Abductive Conditionals as a Test Case for Inferentialism*

Patricia Mirabile
SND / Sorbonne University
patricia.mirabile@sorbonne-universite.fr

Igor Douven
SND / CNRS / Sorbonne University
igor.douven@sorbonne-universite.fr

Abstract

According to inferentialism, for an indicative conditional to be true, there must be a sufficiently strong inferential connection between its antecedent and its consequent. Previous experimental research has found support for inferentialism, but the materials used concerned a fairly abstract context, leaving open the question of how accurately the account can predict semantic judgments about more realistic materials. To address this question, we conducted three experiments using abductive conditionals, which are conditionals featuring an explanatory–inferential connection between their antecedent and consequent (typically, the event cited in the consequent is, or purports to be, the best explanation of the event cited in the antecedent). Two experiments try to predict truth ratings for such conditionals on the basis of judgments of explanatory goodness. Inferentialism predicts about our materials that participants will tend to agree more with a conditional, the better the consequent explains the antecedent and so the stronger the inferential connection between antecedent and consequent is. The first two experiments allow us to contrast inferentialism with a version of the mental models account that aims to explain truth ratings in terms of salient alternatives and disablers. A third experiment looks at abductive conditionals in the context of modus ponens arguments. Inferentialism predicts that endorsement rates for such arguments co-depend on the strength of the inferential connection between the component parts of the major premise and so, again given our materials, on how well that premise's consequent explains its antecedent. The experiment aims to determine whether there is any support for this prediction, and it also contrasts inferentialism with the suppositional account of conditionals as well as with accounts that postulate a more complex probabilistic connection between a conditional's antecedent and consequent. To preview our results, we find strong support for inferentialism across the three experiments.

Keywords: abductive conditionals; explanation; inference; inferentialism; modus ponens; semantics.

*All Supplementary Materials for this paper can be downloaded from https://osf.io/8nh2x/?view_only=2dcfc098369e428388120880664 68e55.

1 Introduction

Conditionals without an apparent connection between their antecedent and consequent tend to raise interpretational difficulties. Consider these examples:

(1) a. If Trump wins the 2020 presidential election, then goldfish need food to survive.

b. If sea levels keep rising, then Brazil will win the 2022 FIFA World Cup.

How in the world could the result of the 2020 United States presidential election be related to the dietary requirements of goldfish? Similarly for (1b): how could there be a connection between sea levels and Brazil's winning chances for the 2022 World Cup? And thus, more generally, what could ever occasion the assertion of either conditional?

To understand the strangeness of such "missing link conditionals" (Douven, 2017a), we must go beyond the simple statement that there is no apparent connection between their constituent parts, or that their consequent does not depend in any plausible sense on the truth of their antecedent, or that there is no intuitive sense of conditionality at play in this type of sentence. Attempts to more fundamentally explain the said phenomenon can go in two different directions: one could argue that what is wrong with missing link conditionals is of a pragmatic nature, in that they violate certain principles of good language usage; or one could hold that such conditionals are odd because they blatantly fail to be true, making what is amiss with them a semantic issue.

Of late, there has been increasing attention to the second approach, in particular for a view called "inferentialism," according to which the truth of a conditional requires the presence of a sufficiently strong inferential connection between its antecedent and its consequent. A number of recent studies report experimental results supporting inferentialism (e.g., Vidal & Baratgin, 2017; Douven, Elqayam, Singmann, & van Wijnbergen-Huitink, 2018, 2019; Skovgaard-Olsen, Kellen, Hahn, & Klauer, 2019). So far, the most explicit comparison of inferentialism with other semantics of conditionals is to be found in Douven et al. (2019), which re-analyzes data gathered in the context of testing Douven et al.'s (2018) processing account of conditionals. The re-analysis showed those data to be much better explained by inferentialism than by the rival semantics of conditionals.

However, the data from Douven et al.'s (2018) study concerned fairly abstract materials, and this could raise legitimate concerns about the generality or ecological validity of those same authors' (2019) findings. In particular, one might wonder whether studies with realistic materials, which evidently would constitute a more important test of inferentialism, would corroborate or undermine the support that Douven and coauthors found for that position. The present paper addresses this concern by presenting three new experiments, all using the same set of unambiguously realistic conditionals. To be more exact, the materials were modifications of the materials from earlier research concerning causal reasoning (Cummins, Lubart, Alksnis, & Rist, 1991; Cummins, 1995; De Neys, Schaeken, & D'Ydewalle, 2003). In that research, the materials consisted of conditionals that were all of the schematic form "If ⟨*cause*⟩ then ⟨*effect*⟩." In our research, the conditionals are of the form "If ⟨*effect*⟩, then ⟨*explanation*⟩," where the type of explanation is always a causal explanation. In other words, we are using *abductive conditionals*, in the sense of Douven and Verbrugge (2010).

The first two experiments use these materials to elicit both truth ratings of the conditionals and judgments of how well the consequent explains the antecedent, the first experiment using a between-subjects design, the second a within-subjects design. The judgments of explanation quality can be interpreted as measuring the perceived strength of the inferential connection between antecedent and consequent (Glass, 2007, 2012; Douven, 2013; Douven & Wenmackers, 2017; Douven

& Mirabile, 2018). Thus, supposing inferentialism to be correct, those judgments should accurately predict the truth ratings.

In the third experiment, the conditionals from our materials serve as major premises in modus ponens arguments. In an inferentialist account, conditionals can be thought of as "inference tickets" (Ryle, 1950). And just as a more expensive train ticket will (typically, at least) allow a passenger to travel farther, the strength of the inferential connection between a conditional's constituent parts will determine how far the conditional will allow a reasoner to go. In a modus ponens argument, this means that how confident participants will be in endorsing the conclusion will depend not only on how confident they feel about the minor premise, but also on their judgment of the strength of the inferential connection between the major premise's antecedent and consequent—which, again, for our materials, depends on explanation quality. Before turning to our experiments and results, we lay out the inferentialist account of conditionals and the theoretical background of this research more generally.

2 Theoretical background

Whereas the study of logic as initiated by the Greek philosophers was not especially concerned with mathematical reasoning, formalizing mathematical reasoning was the *only* concern of Frege, Russell, and others when they devised *modern* logic. From that later perspective, logical operators like "and" (conjunction), "or" (disjunction), "not" (negation), and "if" (implication) were not required to match what would appear their obvious natural-language counterparts. By contrast, such a match was a vital point of interest for the philosophers and linguists who, impressed by the successes of logic in the first half of the twentieth century, sought to extend the use of logic to natural language, or at least important parts of it.

The finding that there were in fact considerable mismatches between the logical operators and their natural-language counterparts inspired Grice (1989) and others to develop the field now known as "pragmatics," which proposed to bridge the gap between mathematical and natural language by reference to principles of usage (referred to as conversational "maxims"), all flowing from the general premise that communication relies on an assumption of cooperativeness between participants: we expect each other to be *helpful*, which creates certain further expectations about what we contribute to a conversation. When conjoined with knowledge of the meaning of the logical operators, these expectations can be used to calculate what a speaker means to convey. For instance, logically speaking, the statement "Jim has tea or coffee" is true when Jim has either beverage and when he has both beverages, but we take it to mean that he has *either* tea *or* coffee (but not both) because we reckon that if the speaker had meant that Jim had both, she could have easily been more informative and thus more helpful, viz., by simply saying that Jim has tea *and* coffee.

Although the development of pragmatics has contributed much to our understanding of natural language, there remain linguistic phenomena which are not well explained even by the combination of logic and pragmatics. A prime example is our use of the natural-language conditional-forming operator "if" (or "if . . . then . . . "). Grice and others thought that this operator could be modeled by means of the material conditional (the horseshoe from propositional and predicate logic), where pragmatics then had to take care of apparent counterexamples. But this approach is known to face some seemingly insurmountable problems (Evans & Over, 2004; Douven, 2008, 2016, 2017a). Here, we want to highlight a problem related to missing link conditionals.

A reasonable diagnosis of why we find (1a) and (1b) perplexing is that we are unable to see *any* connection between a possible second Trump victory and what goldfish need to survive, or between rising sea levels and a 2022 win for the Brazilian football team. However, our expectation of finding these connections is not accounted for by anything in the logic of the material conditional, and as far as anyone has shown bringing in Gricean pragmatics is not helpful either. One might initially think otherwise. For among the Gricean maxims, there is the maxim of relation, which requires a contribution to a conversation to be *relevant*, and—some might say—what makes missing link conditionals appear odd is precisely that their antecedent bears no relevance to their consequent. It is to be noted, though, that the lack of a relevance relation between their constituent parts is not itself sufficient to explain why missing link conditionals do not contribute relevantly to an ongoing conversation—and the latter is all the Gricean maxim cares about (Douven, 2016, Ch. 4). To put this differently, *that* one will typically make an irrelevant contribution to a conversation by asserting a conditional whose constituent parts are not related in any comprehensible sense is not something that follows from the maxim of relation.[1]

Recent attempts to explain semantically the oddness of missing link conditionals build on a tradition that long pre-dates the emergence of pragmatics. The Stoic philosophers, most notably Chrysippus, held that for a conditional to be true, the consequent must be deducible from the antecedent (Kneale & Kneale, 1962, Ch. 3). In a similar vein, John Stuart Mill wrote in his *System of Logic* (1843) that what we assert when we assert a conditional is the "inferribility" of the consequent from the antecedent. Still later, a related view was defended by Ryle, who argued in his (1950, pp. 308–310) that the conditional "If φ then ψ" is an *inference ticket* which allows us to "travel" from φ to ψ (which we may never actually do), and that asserting "If φ then ψ" is like asserting "φ, so ψ" without committing oneself to the truth of either φ or ψ. Similar "inferentialist" ideas were most recently defended by Spohn (2013), Krzyżanowska, Wenmackers, and Douven (2013), Krzyżanowska, Wenmackers, and Douven (2014), Krzyżanowska (2015), Douven (2016), Skovgaard-Olsen (2016), and van Rooij and Schulz (2019).[2]

In all these accounts, the idea of an inferential connection between a conditional's antecedent and its consequent is baked into the truth conditions of the conditional. Still, the exact nature of that connection differs among accounts. For the Stoics, the inferential connection had to be deductive: the consequent had to follow by necessity from the antecedent. Mill and Ryle may have had a more

[1] See for more on this, Douven (2008). That Gricean pragmatics is unable to explain the oddness of missing link conditionals is also indicated by experimental evidence; see Krzyżanowska, Collins, and Hahn (2017), where it is argued that the pragmatic requirement of discourse coherence is not sufficient to explain why missing link conditionals are odd. These authors found that, in order to be judged assertable by their participants, conditionals needed, in addition to meeting the requirement of discourse coherence that is more generally expected from consecutive elements of discourse, also to have a relationship of probabilistic relevance between antecedent and consequent.

