset of inquiry, though, we can't really know which task norm is the best approximation of the cognitive norm. A lot of it comes down to luck.

This explains why there isn't a problem of arbitration *across* tasks: we don't have to justify the choice of norm that we use to design initial tasks. Despite what Evans & Elqayam (2011) might suggest, Wason is blameless for designing tasks using the norms of classical logic to study human reasoning about conditionals. His guess might have been wrong, but guesses are often wrong: he designed a task for testing conditionals before much evidence was available about the context, constraints, and goals of human reasoning about conditions. Moreover, Wason's incorrect guess triggered a learning process that led to Oaksford & Chater (2007) finding (contested) evidence that human subjects use Bayesian norms to reason about conditionals and Klauer et al. (2007) finding (contested) evidence that human subjects use a variety of strategies to reason about conditionals.

Wason's guess could have been misleading, which would have been worse, but it would have triggered a longer learning process that would still eventually have culminated in the truth. After all, our strategy for explaining success, fragile failure, and robust failure is an error-correcting search process: it searches for similar task norms that make relatively minor modifications to an existing task until the rates of robust failure reduce to zero. When the rates of robust failure converge on zero, the result is an explanation of robust failure on the original task: subjects robustly failed on the original task because they substituted it with the modified task and then correctly solved that modified task. Obviously, the greater the distance between the task norm and the nearest cognitive norm, the longer the search process tends to be. Again, though, the distance between the task norm and the nearest cognitive norm is unknown prior to the search.

For descriptivists and soft normativists to argue that we must justify the choice of initial task norms is impractical and counterproductive—if it arbitrarily constrains or otherwise discourages the design of novel tasks. So, we should insist that there are no constraints on arbitrating task norms during task design at the start of a new search for cognitive norms. But if one is continuing an ongoing search for cognitive norms, then we should insist on only one constraint for selecting task norms: task norms must be similar to but different from task norms used to design tasks that elicited high rates of robust failure. After all, this is the only condition necessary for continuing a blind search for task norms that we need to design tasks that elicit lower rates of robust failure.[18]

## §6. Conclusion

The H&B program has been criticized for categorizing human responses to reasoning tasks using the norms of classical logic, elementary algebra, probability theory, etc., despite the fact that these norms aren't sensitive to the context, constraints, and goals of human reasoning. In this paper, I've argued that this criticism is confused. The design of novel tasks requires the selection of norms, but the selection of norms precedes the discovery of the context, constraints, and goals that causally influence human reasoning, so it can't be sensitive to these features. Hence, norms for novel tasks tend to be *hard*: they tend to define correctness such that there isn't a strong correlation between correct and actual responses.

The use of these task norms has been criticized for raising the arbitration problem: responses to any task can be evaluated using any number of task norms and there isn't a further standard that can justify selecting any single task norm. I've argued that this is false. Task norms specify what count as solutions to a task. A well-defined task will specify in unambiguous terms what counts as a solution and the information required to identify determinate solutions. One objection to this solution is that it rules out studies about how subjects use their own normative commitments to resolve norm indeterminacies in tasks. I've argued that this is mistaken: a well-designed task may permit indeterminacies among lower-order norms so long as it specifies a determinate higher-order norm, which specifies in unambiguous terms what counts as a solution to the overall task.

[18] Finally, if one has found a task norm that designs tasks that elicit very low rates of robust failure, then we should test whether the task norm corresponds to a cognitive norm by searching for non-normative modifications to the task (which hold the task norm fixed) that restore high rates of robust failure. This can correct for misleading task norms, as in the case where high rates of success can be observed on the WST under trivalent logic even though subjects are maximizing the expected information gains.

Ultimately, the choice of task norms makes a difference to the strategy that we use for identifying cognitive norms. I've argued that task norms distinguish among three outcomes: success, fragile failure, and robust failure. Each requires a unique response: (a) success requires follow-up studies that vary non-normative features of the task to drive success down, (b) fragile failure requires some kind of rational analysis to develop a bounded rationality explanation, and (c) robust failure requires follow-up studies that vary the norms used to design the task in order to restore success. Although the choice of task norms makes an important difference to subsequent research, it can't be optimized: we cannot know in advance which choice of task norms will end up being optimal. This is a critical lesson: descriptivists and soft normativists may impede new research programs by discouraging the design of tasks using hard, *a priori* task norms.[19] On the contrary, we should encourage new research programs by promoting a laissez-faire approach to task norms.

## §7. Appendix: Is-Ought Inferences

Elqayam & Evans (2011) raise the objection that the psychology of reasoning cannot use norms without committing the *is-ought fallacy* (Hume, 1739). After all, it is impossible to draw a valid inference from premises about what subjects actually do to a conclusion about what they *ought* to do. Moreover, we can only observe what it is that subjects actually do—not what it is that subjects *ought* to do. Since norms are neither observable nor deducible from what's observable, Elqayam and Evans argue that norms are unempirical. So, Evans (2007: 161) concludes that "normative rationality is essentially a philosophical and not a psychological concept". This is a bold objection, but it should strike any philosopher as confused. Still, I worry that if it isn't addressed, it may do serious damage by discouraging psychologists from using *a priori* norms to create well-designed tasks, so I'll address it in this appendix.

The is-ought gap is an unremarkable example of a general property of classical deductive logic, known as *conservativity*: in any valid inference, the conclusion must contain only information that is already contained in the premises, such that it's impossible for the premises to be true and the conclusion to be false (Maguire, 2015).[20] If we limit what counts as empirical to (a) what is directly observable and (b) what can be deduced in classical logic from premises about what's directly observable, then normative claims won't count as empirical, yes, but neither will theoretical claims. Claims about causation, cognition, and normativity are all claims about unobservable theoretical entities, so none of these things will count as empirical on a deductive conception of empiricism. Hume (1739) was prepared to accept these radical implications—in fact, he built his entire philosophy around it—but most psychologists probably aren't.

[19] For example, I recently developed a novel kind of task design that explicitly requires the use of hard norms (Dewey, 2022). During the review process for that paper, though, I received strong criticism from some reviewers for using hard norms and was directed to Elqayam & Evans (2011). As a philosopher, though, I was confident with insisting on the use of normative assumptions, but I worried that non-philosophers might not be so confident. That convinced me to write the current paper, to offer an alternative to subjectivism, soft normativism, and descriptivism.

[20] To their credit, Elqayam & Evans (2011) are sensitive to this. They recognize that the is-ought gap can be closed without creating an is-ought fallacy by adding an "implicit normative premise", often known as a *bridge premise*. But they don't seem to be sensitive to the fact that this undermines their original concern: whenever we draw normative conclusions from empirical premises, it's always possible to rationalize our inference post hoc by adding an implicit normative bridge premise that validates our argument. For this reason, it seems quite infelicitous to accuse anyone of committing an is-ought fallacy. If we want to disagree with someone's is-ought inferences, it's much more felicitous to explicate their implicit normative bridge premise and then argue that the premise is *false*.

Moreover, Elqayam and Evans are critical of classical deductive logic, so it's a bit ironic that they commit to a classical deductive conception of empiricism. Instead, most philosophers now endorse non-deductive conceptions of empiricism, which invoke non-deductive norms of inference. For example, many empiricists today limit what counts as empirical to (a) what is directly observable and (b) whatever is posited by theories that we construct from inference to the best explanation of direct observations. This accounts for the empirical status of unobservable theoretical entities. For example, cognition counts as empirical despite the fact that it's unobservable because the best explanations of observable behaviour claim that there is such a thing as cognition. Normative entities are unobservable entities, so any non-deductive conception of empiricism that attributes empirical status to theoretical entities can be expected to attribute empirical status to norms too.[21]

Throughout this paper, I've taken the liberty to posit norms whenever there are good experimental and explanatory reasons for doing so. This moderate, laissez-faire approach to empiricism has been the orthodox view in the metaphysics of science since it was first developed in a seminal paper by Quine (1948). He argued that we should take the liberty to posit any unobservable entities and properties that we need to best explain what we observe and then accept that they actually exist. The motivating idea is that imposing further criteria on existence (e.g., that claims about entities must be deducible from observations) is arbitrary and impairs explanation. Quinean empiricism was a significant causal factor in the downfall of radical, deductive forms of empiricism, which were popularized by Hume and dominated the first half of the 20th century. Given this history, the radical empiricist arguments that Elqayam and Evans seem to be making against objective norms in psychology might strike modern philosophers as quite unmotivated.

A related version of this argument seems to explain the popularity of subjectivism (or relativism) (Elqayam, 2012; Stupple & Ball, 2014; Skovgaard-Olsen et al., 2019): the idea that it's better to evaluate a subject's behaviour using a norm that they themselves have endorsed than to evaluate it using a norm that the experimenter has chosen. After all, the is-ought gap seems smaller: not only do we observe a subject's behaviour, but we can also observe that the subject avows a certain norm. As a result, it might *seem* like less of an inferential leap to take that avowed norm and then use it to evaluate the subject's own behaviour. By comparison, it might *seem* like more of an inferential leap to single out a hard norm that the subject hasn't avowed and then use it to evaluate the subject's behaviour. If true, this would justify a general preference for soft vs. hard norms.