[2] Versions of inferentialism, albeit not under that name, are also to be found in Barwise and Perry (1983) and Oaksford and Chater (2010). In particular, that Oaksford and Chater relate conditionals to "methodological policies" (p. 110), and describe them as "structure building operators" (p. 114), and above all their claim that "the assertion of a conditional directly proposes the existence of a causal or related dependency between the antecedent and the consequent" (p. 116), makes their view *very* close to our own favored version of inferentialism, possibly even identical with it, depending on how Oaksford and Chater would explicate the reference to "related dependencies." In later work (e.g., Oaksford & Chater, 2013, 2014, 2017), these authors mainly focus on causal connections, which are important from an inferentialist viewpoint in that such connections often underlie abductive and inductive *inferential* connections, but which cannot be the whole story from that viewpoint, given that, for instance, not all abductive inferential connections rest on causal explanations (e.g., some explanations are functional or teleological rather than causal). In Oaksford and Chater (2020a), the authors explicitly commit to inferentialism, even though in that paper, too, the emphasis is on causal connections between antecedent and consequent.

informal notion of inference in mind. Krzyżanowska and colleagues are explicit that the inferential connection need *not* be deductive, but may also be inductive (based on frequency information) or abductive (based on explanatory considerations). Indeed, in their view the connection ought to be ensured by a compelling *argument* from antecedent to consequent, where this argument may involve deductive, inductive, and abductive steps simultaneously.

It is worth emphasizing that, by these accounts, the oddness of missing link conditionals is not explained simply by those conditionals' perceived lack of truth. Nor is it explained by the fact that we might be unable to *reconstruct* the argument connecting their constituent parts, since we have no issue accepting some conditionals as true despite not having effectively identified the connecting argument. For instance, if we trust the speaker, then an assertion of "If φ then ψ" will normally make us believe that there is a compelling argument from φ plus background premises to ψ even if the speaker does not provide that argument and we fail to see it ourselves. In the case of missing link conditionals, by contrast, the problem is not that we are unable to identify the connecting argument but rather that it is glaringly obvious that there is no such argument. With some effort, we might be able to come up with an argument connecting a 2020 Trump victory with what goldfish need to survive. But we know in advance that the argument would be phony and anything but compelling. *That* is why missing link conditionals perplex us, or at least it is one reason why they do so.

In the following, we concentrate on Krzyżanowska and coauthors' recent implementation of the idea that the truth of a conditional requires an inferential connection between its antecedent and consequent. To be more precise, in this account the truth of a conditional requires two conditions: first, there must be a compelling argument from the conditional's antecedent to its consequent—which may contain deductive, inductive, and abductive steps—and second, the antecedent must be essential to making the argument compelling, meaning that the argument for the consequent cannot rely on background premises alone.

That, on this account, the argument only needs to be compelling, and not necessarily conclusive, is of particular interest to us and will become relevant in the description of the third experiment. Because of this feature, the account is consistent with there being true conditionals with a true antecedent and a false consequent: inductive and abductive inferences are what is called "ampliative," meaning that the truth of their conclusions is not guaranteed by the truth of their premises. As a consequence, the argument form of modus ponens (MP), which licenses the inference of ψ from the premise set $\{\varphi, \text{If } \varphi \text{ then } \psi\}$, is logically invalid on this account, as logical validity requires that the conclusion of an argument be necessarily true any time its premises are true. For suppose that (i) φ is true; (ii) there is a compelling yet inconclusive argument from φ to ψ; yet (iii), as it happens, ψ is false. Then together φ and "If φ then ψ" give rise to an instance of MP with true premises and a false conclusion.

Some might see this as a problem for inferentialism, at least in the version of Krzyżanowska et al., given how natural MP appears to us, and also given that conclusions from MP arguments tend to be highly endorsed in experimental settings (for an overview, see Evans & Over, 2004, Ch. 3). However, for reasons given in McGee (1985), that criticism would be unwarranted. Specifically, McGee gave independent reasons (having to do with the nesting of conditionals) for holding that MP is invalid. As for the concern that MP appears *intuitively* valid, McGee notes that our intuition would be unable to discriminate between applications of MP leading from true premises to a true conclusion *always*, or only *almost always*. And that is a response the inferentialist can fully appropriate, given that on this account, MP will *reliably* lead from true premises to a true conclusion, as reliably as we

Figure 1: The soritical color series from the materials of Douven et al. (2018).

take the combined use of deduction, induction, and abduction to be. And that is pretty reliable—which is why in our daily lives we routinely rely on all these modes of inference.

In the meantime, evidence has accumulated for the specific version of inferentialism proposed by Krzyżanowska and colleagues. This version of inferentialism served as one of the pillars of Douven et al.'s (2018) Hypothetical Inferential Theory (HIT), a psychological account of the interpretation of conditionals; the other pillar is Evans' (2006, 2007) Hypothetical Thinking Theory, a dual-process account of reasoning and decision making. According to HIT, our default interpretation of conditionals is one in which we represent their constituent parts as inferentially related to one another, where this connection need only be "strong enough," in the sense of Krzyżanowska et al. (2014).

Douven and coauthors' main experiment concerned the soritical color series shown in Figure 1, with colored patches gradually shifting from clearly green to clearly blue, through various shades of blue and green, including borderline blue–green shades. Participants were asked to evaluate a number of conditionals about this series, all being of the schematic form

If patch number i is X, then patch number j is X,

where $i \in \{2, 7, 8, 9, 10, 13\}$ for all participants, and with X standing for either "blue" or "green," depending on whether the participant was in the blue or in the green condition (a split that was made strictly for control purposes). Finally, j depended on whether the participant was in the small or large condition: in the former case, the patch referred to in the consequent was either one or two steps away from the patch referred to in the antecedent; in the latter case, the distance between the patches was either one or three steps.[3]

It is to be noted that, with each of the resulting conditionals, one can naturally associate an argument. For instance, the argument backing

(2) If patch number 6 is green, then so is patch number 7,

would go something like this: Patches become greener as we move to the right in the color series; on the supposition that patch number 6 is green, and given that patch number 7 is to the right of patch number 6, patch number 7 must be green. And with

(3) If patch number 6 is green, then so is patch number 5,

we can associate an argument to the effect that since adjacent patches are very similar in color, and since patch number 5 is adjacent to patch number 6, patch number 5 must be green on the supposition that patch number 6 is green.

Importantly, these arguments are not all on a par in terms of strength. For example, (2) and (3) both refer to adjacent pairs of patches, but in the former the consequent patch is to the "greener" side of the antecedent patch while in the latter the consequent patch is to the "bluer" side of the

[3] For instance, a participant in the green and small condition would, for the antecedent patch 7, see the instances of "If patch number 7 is green, then patch number j is green" with $j \in \{5, 6, 8, 9\}$, while a participant in the green and large condition would for the same antecedent patch see the instances with $j \in \{4, 6, 8, 10\}$.

antecedent patch. The argument associated with (3) is still a good one, but it is not as good as the one associated with (2): in the former case, there is a consideration at least somewhat going against the conclusion, in the latter, there is not. As Douven and coauthors argued, the important determinants for argument strength in the context of their materials were direction—is the consequent patch to the left or to the right of the antecedent patch?—and distance: how close is the consequent patch to the antecedent patch? The importance of direction is already illustrated by the comparison of (2) and (3) above. To see the importance of distance, compare, for instance, (3) with

(4) If patch number 6 is green, then so is patch number 4.

We can associate an argument with (4) that is *roughly* the same as the one we associated with (3), but because (i) in both sentences the consequent patches are to the "bluer" side of the antecedent patch, and (ii) patches that are two steps away from each other are not quite as similar as patches that are only one step away, the argument associated with (4) is a bit weaker. In their analysis, Douven et al. found that these factors indeed predicted with great accuracy the rates at which their participants had judged the conditionals to be true.

Douven et al. (2018) were primarily interested in HIT, a psychological theory, and so they did not look at the implications of their findings for the semantics of conditionals. They made good on that in their (2019), in which they compared inferentialism with the main other extant semantics of conditionals. Re-analyzing the data from their earlier paper, they found in the later paper that inferentialism did a much better job explaining those data than did any of its rivals.

Although good news for inferentialism, it is to be noted that the data from Douven et al. (2018) concerned a somewhat artificial setting. Even if the materials are not abstract, they are not exactly realistic either. Hence, the question arises how inferentialism holds up when tested using realistic materials. Obviously, we are much more interested in how accounts of conditionals fare when confronted with such realistic materials, and much less in how they fare in artificial settings. Therefore, the current paper relies on realistic materials only.

3 Plan

Douven and Verbrugge (2010) proposed a typology of conditionals based on the type of inferential connection between antecedent and consequent. They distinguished three types of conditionals: deductive, inductive, and abductive. In a *deductive* conditional, the consequent follows deductively from the antecedent plus background knowledge, as in "If France has a king, then France is a monarchy." In *inductive* conditionals, the consequent follows inductively from the antecedent plus background knowledge, so meaning that the inference is based on information about frequencies specified either verbally ("most," "almost all," "virtually always," and so on) or numerically (e.g., "over ninety percent"). An example would be, "If Henry is in class 6A, then he has the flu," supposing it is part of the background knowledge that almost everyone in class 6A has the flu, or that, say, ninety-five percent of the students in class 6A have the flu. Finally, in an *abductive* conditional, the consequent follows abductively from the antecedent and background knowledge, meaning that the inference is based on explanatory considerations, more exactly, on considerations of explanatory superiority. Consider, for instance, "If Judy and Pam are jogging together, then they have patched up their friendship," where it is known that Judy and Pam recently had a flaming row. That they are jogging together is best explained by assuming that they patched up their friendship, after the row they had. Note that there could be other explanations: maybe they had to have a discussion for

professional reasons, and they thought it would be best to have that discussion while jogging, so that they did not have to look each other in the face. What warrants the inference from Judy and Pam's jogging together to their having patched up their friendship is that the latter is the *best* explanation for the former (or so we may assume, for the sake of the example).

We decided to use abductive conditionals for our materials, for a number of reasons. First, given the connection between abductive conditionals and causal conditionals—basically a reversal of antecedent and consequent, as mentioned previously—we could use uncontroversially realistic materials from studies on causal reasoning to serve our current purposes. Second, we had previous experience manipulating explanation quality, and detailed knowledge of how explanation quality relates to truth judgments and probabilities, and also of the effect of alternative explanations on truth and probability ratings (Douven & Mirabile, 2018). Third, we wanted to compare inferentialism with an account that posits that the interpretation of conditionals is primarily influenced by a counterexample search process in semantic memory (Markovits, Fleury, Quinn, & Venet, 1998; Janveau-Brennan & Markovits, 1999; Markovits, 2000; Markovits & Potvin, 2001), such as in the mental models account of conditionals proposed in Johnson-Laird (2006, Ch. 21).[4] Indeed, an alternative account of how judgments of explanation quality and truth judgments relate, viz., in terms of certain mental models being prompted by the consideration of alternative explanations and disabling conditions (see below) appears to be a natural prolongation of how mental models theory has been suggested to apply to causal conditionals (Cummins et al., 1991; De Neys et al., 2003). And fourth, we also wanted to compare inferentialism with the suppositional account of conditionals, and such a comparison requires that one can prise apart judgments of inferential strength and probability judgments. That may be impossible in the case of deductive conditionals (given that probability respects logic) and inductive conditionals (where inferential strength and probability may both be determined straightforwardly by frequency information). However, on the basis of our earlier work on the connection between explanation, acceptability, and probability, we had good reason to believe that the comparison *can* be made for abductive conditionals. As part of the experiments reported in (Douven & Mirabile, 2018), participants were asked to evaluate the quality, probability, and acceptability of two competing possible explanations for an event, and we found that ratings of acceptability for an explanation could be more successfully predicted by explanation-quality ratings than by probability ratings.

We use the same set of abductive conditionals in three experiments, with an eye toward testing inferentialism along two different lines. Both capitalize on a quantitative version of the so-called Inference to the Best Explanation—as defended in, for instance, Glass (2007, 2012), Douven (2013, 2017b, 2019, 2020), Douven and Wenmackers (2017), and Trpin and Pellert (2019)—which says that the strength of an abductive inference is determined by how well the explanans (the part that provides an explanation) explains the explanandum (the fact that requires an explanation): the better a hypothesis explains the available data, the stronger our license to infer that hypothesis. We rely on this idea when we measure the explanation quality of the consequents of our conditionals in light of the corresponding antecedents and then use those data to test predictions in line with inferentialism. More specifically, we derive two hypotheses from inferentialism: first, a hypothesis about how judgments of the explanation quality of ψ in light of φ predict the truth ratings for "If φ then ψ"; and second, a hypothesis about how those same explanation-quality judgments predict the rates at which ψ is endorsed when participants are asked to suppose the conditional "If φ then ψ" while also being given the information that φ holds with some degree of certainty.

[4] See Baratgin et al. (2015) for a general critique of this account.

Experiments 1 and 2 target the first hypothesis. They also compare inferentialism with an alternative account of the semantics of conditionals which suggests that the truth ratings we obtain should be predicted by the results of a search in semantic memory for counterexamples, and more precisely by (i) the number of alternative explanations for the antecedent (i.e., the possible explanations other than the event mentioned in the consequent), and (ii) the number of so-called disablers: possible events that might block the antecedent event even in the presence of the consequent event. Experiment 3 addresses the second hypothesis. It also compares inferentialism with a number of accounts that aim to capture the connection between a conditional's constituent parts probabilistically. The best known of these is the suppositional account, which revolves around the so-called Equation, according to which $\Pr(\text{If } \varphi \text{ then } \psi) = \Pr(\psi \mid \varphi)$, or in words: The probability of a conditional equals the corresponding conditional probability. There is evidence that people's probability judgments do obey the Equation, at least by and large (Hadjichristidis et al., 2001; Oaksford & Chater, 2003, 2007; Over & Evans, 2003; Evans & Over, 2004; Oberauer, Weidenfeld, & Fischer, 2007; Gauffroy & Barrouillet, 2009; Douven & Verbrugge, 2010, 2013; Pfeifer & Kleiter, 2010; Politzer, Over, & Baratgin, 2010; Fugard, Pfeifer, Mayerhofer, & Kleiter, 2011; Over, Douven, & Verbrugge, 2013), although recent work on missing link conditionals has shown the Equation to break down precisely for such conditionals (Skovgaard-Olsen, Singmann, & Klauer, 2016). Nevertheless, we shall see that the suppositional account and kindred probabilistic accounts appear well-poised to explain the variability in endorsement rates of MP arguments that we intend to explain by appeal to inferentialism.

4 Experiment 1

We were interested in testing the following hypothesis:

H1: The strength of the inferential connection between a conditional's antecedent and consequent predicts the rate at which that conditional will be judged true.

Given our choice of materials (consisting only of abductive conditionals), the strength of the inferential connection between antecedent and consequent amounts to the measure to which the consequent explains the antecedent. Thus, for our materials, H1 can be sharpened as follows:

H1*: The explanation quality of the consequent, given the antecedent as explanandum, predicts the rate at which the conditional will be judged true.

While H1 and H1* follow from inferentialism, in the way described previously, the phenomena they explain can also be interpreted in light of mental models theory. Experimental results reported in Cummins et al. (1991), Cummins (1995), and De Neys et al. (2003) suggest an account in terms of retrieval in semantic memory of counter-examples, that is, alternative explanations and potential disablers. Alternative explanations diminish the necessity for the consequent event to have resulted from the antecedent event, and potential disablers are events that undermine the effectiveness of the explanans (the consequent event, in our materials), in that they hinder its ability to produce the explanandum (the antecedent event). To illustrate, a wet sidewalk can be explained by a spell of rain earlier in the day, but also by the flooding of a nearby river; and the former explanation could be "disabled" by an overhanging eave covering the sidewalk.

Concretely, these earlier findings suggest the following hypothesis:

H2: An increased number of available alternative explanations for the antecedent in addition to the explanation provided by the consequent, and/or an increased number of events that could prevent the consequent from successfully producing the antecedent from happening, jointly predict a decrease in the rate at which the conditional will be judged true.

In a mental models framework, this hypothesis would be derived from the assumption that an increased number of available counter-examples (alternative explanations and potential disablers) for an abductive conditional would increase the probability of a subject's retrieving at least one of these counter-examples when considering the truth of that conditional. Each of these retrieved counter-examples would in turn trigger the generation of an alternative mental model in which the conditional fails to hold true. Finally, the number of generated alternative mental models—building upon the results established by De Neys et al. (2003) in the case of causal conditionals—would predict a progressive decrease in the truth ratings of the conditional. Our aim was to investigate whether truth ratings of abductive conditionals can be predicted on the basis of judgments of explanation quality, over and above the mean number of generated counter-examples (alternatives and disablers) for these conditionals.

Following the procedure used by the aforementioned authors, we first ran a pilot study in which participants generated counter-examples for a pool of twenty conditionals. We used these data to compute the mean number of generated alternative explanations as well as the mean number of potential disablers for each conditional. This allowed us to select the materials for the three experiments reported in this paper. We then proceeded with Experiment 1, which had a between-subjects design with two conditions, with participants in the first condition rating the truth of abductive conditionals and participants in the second condition rating how well the consequent of those conditionals explained their antecedent.

4.1 Methods

4.1.1 Materials

The materials were selected on the basis of the outcomes of a pilot study. Participants in this pilot were 81 adults recruited by the INSEAD–Sorbonne University Behavioral Lab (39 females; M_{age} = 23.1, SD_{age} = 2.95) after approval by the INSEAD's Ethical Committee. They completed the study via an on-line questionnaire after giving informed consent and were compensated through a lottery system, with two randomly selected participants receiving € 20 each. An additional 7 participants completed the study but were excluded for failing an attention check.

The materials were adapted from a pool of twenty causal conditional statements initially developed by Cummins et al. (1991). These conditionals pair an everyday event (the consequent) with a possible cause for that event (the antecedent) and were rated as presenting an at least "moderately strong causal relationship" (Cummins et al., 1991, p. 277); some examples are, "If John studied hard, then he did well on the test," and, "If Joe cut his finger, then it bled." In the experiments reported here, these statements were translated into French and transformed into twenty abductive conditionals by flipping their antecedents and consequents: they were introduced as offering a possible explanation (described in the consequent of the conditional) for a given event (described in the antecedent of the conditional), for instance, "If John did well on the test, then he studied hard," and, "If John's finger is bleeding, then he cut his finger."

We constructed twenty vignettes for this pilot study. Each vignette consisted of a statement describing an event as well as a possible explanation for that event (corresponding, respectively, to

the antecedent and to the consequent of one of our twenty abductive conditionals). To give a full example of one of the vignettes used:

Fact: John did well on the test.

Possible explanation: John studied hard.

corresponded to the conditional statement, "If John did well on the test, then he studied hard."

Half of the participants were asked to generate alternative explanations for each vignette, and the other half were asked to generate possible disabling conditions, that is, obstacles that might prevent the explanation from producing the event. In the alternative-explanations generation task, participants were asked: "Can you find other possible explanations for this fact?"; and in the disablers generation task, they were asked: "Can you find examples of events that could have prevented the explanation from producing this fact?" Participants were also instructed that they needed to generate counter-examples that were "reasonably different" from each other.

Participants first read examples of alternative explanations and disabling conditions for a few everyday situations and completed a practice trial. At the end of the practice trial, examples of acceptable and unacceptable responses were discussed in order to clarify what counted as reasonably different responses. Participants then performed the generation task for each of the twenty vignettes, which were presented in an order randomized per participant. They were allowed ninety seconds per item to generate counterexamples, after which they were automatically taken to the next item. Finally, participants responded to a few demographic questions and were asked to indicate whether they had responded seriously to the survey (following a recommendation by Aust, Diedenhofen, Ullrich, & Musch, 2013).

In the next stage, the generated responses were scored by two independent coders, according to scoring criteria determined in advance. These criteria allowed the elimination of "outlandish" responses (e.g., "John did well on the test because a magician told him the answers in advance") and of responses that were not different enough (e.g., "John was very focused while studying" would count as being too similar to "John studied hard"). Scores provided by an additional third coder were used in case the first two coders disagreed; else just the first two coders' scores were used.

Using these data, we computed the mean number of alternative causes and mean number of disabling conditions for each of the twenty conditionals, which allowed us to rank conditionals by mean number of generated alternatives and by mean number of disabling conditions. According to these rankings, we selected sixteen conditionals, with four conditionals for each cell of a 2×2 matrix of alternative explanations (many, few) and disabling conditions (many, few). These sixteen conditionals constituted the materials used in the three experiments reported in this paper.

4.1.2 Participants

Participants in the main experiment were 81 adults recruited by the INSEAD–Sorbonne University Behavioural Lab (58 females; $M_{age} = 22.1$, $SD_{age} = 3.2$) after approval by the INSEAD's Ethical Committee. They completed the study via an on-line questionnaire after giving informed consent and were compensated through a lottery system, with two randomly selected participants receiving € 20 each. An additional 5 participants completed the study but were excluded for failing an attention check.

4.1.3 Procedure

The study used a between-subjects design, where participants were randomly assigned either to the explanation-quality evaluation condition or to the truth rating condition.

In the explanation-quality evaluation condition, participants ($N = 33$) were first given a comprehension task where they had to select a bad explanation amongst a list of possible explanations for an event. They then completed the main part of the survey, in which they received, in an order randomized per participant, all sixteen of the abductive conditionals selected in the pilot: they were asked to suppose that an event had been observed and were offered a possible explanation for that event; for instance, "Suppose we observe that John did well on his exam. We propose to explain this by the fact that he studied hard." The event and the explanation corresponded respectively to the antecedent and the consequent of one of the abductive conditionals that constituted our main materials. The participants then evaluated the quality of the explanation, responding to the question "How would you rate the quality of this explanation?" on an 11-point Likert scale, with the extreme points labeled "Very bad" and "Very good" and the middle point labeled "Neither good nor bad."