But appearances are deceiving: the inferential leap is greater in subjectivism. Subjectivism restricts the norms that are acceptable for the psychology of reasoning to use. To justify this, it requires a philosophical principle. That is, a subjectivist can only justify their use of a norm by appealing to two things: (a) the philosophical principle that it is best to evaluate behaviour using norms that subjects have avowed and (b) the empirical fact that a subject has avowed the norm that they are using. The appeal to empirical fact gives an empirical veneer to subjectivism that distracts from the fact that it ultimately requires commitment to a heavyweight philosophical principle. And this principle is dubious: I've argued in §3 that any well-defined task requires at least one objectivist norm, which subjects haven't been given the opportunity to avow.

[21] This is just an onus-shifting argument: if someone believes that norms aren't empirical (in some sense) but other unobservable things are, then the onus is on them to show that there is a consistent, plausible way to do this.

By comparison, the pluralist normativism that I've advocated here permits the psychology of reasoning to use any task norms—both objectivist and subjectivist—to design tasks. After all, I've argued that the choice of any task norm can be justified just by appealing to the empirical fact that any difference (error) between the task norm that we've chosen and the cognitive norm that we hope to identify will be progressively decreased (corrected) by careful rational analysis. Since pluralist normativism doesn't restrict the use of norms that are acceptable for the psychology of reasoning to use, it doesn't require a heavyweight philosophical principle to do so. Therefore, it requires a smaller inferential leap from the descriptive to the normative than subjectivism does. Even if we did find some way to justify caring about reducing the is-ought gap, then, pluralist normativism still turns out to be superior to soft normativism.

This debate reveals a difficulty that psychologists will face when deciding how to study reasoning: they will need to take positions in ongoing debates in the philosophy of normativity. Unfortunately, this creates the risk of what Ballantyne (2019) helpfully calls *epistemic trespassing*, which occurs when experts in one field pass judgments on live debates in other fields. I worry that descriptivists, soft normativists, and relativists have all been guilty of this. After all, there are ongoing debates in the philosophy of normativity about whether general principles (hard norms) or contextual norms (soft norms) are better for evaluating reasoning (for reviews, see Dancy, 2017; Ridge & McKeever, 2020), and what's necessary for a norm to be binding to an agent (for two influential discussions, see Korsgaard, 1996; Enoch, 2014).[22] To endorse philosophical principles that restrict the use of norms in the psychology of reasoning without acknowledging these live debates is to epistemically trespass on the philosophy of normativity.

Of course, epistemic trespassing is a common mistake: philosophers and psychologists are often guilty of it, and I'm certain that I am too, despite my best efforts to avoid it. Ballantyne (2019) even admits that it may be an inevitable by-product of interdisciplinary research. Nevertheless, the risks of epistemic trespassing can be mitigated. Psychologists who take positions on the study of reasoning should identify the corresponding positions that they are taking in ongoing debates in the philosophy of normativity. Then they should aim to be sensitive to arguments on both sides of these debates. It would probably be prudent for them to maintain neutrality in these debates too in order to defend a decisive course of action (which is what I've tried to do in this paper). This can be a lot of work: the philosophy of normativity is a difficult and enormous literature. One of the easiest ways to do this is to form collaborations with philosophers of normativity, who have already done some of this work. I warmly encourage more of these collaborations in the future!

**Acknowledgements:** I thank Reviewer 1 for their extensive comments and invaluable suggestions, which have significantly improved this paper. I also thank Sara Aronowitz and Mike Oaksford for their feedback, which helped shape earlier versions of this paper.

[22] Note that these resources are mostly focused on the norms used for evaluating *moral* reasoning (a favourite kind of reasoning among philosophers, including myself). However, similar arguments extend to other kinds of reasoning.

# References


Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Inc.

Ballantyne, N. (2019). Epistemic trespassing. *Mind*, *128*(510), 367–395. https://doi.org/10.1093/mind/fzx042

Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*: 317–70.

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition*, *10*(5), 496–502. https://doi.org/10.3758/BF03197653

Dancy, J. (2017). Moral particularism. E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2017/entries/moral-particularism/

Davies, P. S., Fetzer, J. H., & Foster, T. R. (1995). Logical reasoning and domain specificity. *Biology and Philosophy*, *10*(1), 1–37. https://doi.org/10.1007/BF00851985

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489. https://doi.org/10.1111/j.1467-9280.2008.02113.x

Dewey, A. R. (2022). Metacognitive control in single- vs. dual-process theory. *Thinking & Reasoning*, 1–36. https://doi.org/10.1080/13546783.2022.2047106

Elqayam, S. (2011). Grounded rationality: A relativist framework for normative rationality. In K. Manktelow, D. Over, & S. Elqayam (eds.), *The science of reason: A festschrift for Jonathan St B. T. Evans* (pp. 397–419). Psychology Press.