In the truth rating condition, participants ($N = 48$) were first shown an example of the task. They were then presented with all sixteen conditionals in a randomized order and asked to rate the truth of those conditionals. For instance, they read, "If John did well on his exam, then he studied hard," and were asked to respond to the question "How strongly do you agree that this statement is true?" on an 11-point Likert scale, with the extreme points labeled "Strongly disagree" and "Strongly agree" and the middle point labeled "Neither agree nor disagree." An attention check that had participants count the number of objects in a picture was included half-way through the survey. After completing the main part of the survey, participants in both conditions responded to a few demographic questions and were asked to indicate whether they had responded seriously to the survey, which constituted a second attention check.

4.2 Results and discussion

In the pilot, participants generated an average of 2.38 (SD = 1.25) alternatives and an average of 0.93 (SD = 0.62) disablers. In the main experiment, the average truth rating for the conditionals was 7.21 (SD = 3.58) and the average explanation-quality rating was 9.04 (SE = 2.44). Figure 2 plots the truth responses versus the explanation-quality responses and versus the number of alternatives and number of disablers.

For each of the 16 conditionals in our materials, we calculated mean truth ratings and mean explanation-quality ratings as based on the responses from the main experiment, and we calculated the mean number of alternatives and mean number of disablers as based on the responses from the pilot. Here and in the other experiments, analyses were performed using both Bayesian and frequentist methods. In the paper, we report the results from the Bayesian analyses. Interested readers may consult the Supplementary Materials for the frequentist analyses. All Bayesian analyses were conducted using the `R` package `brms` (Bürkner, 2017).

For this experiment, we fitted three Bayesian regression models, using the weakly informative priors as provided by default by the `brms` package.[5] All models had mean truth responses as dependent variable, one with explanation quality (EQ) as predictor variable, a second with both mean number of alternatives (A) and mean number of disablers (D) as predictor variables, and finally a

[5] An improper flat prior was used for all predictors, a t distribution ($\mu = 3$, $\sigma = 8$, df = 10) for the intercept, and another t distribution ($\mu = 3$, $\sigma = 8$, df = 10) for the standard deviation.

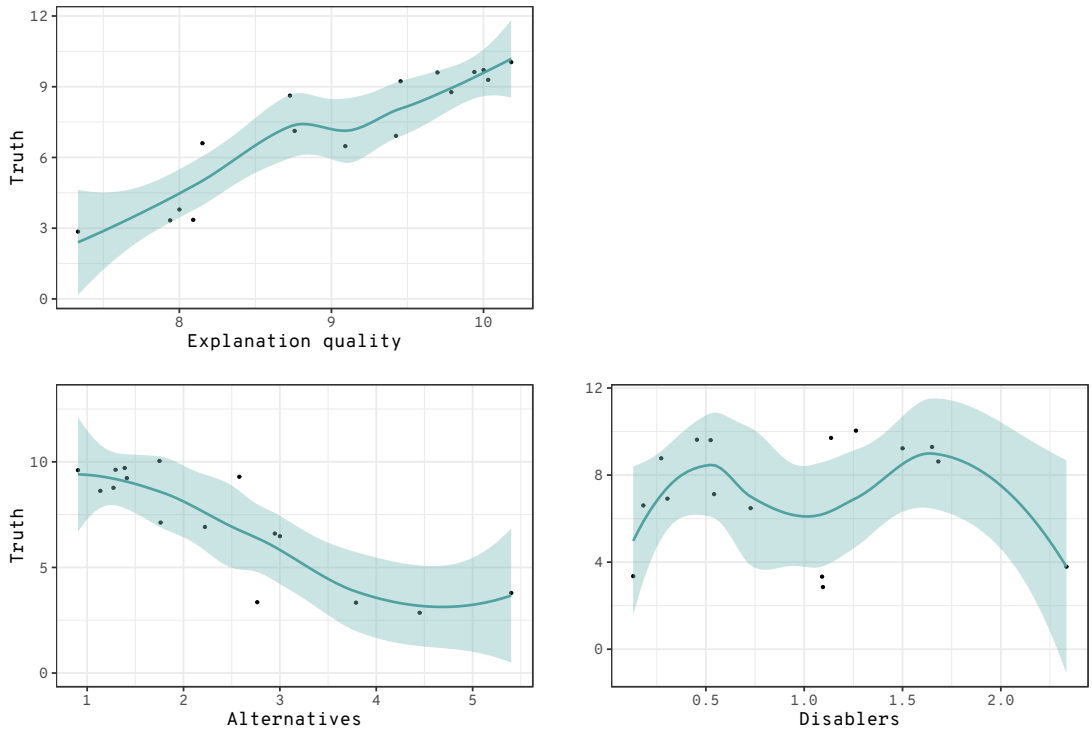

Figure 2: Mean truth ratings versus mean explanation-quality ratings (top left), number of generated alternatives (bottom left), and of generated disablers (bottom right) for each of the sixteen abductive conditionals, with smoothers added to highlight trends (shaded areas indicate 95 percent confidence intervals).

"full" model, with EQ, A, and D as predictor variables. To facilitate interpretation of the regression results and to make the coefficients of the predictors better comparable, we followed a recommendation by Gelman (2008) and standardized those predictors by centering them at their means and then dividing by twice their standard deviation. Diagnostic tests (collinearity diagnostics, posterior predictive checks, $\widehat{R}$ statistics, and caterpillar plots) raised no red flags.

Currently, the leave-one-out cross-validation information criterion (LOOIC) is recommended for comparing Bayesian models. Using this criterion, the full model came out best, as is seen in

Table 1: Comparison of Bayesian cumulative ordinal regression models.

predictor(s)	LOOIC	(SE)	ΔLOOIC	(SE)
EQ, A, D	46.2	(10.7)	0.0	—
A, D	58.2	(5.0)	12.0	(9.8)
EQ	51.6	(5.1)	5.4	(9.2)

Note: EQ = explanation quality; A = number of alternatives; D = number of disablers. The LOOIC is used to estimate the expected out-of-sample predictive accuracy for a fitted Bayesian model (Vehtari, Gelman, & Gabry, 2017); smaller values indicate better expected out-of-sample accuracy. ΔLOOIC is the difference in expected predictive accuracy, according to LOOIC, between each model and the model with the smallest LOOIC value.

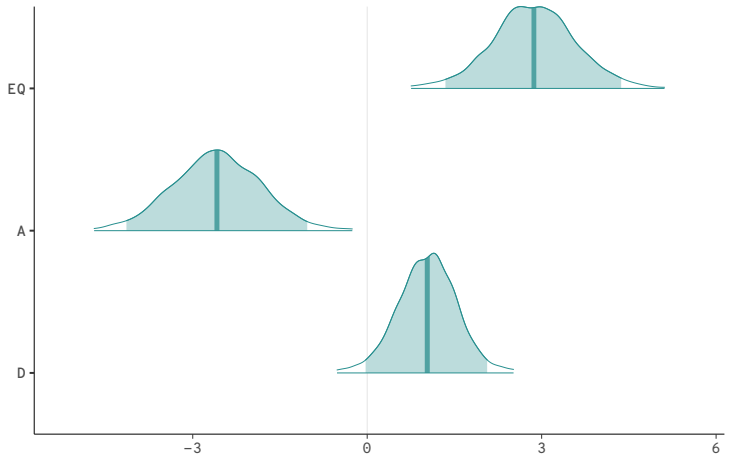

Figure 3: Posterior distributions of explanation quality (EQ), number of alternatives (A), and number of disablers (D) in the full model for Experiment 1.

Table 1. In this model, both EQ and A were consistently associated with the dependent variable: the coefficient for EQ equalled 2.87 (posterior SD = 0.76), with a 95 percent credible interval from 1.35 to 4.37, and that for A equalled −2.58 (posterior SD = 0.80), with a 95 percent credible interval from −4.14 to −1.03. The remaining predictor D had a coefficient of 1.03, with a 95 percent credible from −0.03 to 2.06. That this interval includes 0 indicates that there was no consistent association between that predictor and mean truth responses. Figure 3 plots the posterior distributions for the three predictors, as estimated in the full model.

The outcomes provide evidence in favor of H1, and hence in favor of inferentialism: truth ratings of abductive conditionals increase, ceteris paribus, with an increase of the rating of how well the conditional's consequent explains its antecedent. At the same time, there is partial support for H2, given that number of alternatives also has an effect on truth ratings, even though number of disablers does not.

However, the approach taken so far is rather crude, considering only aggregate responses. Nothing said in the above precludes that, at an *individual* level, numbers of alternatives and disablers both impact truth responses, perhaps even to a greater extent than explanation-quality judgments do. We investigated this possibility in Experiment 2.

Before turning to this experiment, we note that there is no inconsistency between our results and the results from Cummins, De Neys, and their respective co-authors, given that they were concerned with causal conditionals, not abductive conditionals. Their results can naturally lead to the supposition that number of alternatives and number of disablers are also predictive of truth ratings of abductive conditionals, and we saw that at least number of alternatives seems to be. Conversely, our results suggest that it could be worthwhile to investigate the extent to which truth ratings of conditionals of the form "If ⟨*cause*⟩ then ⟨*effect*⟩" can be predicted on the basis of judgments of how well ⟨*cause*⟩ explains ⟨*effect*⟩.

5 Experiment 2

In Experiment 2, our goal was to replicate the results from Experiment 1 at the individual level: we used a within-subjects design to investigate whether the number of counter-examples (alternative explanations and disabling conditions) generated by participants for a conditional would be a better predictor of their judgments of the truth of that conditional than their judgments of how well the conditional's consequent explained its antecedent. In other words, we wanted to subject H1 and H2 to a further, more severe test.

5.1 Method

5.1.1 Participants

Participants in Experiment 2 were 27 adults recruited by the INSEAD–Sorbonne University Behavioural Lab (19 females; M_{age} = 21.6, SD_{age} = 3.6), after approval by the INSEAD's Ethical Committee. They each gave informed consent and received € 12 as compensation for participating. An additional 14 participants completed the study, but 8 were excluded for failing the comprehension check or one of the three attention checks and 6 for returning incomplete response sets.

5.1.2 Procedure

The study consisted of two phases, which were separated by a period of approximately one week in order to avoid any carry-over or ordering effects: the truth and explanation-quality ratings phase, and the counter-examples generation phase.

The explanation-quality and truth ratings phase was completed first, via an on-line survey, as a preliminary to the participants being invited to visit the lab in person for the second phase. This first phase started with a comprehension task in which participants had to select a bad explanation amongst a list of possible explanations for an event. Then participants were asked to respond to a questionnaire consisting of three parts: an explanation-quality evaluation part, a truth rating part, and a distraction part with attention checks that doubled as distraction items.

The distraction part contained three questions: participants were asked to count the number of objects in a picture, to identify the color of a marble on a second picture, and to write a few sentences about school uniforms. The distraction part always appeared between the other two parts, which appeared first or last (randomized across participants).

The explanation-quality evaluation part was identical to the explanation-quality evaluation condition described in Experiment 1, and the truth rating part was identical to the truth rating condition described in Experiment 1. In both of these parts of the questionnaire, participants received all sixteen abductive conditionals that were selected in the pilot study described in Experiment 1.

The counter-examples generation phase was completed in person at the INSEAD–Sorbonne University Behavioral Lab. In this phase, participants completed the same task as the one described in the pilot study of Experiment 1, with three important differences. First, participants completed an additional practice trial, in which they were asked to select acceptable responses amongst a list of generated responses and received feedback explaining why a response was acceptable or not (e.g., why a counter-example was not sufficiently different from a previously given one). They then completed the practice trial described in the pilot study. A second difference was that participants were allowed sixty seconds per trial instead of ninety seconds, a change which was based on the observation in

the pilot that participants ran out of possible responses well before the ninety seconds mark. The final difference was that each participant completed both the alternative-explanations generation and the disabling-conditions generation tasks from the pilot study. These two tasks were presented in an order that was randomized per participant. Participants were allowed a short break between the two tasks. The generated responses were scored using the same procedure and criteria as the ones described in the pilot study. Due to experimenter error, one of the abductive conditionals was not included in this phase and it was therefore excluded from all analyses in this experiment.

5.2 Results and discussion

The average rating for the truth of the conditionals was 7.91 (SD = 2.97) and that of how well the conditionals' consequents explained their antecedents was 8.67 (SD = 2.57). The mean number of alternatives, averaged over all responses (i.e., over all participants and all conditionals), was 1.80 (SD = 1.31) and the mean number of disablers, also averaged over all responses, was 0.79 (SD = 0.96). Table 2 gives the correlations between all pairs of these variables, and Figure 4 provides a graphical overview of the responses, plotting truth versus the possible predictors.

While it is still common practice in psychology to analyze Likert-scale responses by means of metric models, Liddell and Kruschke (2018) have warned that doing so can have various untoward consequences (e.g., inflation of error rates, or inversions of between-group differences) and recommend cumulative ordinal regression models as a better alternative. Following their recommendation, we used the `brms` package to fit a number of Bayesian cumulative ordinal regression models (instead of Bayesian linear regression models, as we did in Experiment 1), also using again the weakly informative priors the `brms` package assumes by default.[6] (Note that the analyses from Experiment 1 were of aggregate Likert scale responses, which are represented on a continuous scale; the use of metric models was entirely appropriate there.)

Specifically, we fitted three cumulative ordinal regression models paralleling the linear models used in the analyses for Experiment 1. Thus, the models all had truth responses as dependent variable and had either explanation quality (EQ) as fixed effect, or both number of alternatives (A) and number of disablers (D) as fixed effects, or all three of those predictors. We again standardized

Table 2: Correlation matrix for all variables in Experiment 2.

	T	EQ	A	D
T	–	0.70	−0.52	0.04
EQ	0.70	–	−0.40	0.04
A	−0.52	−0.40	–	0.11
D	0.04	0.04	0.11	–

Note: T = truth; EQ = explanation quality; A = alternatives; D = disablers. All correlations except those between D and any of the other variables were significant at $\alpha = .0001$. Of the correlations involving D only the one with A was significant, and then only at $\alpha = .05$.

[6] An improper flat prior was used for all fixed effects, a t distribution ($\mu = 3$, $\sigma = 0$, df = 10) for the intercept of each of the thresholds and for each standard deviation, and a uniform LKJ distribution for the correlations between random effects.

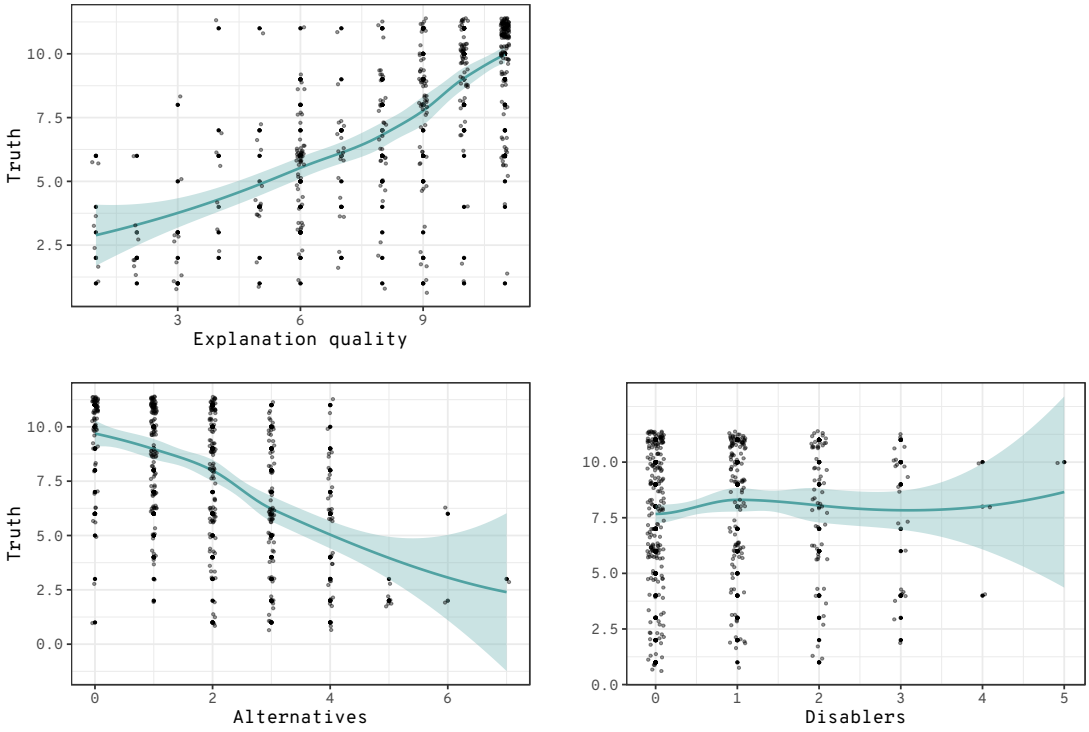

Figure 4: Participants' truth responses versus explanation-quality responses (top left), number of generated alternatives (bottom left), and of generated disablers (bottom right), with smoothers added to highlight trends (shaded areas indicate 95 percent confidence bands).

the predictors in the way recommended by Gelman. Also, following a recommendation by Barr, Levy, Scheepers, and Tily (2013), the models had a full random-effects structure, meaning that they included random slopes as well as random intercepts for both participants and items (abductive conditionals). Diagnostic tests gave no reason for concern.

A cumulative-link model estimates the distribution of responses provided on an ordinal scale, that is, a scale where the ordering of points is significant. It estimates the cumulative proportion of responses in log-odds at different thresholds, which correspond to the points on the scale. For instance, responses on a 7-point Likert scale will be estimated based on six thresholds, and for each threshold, the model will estimate the density of responses that are less than or equal to that threshold (the model provides no estimate for the highest point of the scale, given that all the responses are either less than or equal to it).

The LOOIC values for the three models are given in Table 3. We see that the model with only EQ as a predictor does best by this criterion. The coefficient for EQ equals 0.70 (posterior SD = 0.11), with a 95 percent credible interval from 0.50 to 0.92. This constitutes evidence that explanation quality influences truth ratings, and in particular, that the higher the truth rating of an abductive conditional tends to be, the better the conditional's consequent explains its antecedent. The credible intervals for A and D in either of the other two models both include 0—in the full model, they range from −0.52 to 0.10 and from −0.18 to 0.41, respectively—meaning that there is no consistent association between either alternative explanations or potential disablers generated by participants

Table 3: Comparison of Bayesian cumulative ordinal regression models.

predictor(s)	LOOIC	(SE)	ΔLOOIC	(SE)
EQ, A, D	1333.3	(43.1)	8.8	(9.0)
A, D	1448.7	(39.8)	124.2	(26.2)
EQ	1324.5	(42.4)	0.0	—

Note: EQ = explanation quality; A = number of alternatives; D = number of disablers. For further explanation, see the note to Table 1.

and their truth rating for the conditional. Figure 5 plots the posterior distributions for the three predictors, as estimated in the full model.

It is worth adding here that the interpretation of the coefficients in an ordinal regression model is different from that of coefficients in a linear model. For instance, the estimate of 0.70 for EQ in the best model indicates that for each increase by one unit in that variable, there will be a multiplicative effect of $\exp(0.7) \approx 2.01$ on the cumulative odds of obtaining a truth rating greater than a given point on the Likert scale as opposed to obtaining a truth rating not greater than that point, ceteris paribus (see, e.g., Agresti & Tarantola, 2018).

Thus our findings in the first experiment were more than reconfirmed as far as inferentialism is concerned, providing even stronger support for it. They show that the degree to which a participant deemed ψ a good explanation of φ (which for an abductive conditional "If φ then ψ" determines the strength of the inferential connection between its component parts) predicted to what extent that participant would agree that the conditional is true. However, they also show that at the individual level there is not even partial support for a mental models account according to which the participant's truth rating depends on the results of a search in semantic memory for counterexamples, which will have a higher likelihood of being retrieved when that participant is able to generate a

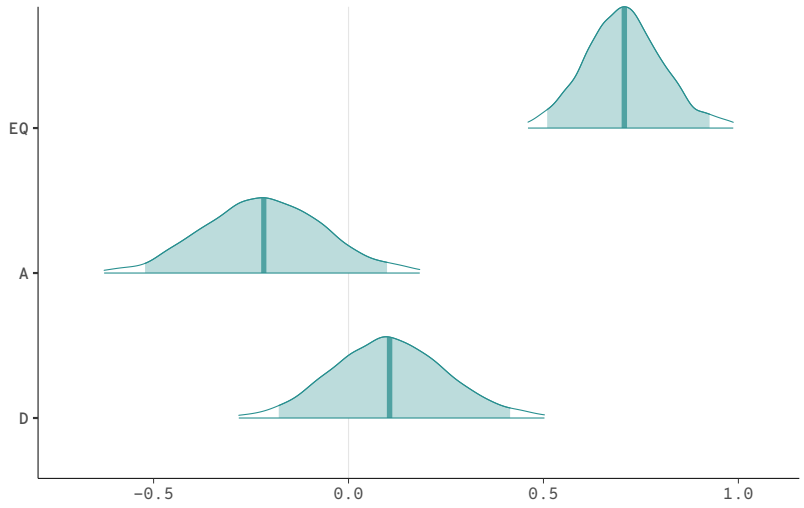

Figure 5: Posterior distributions of explanation quality (EQ), number of alternatives (A), and number of disablers (D) in the full model for Experiment 2.

higher number of alternative explanations of the antecedent and/or a higher number of events that could prevent the consequent from producidng the antecedent.

There is still no inconsistency here with the results from Cummins, De Neys, and their co-authors, who—to repeat—were interested in causal rather than in abductive conditionals. Furthermore, while we only looked at the *number* of generated alternatives, following the procedure of the aforementioned authors, it would be interesting to examine whether controlling for the *quality* of each generated alternative might lead to a different outcome, and might show an impact of considered alternatives. We flag this possibility here to set it aside for future research.[7]

6 Experiment 3

Experiment 3 aims to investigate the inferentialist hypothesis that the strength of the inferential connection between a conditional's antecedent and consequent contributes to the role that conditional plays in MP arguments. This puts inferentialism to the test in a way that was not previously attempted, even though the experimental paradigm we use is about as old as the psychology of reasoning.