Elqayam, S. (2012). Grounded rationality: Descriptivism in epistemic context. *Synthese*, *189*(1), 39–49. https://doi.org/10.1007/s11229-012-0153-4

Elqayam, S., & Evans, J. S. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X

Elqayam, S., & Over, D. E. (2016). Editorial: From is to ought: The place of normative models in the study of human thought. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00628

Enoch, D. (2014). Authority and reason-giving. *Philosophy and Phenomenological Research*, *89*(2), 296–332.

Evans, J. St. B. T. (1993) Bias and rationality. In K. I. Manktelow & D. E. Over (eds.), *Rationality: Psychological and philosophical perspectives* (pp. 6–30). Routledge.

Evans, J. S. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. https://doi.org/10.1080/13546780601008825

Evans, J. St. B. T. & Over, D. E. (1996). *Rationality and reasoning*. Psychology Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. https://doi.org/10.1037/0033-295X.98.2.254

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Hattori, M. (1999). The effects of probabilistic information in Wason's selection task: An analysis of strategy based on the ODS model. In *Proceedings of the 16th Annual Meeting of the Japanese Cognitive Science Society*, *16*, 623–626.

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, *24*(6), 1922–1928. https://doi.org/10.3758/s13423-017-1241-8

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*, *6*(4), 369–380. https://doi.org/10.1037/dec0000107

Hoover, J. D., & Healy, A. F. (2021). The bat-and-ball problem: A word-problem debiasing approach. *Thinking & Reasoning*, *0*(0), 1–32. https://doi.org/10.1080/13546783.2021.1878473

Hume, D. (1739) A treatise of human nature. Being an attempt to introduce the experimental method of reasoning into moral subjects. Available at https://www.gutenberg.org/ebooks/4705

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008
Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*(2), 134–148. https://doi.org/10.1016/0010-0285(70)90009-5
Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus & Giroux.
Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). Cambridge University Press.
Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 680–703. https://doi.org/10.1037/0278-7393.33.4.680
Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.
Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311. https://doi.org/10.1111/tops.12086
Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X1900061X
Maguire, B. (2015). Grounding the autonomy of ethics. In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, vol. 10* (pp. 188–215). Oxford University Press.
Manktelow, K. I., & Evans, J. ST. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology*, *70*, 477–488. https://doi.org/10.1111/j.2044-8295.1979.tb01720.x
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608
Oaksford, M., & Chater, N. (2007). *Bayesian rationality the probabilistic approach to human reasoning.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524496.001.0001
Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition*, *31*(1), 143–154. https://doi.org/10.3758/BF03196089
Oberauer, K., Wilhelm, O., & Diaz, R. R. (1999). Bayesian rationality for the WST? A test of optimal data selection theory. *Thinking & Reasoning*, *5*(2), 115–144. https://doi.org/10.1080/135467899394020
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001
Pollard, P., & Evans, J. S. (1987). Content and context effects in reasoning. *The American Journal of Psychology, 100*(1), 41–60. https://doi.org/10.2307/1422641
Putnam, H. (1974). The 'corroboration' of theories. In A. Schilpp (ed.), *The philosophy of Karl Popper*, Vol. 2. La Salle, IL: Open Court.
Quine, W. V. (1948). On what there is. *The Review of Metaphysics, 2*(5), 21–38.
Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin, 144*(8), 779–796. https://doi.org/10.1037/bul0000146
Ridge, M. & McKeever, S. (2020). Moral particularism and moral generalism. E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2020/entries/moral-particularism-generalism/
Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1–2), 57–77. https://doi.org/10.1016/S0004-3702(97)00026-X
Samuels, R., Stich, S., & Bishop, M. (2012). Ending the rationality wars: How to make disputes about human rationality disappear. In *Collected Papers, Volume 2*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199733477.003.0009
Schurz, G., & Hertwig, R. (2019). Cognitive success: A consequentialist account of rationality in cognition. *Topics in Cognitive Science, 11*(1), 7–36. https://doi.org/10.1111/tops.12410
Simon, G., Lubin, A., Houdé, O., & Neys, W. D. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*, *6*(4), 158–168. https://doi.org/10.1080/17588928.2015.1036847
Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, *126*(5), 611–633. https://doi.org/10.1037/rev0000150
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Stich, S. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. Cambridge: MIT Press.

Stupple, E. J. N., & Ball, L. J. (2014). The intersection between Descriptivism and Meliorism in reasoning research: Further proposals in support of 'soft normativism'. *Frontiers in Psychology*, *5*, Article 1269. https://doi.org/10.3389/fpsyg.2014.01269

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 171–195). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0008

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology*,*60*, 471–80.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Harvard University Press.