To understand how a conditional contributes to an MP argument, it is worth briefly noting that the inferentialist's take on the conditional is very close to the intuitionists' one. For intuitionists, a conditional is a method that enables us to turn any proof we have for the antecedent into a proof for the consequent (van Dalen, 2001; Sundholm & van Atten, 2008; van Atten, 2017). Where the intuitionists' account was developed for conditionals in mathematical contexts, we consider instead natural language conditionals. By substituting the intuitionists' notion of proof with the more informal one of support, we basically obtain inferentialism: a conditional can be regarded as providing a method for carrying over whatever support one has for the antecedent to the consequent. After all, the truth of a conditional requires the existence of a compelling argument from antecedent to consequent, and compelling arguments are supposed to turn one's grounds for believing the premises into grounds for believing the conclusion.

On this understanding of conditionals, we can picture them as *conduits*, or *pipes*, or indeed *inference tickets*, in the terminology of Ryle (1950) encountered earlier. Because the connecting argument need only be compelling, however, and not conclusive, the pipe can be somewhat leaky, in that some of the support may seep away in the transfer; or to use a different metaphor, the inference ticket may not always allow you to travel all the way to your destination, but will sometimes only bring you close to where you need to be. Less prosaically, on an inferentialist understanding of the conditional a true conditional will pass on one's support (if any) for the antecedent onto the consequent, albeit in a possibly not entirely lossless manner. In these terms, we can also easily further our understanding of why we do not encounter missing link conditionals in quotidian speech: such conditionals cannot fulfill the important support-passing function of normal conditionals, because what is supposed to do the passing—the argument connecting antecedent and consequent—is lacking. A missing link conditional is a ticket that does not even allow you to hop on the inference train.

More to the point, the inferentialist's take on conditionals as possibly imperfect conduits has interesting consequences for how they may figure as premises in arguments. Consider modus ponens (MP), which—as said—licenses the inference of ψ from φ together with "If φ then ψ." Previous

[7] We thank an anonymous reviewer for suggesting this possible extension of the experiment.

experimental work concerning MP has shown endorsement rates of the conclusion to be almost at ceiling—almost, but not quite (see Evans & Over, 2004, and references given there). Naturally, the fact that MP arguments are not unanimously endorsed by participants might be a matter of noise, which will creep in no matter how carefully we design our experiments. But from an inferentialist perspective, the admission that we have no way of keeping noise entirely at bay may not be the only way to account for the finding that MP endorsement tends to be not *exactly* at ceiling.

Another way to explain this finding is to build on two implications of inferentialism. First, that MP is not a valid argument form: a conditional can be true even if the argument leading from antecedent to consequent is not conclusive, and so it can happen that φ and "If φ then ψ" are both true while ψ is nevertheless false. Second, that conditionals can be somewhat leaky pipes, and that it is therefore reasonable to suppose that, in an MP argument, our willingness to endorse the conclusion will be affected by *how* leaky we assess the pipe to be.

To continue the conditional-as-conduit metaphor, we can think of going through an MP argument as a process in which water gets transferred from one bucket to a second bucket. The minor premise and the conclusion of the MP argument are those buckets, and at the start of the argument, the minor-premise bucket contains a certain amount of water while the conclusion bucket is still empty. The MP argument apparatus works by pouring the water contained in the minor-premise bucket into a conduit—the major-premise conditional—which we can mentally picture as being slightly tilted toward the conclusion bucket. Let amounts of water represent how confident you are in the truth of a statement: the MP argument will carry the amount of confidence you have in the minor premise through the conditional over to the conclusion. How much water ends up in the conclusion bucket, however, will depend not just on the amount of water in the minor-premise bucket, but also on the leakiness of the pipe, on how much water is lost along the way. The leakiness is a matter of degree, with deductive conditionals constituting one endpoint—no water is spilled—and missing link conditionals constituting the other, being pipes so leaky that, no matter the amount of water poured into them, not a drop will reach the conclusion bucket. To put it more succinctly: your confidence in the conclusion of an MP argument depends not only on your confidence in the minor premise but also on your assessment of how compelling the argument from the major premise's antecedent to its consequent is. That, at any rate, is what inferentialism implies.

We test this implication in the form of the following hypothesis:

H3: The strength of the inferential connection between the antecedent and consequent of the major premise in an MP argument predicts the rate of endorsement of the conclusion of that argument, ceteris paribus (in particular, keeping fixed how confident one is in the minor premise).

As in the case of H1, our focus on abductive conditionals allows us to make H3 a little more specific:

H3*: The explanation quality of the consequent of the major premise of an MP argument, with the premise's antecedent given as explanandum, predicts the rate of endorsement of the conclusion of the argument, ceteris paribus.

To the best of our knowledge, there has so far been no research on this hypothesis, unlike in the case of H1.

There has been closely related research, however. George (1995, Expt. 1) found that the degree to which his participants deemed the major premise in an MP argument uncertain reliably predicted the degree to which they were uncertain about the conclusion. And Stevenson and Over (1995)

showed that when the major premise in an MP argument is qualified (either by adding premises that make it uncertain or by qualifying it directly), participants will prefer a qualified conclusion that expresses a degree of uncertainty (see also, Stevenson & Over, 2001). Accordingly, the suppositional account might also be able to predict the "noise" supposedly responsible for the experimental results that find MP endorsement to be typically "not-quite-at-ceiling." Most notably, advocates of this account might hold that, even if participants are asked to suppose the major premise, they may have their own ideas about how *certain* that premise is, therefore assigning a probability to it that is not necessarily 1. And to the extent that the participants are uncertain about the major premise, their certainty in the conclusion of the argument may diminish. So, the hypothesis

H4: The probability assigned to the major premise in an MP argument predicts the endorsement rate of the conclusion, ceteris paribus,

could be proposed as the suppositional theorists' rival to H3. One interesting consequence of the prediction made by H3 is that it can serve a double purpose: to evaluate inferentialism and also to help discriminate between inferentialism and the suppositional account.

The discrimination criterion is, which of explanation quality and probability best predicts the "not-quite-at-ceiling" effect, if either yields accurate predictions to begin with. Importantly, H3 and H4 are not inconsistent with each other, nor is H4 incompatible with inferentialism. There is nothing in inferentialism to preclude that *both* explanation-quality ratings *and* probabilities contribute significantly to predicting endorsement rates. Indeed, participants may be uncertain about the major premise for any number of reasons in addition to the ones identified by inferentialism, and such reasons may affect truth ratings as well.

More generally, although the notion at the heart of inferentialism—the strength of the argument connecting a conditional's antecedent and consequent—is not an *explicitly* probabilistic notion, inferentialism is not committed to some form of anti-reductionism according to which argument strength cannot possibly be captured in purely probabilistic terms. A number of recent publications suggest that this may be possible indeed (see, e.g., Hahn & Oaksford, 2007a, 2007b; Hattori & Oaksford, 2007; Douven, 2008; Eva & Hartmann, 2018; Skovgaard-Olsen et al., 2019). While the suggestions point in somewhat different directions, the idea common to them is to define probabilistically the notion of one proposition being *relevant* to another, which the conditional probability of consequent given antecedent—the notion central to the suppositional account—is unable to do: any given conditional probability is consistent with the propositions at issue being (probabilistically as well as intuitively) irrelevant to each other.

As Hattori and Oaksford (2007) note, the current literature features more than forty candidate-definitions of probabilistic relevance. Here, we consider only the ones that have received attention in the debate about conditionals, which are the difference measure of confirmation (Carnap, 1962; Douven, 2008; Douven & Verbrugge, 2012), the Δp rule (Shanks, 1995; Evans & Over, 2004; Skovgaard-Olsen et al., 2019), and Cheng's (1997) power PC measure, which plays an important role in the causal Bayes nets approach as advocated in Ali, Schlottmann, Shaw, Chater, and Oaksford (2010), Ali, Chater, and Oaksford (2011), Fernbach and Erb (2013), Oaksford and Chater (2013, 2014, 2017, 2020a), Hall, Ali, Chater, and Oaksford (2016), and van Rooij and Schulz (2019). According to the difference measure, the degree to which φ is relevant to ψ is given by $\Pr(\psi \mid \varphi) - \Pr(\psi)$. According to the Δp rule, it is given by $\Pr(\psi \mid \varphi) - \Pr(\psi \mid \neg\varphi)$. And where $\Delta p(\varphi, \psi) := \Pr(\psi \mid \varphi) - \Pr(\psi \mid \neg\varphi)$, Cheng's measure defines the degree to which φ is relevant to ψ to be $\Delta p(\psi, \varphi) \div \Pr(\neg\psi \mid \neg\varphi)$ if $\Delta p(\psi, \varphi) \geqslant 0$ and $\Delta p(\psi, \varphi) \div \Pr(\psi \mid \neg\varphi)$ otherwise. Cheng's proposal is meant to measure *causal*

connectedness.[8] Given that our materials consist strictly of abductive conditionals which build on causal connections (they were all derived from causal conditionals), Cheng's measure might, in the present context, be particularly suited to measure *inferential connectedness* as well.

We are not proposing that inferentialists commit to any of the above measures. In the following, we use these measures to construct additional predictors of truth ratings, next to explanation-quality ratings (as measured directly) and conditional probabilities. Whether those further predictors should be conceived as probabilistic explications of argument strength, or degree of inferential connectedness, or rather as offering the building blocks for possible alternatives to both inferentialism and the suppositional account is a question we leave open here.[9]

6.1 Method

6.1.1 Participants

Participants in Experiment 3 were 120 adults recruited by the INSEAD–Sorbonne University Behavioural Lab (70 females; $M_{age} = 22.6$, $SD_{age} = 3.5$) after approval by the INSEAD's Ethical Committee. They completed the three phases of the study via an on-line questionnaire after giving informed consent and were compensated through a lottery system, with nine randomly selected participants receiving € 25 each. An additional 26 participants completed the three phases of the study but were excluded for failing any of three attention checks or for having received an advance training in logic.

6.1.2 Materials and procedure

The study consisted of three phases, which were separated from each other by a period ranging from three to seven days: the explanation-quality evaluation phase, the argument evaluation phase, and the probabilistic truth-table phase. Because of practical reasons, the order of these three phases was not randomized, instead they were separated by a sufficiently long delay to avoid carry-over and ordering effects.

Participants first completed the explanation-quality evaluation phase of the experiment, which was identical to the explanation-quality evaluation condition described in Experiment 1. Participants rated the quality of the explanation provided on a 7-point Likert scale.

The survey for the argument evaluation phase was sent to participants approximately five days after completion of the explanation-quality evaluation phase. In this second phase, participants were asked to rate the truth of the conclusions of sixteen MP arguments. These MP arguments were constructed such that the abductive conditionals selected in the pilot study served as their major premises and information described as being provided by a witness as their minor premises; for instance, "Dennis tells you that John did well on his exam. Now suppose that if John did well on his exam, then he studied hard."

For control purposes, we varied minor-premise probability. We told participants at the beginning of this phase that they were going to receive information from four different sources ("witnesses"), each with a different track record of truth-telling: 100 percent, 75 percent, 50 percent, and 25 percent. The name (and gender) of the witnesses was picked randomly for each participant. Participants were shown an example of the task they would complete in the study and responded to

[8] Accordingly, she speaks of the *generative / inhibitory power* of φ for ψ if $\Delta p(\psi, \varphi) \geqslant / < 0$.

[9] In view of the success of the new paradigm in the psychology of reasoning (Over, 2009; Elqayam & Over, 2013; Oaksford & Chater, 2020b), it would certainly be *desirable* to have a purely probabilistic definition of argument strength.

a memory check about the reliability of each witness. They then completed the main part of the survey. A reminder of the reliability of the four witnesses was indicated at the top of each question. After reading each argument, participants were asked, "How strongly do you agree that it is true that . . . [e.g., John studied hard]?" Participants provided their responses on a 7-point Likert scale, with every point labeled, ranging from "Strongly disagree" to "Strongly agree," with the midpoint labeled "Neither agree nor disagree."

After completing the main part of the survey, participants were asked to estimate on a 100 point scale how reliable each of the witnesses was, in their opinion. They were instructed that 0 meant that the witness was not reliable at all and never told the truth, that 50 meant that the witness was moderately reliable and told the truth half the time, and that 100 meant that the witness was absolutely reliable and always told the truth. Participants had been reminded throughout the experiment how reliable each witness was, so this part of the questionnaire served as an attention check.[10] Participants also responded to a few demographic questions and indicated whether they had responded seriously to the survey.

The survey for the probabilistic truth-table phase was sent to participants approximately three days after they had completed the argument evaluation phase. This task asked participants to rate the probability of four mutually exclusive and jointly exhaustive situations, which were presented in the schematic form of $\varphi \,\&\, \psi$, $\varphi \,\&\, \neg\psi$, $\neg\varphi \,\&\, \psi$, and $\neg\varphi \,\&\, \neg\psi$, where φ was always the antecedent of a conditional from our materials and ψ the consequent of the same conditional. Participants completed sixteen such truth tables, corresponding to the sixteen abductive conditionals from the previous phases. These truth tables were presented in an order randomized across participants. For each situation, participants indicated a value in a text box and were asked to rate each situation on a probability scale ranging from 0 to 100 percent, where they were instructed that 0 percent meant that the situation would certainly not occur and 100 percent meant that the situation would certainly occur. They were also instructed that the probabilities had to sum to 100 percent and they were not allowed to proceed with the survey until their answers matched this constraint.

6.2 Results and discussion

The purpose of this experiment was twofold. First, the aim was to test H3 and see whether the truth ratings of the conclusions of the MP arguments collected in the second phase of the experiment could indeed be predicted on the basis of the explanation-quality ratings collected in the first phase. Second, we wanted to compare the accuracy of those predictions with the corresponding predictions based on the conditional probabilities or on one of the other probabilistic measures that could be derived from the responses to the probabilistic truth table task the participants were asked to complete in the third phase of the experiment.

The mean for the truth ratings was 5.05 (SD = 1.63), and that for the explanation-quality ratings was 5.71 (SD = 1.60). For any of the major premises "If φ then ψ," we could derive $\Pr(\psi \mid \varphi)$ from the lines in the truth table of the form $\varphi \,\&\, \psi$ and $\varphi \,\&\, \neg\psi$. Summing the numbers on those lines yielded $\Pr(\varphi)$, and dividing the probability assigned to $\varphi \,\&\, \psi$ by that sum yielded the sought conditional probability, and hence, assuming the Equation, also what according to the suppositional account

[10] An anonymous reviewer pointed out to us that this attention check may actually have not been very useful, because we were asking participants to estimate a quantity they had already been told, which could be pragmatically strange and therefore confuse the participants. We reran the analysis with the fourteen participants who failed the attention check included, finding no qualitative differences and only very minor quantitative differences. (For interested readers, the relevant `R` file in the Supplementary Materials contains instructions indicating how to carry out the rerun.)

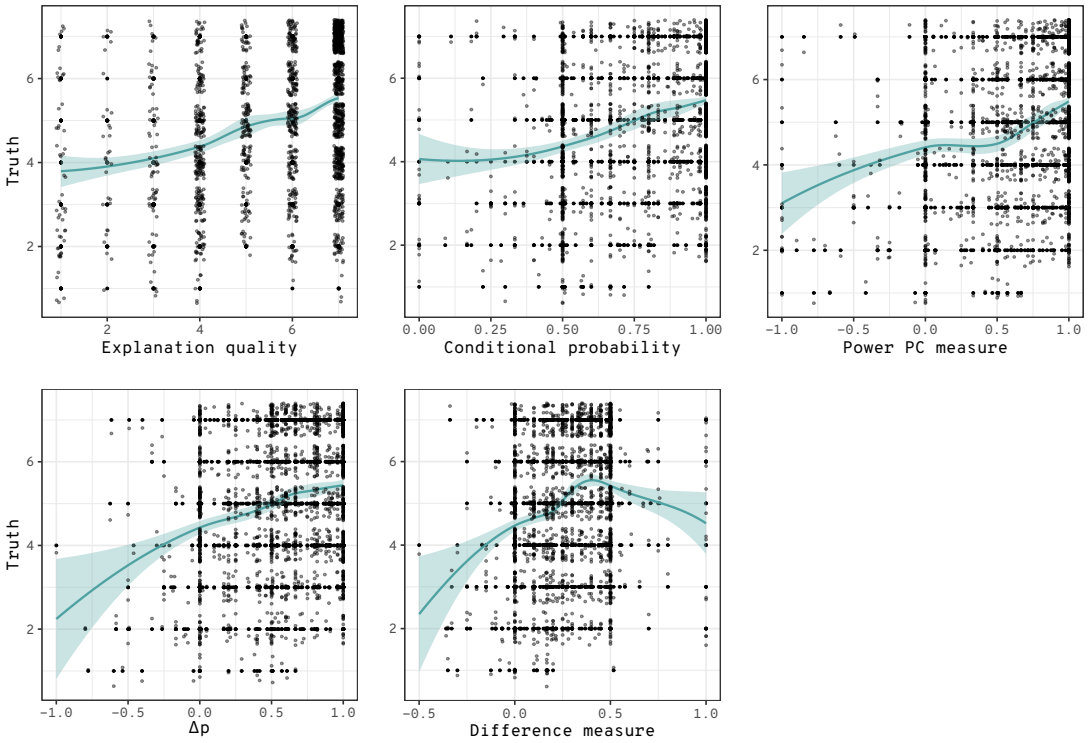

Figure 6: Participants' truth responses versus their explanation-quality responses and their computed values for the various probabilistic measures, with smoothers added to highlight trends (shaded areas indicate 95 percent confidence intervals).

is the probability of the major premise. The mean conditional probability was .80 (SD = .24). In a similar way, we derived values for the other variables of interest: values for the difference measure of confirmation (M = 0.28, SD = 0.21), Δp values (M = 0.58, SD = 0.38), and values for Cheng's power PC measure (M = 0.68, SD = 0.42). Relatively high means for ratings of explanation quality and truth, and for conditional probabilities as well as for the other probabilistic measures (which all have a range from −1 to 1), can be explained by the fact, already mentioned above, that we based our materials on causal conditionals that presented an at least "moderately strong causal relationship" (according to the results provided in Cummins et al., 1991).

Figure 6 gives a graphical summary of the data, plotting truth responses against explanation-quality responses and the values for the probabilistic measures computed as just explained. The smoothers suggest that truth ratings increase with increases in each of the other variables (this is also true for the difference measure in the part of the plot where most of the data are). However, it is not obvious from the figure whether any variable predicts truth ratings reliably, nor which does so most accurately.

We again used the `brms` package to fit a number of cumulative ordinal regression models, all using the default weakly informative priors,[11] and also all with the truth responses as measured on

[11] An improper flat prior was used for all fixed effects, a t distribution ($\mu = 3$, $\sigma = 0$, df = 10) for the intercept of each of the thresholds and for each standard deviation, and a uniform LKJ distribution for the correlations between random effects.

Table 4: Correlation matrix for all variables in Experiment 3.

	T	EQ	CP	PPC	Δp	DIF
T	–	0.33	0.30	0.29	0.25	0.23
EQ	0.33	–	0.42	0.40	0.38	0.36
CP	0.30	0.42	–	0.93	0.84	0.83
PPC	0.29	0.40	0.93	–	0.91	0.86
Δp	0.25	0.38	0.84	0.91	–	0.91
DIF	0.23	0.36	0.83	0.86	0.91	–

Note: T = truth; EQ = explanation quality; CP = conditional probability; PPC = power PC measure; DIF = difference measure. All correlations were significant at $\alpha = .0001$.

a 7-point Likert scale as dependent variable and with witness reliability as fixed effect. Because all probabilistic measures turned out to be highly correlated with each other, while being only modestly correlated with explanation quality—as seen in Table 4—we did not include more than one probabilistic measure as a predictor in any model. We considered all models that decision left us with, these being five models with either explanation quality *or* one of the probabilistic measures as a fixed effect next to witness reliability, and four models with explanation quality *and* one of the probabilistic measures as additional fixed effects. All models had a full random-effects structure, meaning that they had, for all their fixed effects, both random intercepts and random slopes for participants as well as items. Here, too, we followed Gelman's recommendation and standardized the predictors by centering them at their means and dividing by twice their standard deviation. Diagnostic tests gave no cause for concern.

LOOIC values for the models we fitted are given in Table 5. It can be seen that all the models that include EQ as a fixed effect do better than any model without it. It can also be seen that the model with both EQ and CP (besides WR) as predictors does best, closely followed by the model with both EQ and PPC. Of the models with only one fixed effect next to WR, the one with EQ does markedly better than any of the others.

Table 5: Comparison of Bayesian cumulative ordinal regression models.

predictors	LOOIC	(SE)	ΔLOOIC	(SE)
WR, EQ, CP	4057.6	(74.4)	0.0	—
WR, EQ, PPC	4068.9	(74.5)	11.4	(8.2)
WR, EQ, Δp	4084.9	(74.1)	27.4	(12.0)
WR, EQ, DIF	4079.9	(75.0)	22.2	(12.6)
WR, EQ	4108.0	(73.9)	50.4	(17.2)
WR, CP	4173.2	(71.6)	115.6	(29.0)
WR, PPC	4209.5	(71.0)	151.8	(31.4)
WR, Δp	4246.2	(70.1)	188.6	(34.2)
WR, DIF	4226.7	(71.1)	169.0	(33.4)

Note: WR = witness reliability; EQ = explanation quality; CP = conditional probability; PPC = power PC measure; DIF = difference measure. For further explanation, see the note to Table 1.

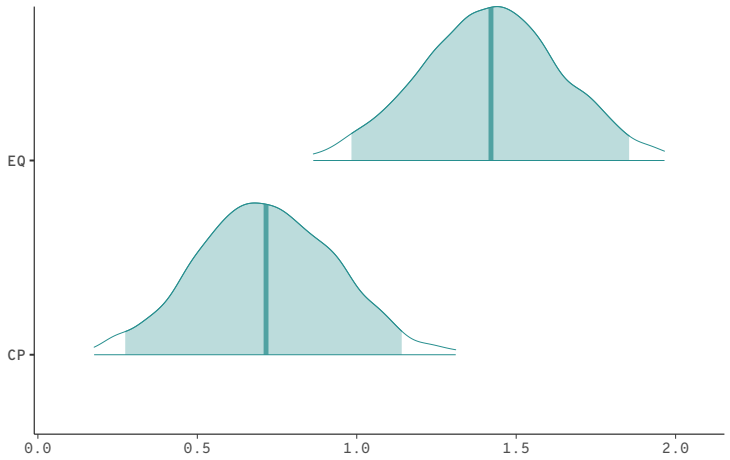

Figure 7: Posterior distributions of conditional probability (CP) and explanation quality (EQ) in the best model from Experiment 3.

As for the estimates of the fixed effects in the best model, we found a posterior mean of 1.42 for EQ (posterior SD = 0.22), with a 95 percent credible interval from 0.98 to 1.85, and a posterior mean of 0.72 for CP (posterior SD = 0.22), with a 95 percent credible interval from 0.27 to 1.14. The posterior distributions for these parameters are shown in Figure 7.

The most important observation to make is that, in the best model, EQ clearly had a larger impact on truth ratings than CP, the difference between EQ and CP being 0.71, with a 95 percent credible interval from 0.13 to 1.26. To bring this further into relief, we derived outcome probabilities from the model (Kruschke, 2015, Ch. 23) and determined the overall effect on truth ratings of the predictors, using again the `brms` package. Figure 8 gives the graphical representation of the result. The difference in impact on truth ratings between EQ and CP is immediately manifest from this figure, as the steeper slope of EQ is easy to observe.

In summary, this analysis of our data arrived at the following conclusions: There is strong evidence that how well the consequent of the major premise explains the antecedent of that premise

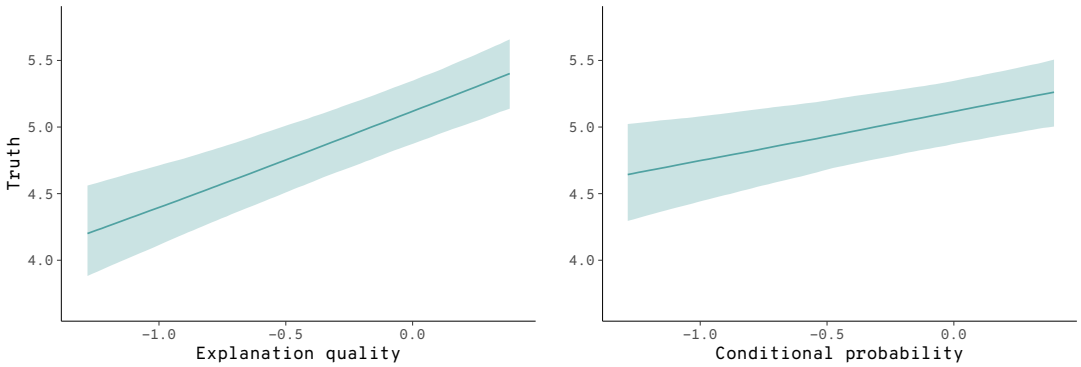

Figure 8: Overall marginal effects of explanation quality (left) and conditional probability (right) on truth ratings, as estimated in the best Bayesian regression model. The bands represent 95 percent credible intervals.

predicts the rate at which the consequent will be endorsed as the conclusion of an MP argument. This is exactly as predicted by inferentialism, and it is the content of H3. There is evidence for H4 as well. Importantly, however, we saw that EQ has an impact on truth ratings roughly twice as large as that of CP. We thus take our results to support inferentialism over the suppositional account. As for the other probabilistic measures, our results give little reason to suppose that any of them captures adequately the notion of explanation quality in probabilistic terms. After all, in terms of LOOIC values, all models having one of those measures as their fixed effect next to WR do much worse than the model that has EQ as its fixed effect next to WR.

These conclusions merit some comments, however. First, it is worth repeating a point already made in Douven et al. (2019), to wit, that inferentialism still needs to be complemented by a story about the probabilities of conditionals. Only then will we be able to give a proper verdict about the inferentialist position. Only then will we be able to say, for instance, whether the impact on truth ratings that conditional probabilities had in the third experiment can be explained by inferentialism or whether it is perhaps an indication that the position can at best only be part of the truth about conditionals. At the moment, inferentialism has nothing informative to say about that finding. In the meantime, it may still be noted that even the complete absence of an explanation of why conditional probability was a reliable predictor of truth ratings does not undercut our conclusion regarding how inferentialism does compared with the suppositional account: for all we know, that account has nothing informative to say about how truth ratings were impacted by judgments concerning how well the consequent of an MP argument's major premise explains that premise's antecedent.

Second, all probabilistic measures (including conditional probabilities) were derived from participants' responses to a probabilistic truth table task. This kind of task has been used in previous research, apparently without problems (e.g., in Evans, Handley, & Over, 2003; ?, ?). Nevertheless, it is to be admitted that the mental arithmetic needed to complete this task probably makes it more challenging, and more susceptible to errors, than the task of judging the quality of an explanation.[12] Future research could consider simpler ways of eliciting probabilities (e.g., conditional probabilities can be obtained via the so-called Ramsey test, which asks participants to *suppose* the antecedent of a conditional probability and then to assess, under that supposition, the probability of the consequent). Such research might lead to results qualitatively different from ours, and might support the claim that, for instance, Cheng's measure is superior in predicting truth ratings after all.

7 General discussion

Across three experiments, we found strong support for inferentialism, a position according to which, in its latest version, the truth of a conditional requires the presence of a strong enough argument from antecedent (plus background premises) to consequent. There was already some support for this position, but we wanted to contribute to the literature by using more realistic materials than had been used in previous work. Our materials involved abductive conditionals, that is, conditionals in which the connection between antecedent and consequent consists of an explanatory link: the consequent explains, to a lower or higher degree, the antecedent, thereby creating an abductive inferential relation between the two, where the strength of the argument is a function of exactly *how* well the consequent explains the antecedent. The materials were produced by simply switching the

[12] Thanks to Mike Oaksford for bringing this to our attention.

role of antecedent (cause) and consequent (effect) in conditionals previously used in research on causal reasoning.

We put inferentialism to the test in two different ways: one which offered an important extension of the work reported in Douven et al. (2018, 2019) by using realistic materials, and one which considered what inferentialism implies for the evaluation of MP arguments. The first centered around H1, the hypothesis that the truth rating of an abductive conditional is predicted by how well the consequent of that conditional is perceived to explain its antecedent. This hypothesis was tested both between subjects (Experiment 1) and within subjects (Experiment 2), both tests yielding favoring evidence. The analyses of these experiments also contrasted inferentialism with a rival, mental-models-inspired account. According to such an account, the various abductive conditionals in our materials would be deemed true to a degree that depended on the number of counterexamples (alternative explanations and disabling conditions) that people were able to generate for them. That account turned out to receive not nearly as much support from our data.

In Experiment 3, we tested a further hypothesis, H3, and investigated whether we could predict people's truth ratings of the conclusion of an MP argument based on the inferential strength of the embedded conditional (the major premise), considering also the reliability of the minor premise, which was varied for control purposes. We compared this hypothesis with a probabilistic hypothesis, H4, according to which people's agreement with the conclusion should be better predicted by the conditional probability of the argument's major premise. In a within-subjects design, participants evaluated the quality of sixteen explanations, given an explanandum (these corresponded respectively to the consequent and the antecedent of abductive conditionals); they rated the truth of the conclusions of MP arguments with abductive conditionals as their major premises; and they completed a probabilistic truth-table task for each of the sixteen conditionals, on the basis of which we could compute what, according to the so-called Equation, should be the participants' probabilities for the major premises, to wit, their probabilities for the consequents of those premises conditional on their respective antecedents.

We saw that explanation quality was a stronger predictor of truth ratings than conditional probability, but also that conditional probability was a reliable predictor even when explanation quality was present in a model. While this supports inferentialism, and also supports inferentialism over the suppositional account, which revolves around the Equation, it underscores the need for further theoretical work on inferentialism. Namely, the proponents of that position need to equip their position with a story that connects the semantics—the account of truth conditions of conditionals, according to inferentialism—with probability theory. The connection would be straightforward if the notion at the core of inferentialism—that of argument strength, or inferential connectedness—could be captured in probabilistic terms. While we are still open to the possibility that it can, our results gave reason to believe that argument strength is not captured by any of the prima facie most promising probabilistic candidate definitions.

One limitation of this research is that the materials used, while concrete, did not refer to actual situations that the participants had experienced, nor to actual uses in natural conversational contexts, which are often more complex than modus ponens arguments. Further research could therefore also examine the interpretation of conditionals when they are used in more natural conversational contexts with a richer background of information, and it could extend the study of conditionals to a more varied and complex set of arguments (for examples, see Douven, 2016, p. 129) than the simpler argument forms that have been studied until now. Further research could also try

to elicit the relevant probabilities in a way that might be easier or more intuitive for participants than a probabilistic truth table task, thereby addressing another limitation that was noted above.[13]

References

Agresti, A., & Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, *72*(3), 210–223.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Ali, N., Schlottmann, A., Shaw, A., Chater, N., & Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 117–134). Oxford: Oxford University Press.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535.

Baratgin, J., Douven, I., Evans, J. S. B. T., Oaksford, M., Over, D., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, *19*(10), 547–548.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge MA: MIT Press.

Bürkner, P.-C. (2017). `brms`: An `R` package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carnap, R. (1962). *Logical foundations of probability*. Chicago: University of Chicago Press.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646–658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*(3), 274–282.

De Neys, W., Schaeken, W., & D'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, *31*(4), 581–595.

Douven, I. (2008). The evidential support theory of conditionals. *Synthese*, *164*(1), 19–44.

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philosophical Quarterly*, *63*(252), 428–444.

Douven, I. (2016). *The epistemology of indicative conditionals: Formal and empirical approaches*. Cambridge UK: Cambridge University Press.

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

[13] We are greatly indebted to Christopher von Bülow and to three anonymous referee for valuable comments on a previous version of this paper, as well as to Mike Oaksford for excellent editorial advice. We also thank the staff of the INSEAD–Sorbonne University Behavioral Lab for their generous help with all logistic matters, and we wish to acknowledge the contributions of Manon André, Adrien Avramoglou, Robin Duclermortier, Louis Dussarps, Camille Fouché, and Théo Jesu, who coded participant responses, and of Guillain Potron, who wrote a Python script to format the coding spreadsheets. Finally, we are grateful to audiences at the University of Amsterdam and at the London Reasoning Workshop for stimulating remarks and discussion.

Douven, I. (2017b). Inference to the best explanation: What is it? And why should we care? In T. Poston & K. McCain (Eds.), *Best explanations: New essays on inference to the best explanation* (pp. 4–22). Oxford: Oxford University Press.

Douven, I. (2019). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence*, *275*, 235–251.

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science*, in press.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2019). Conditionals and inferential connections: Toward a new semantics. *Thinking & Reasoning*, in press.

Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1792–1813.

Douven, I., & Verbrugge, S. (2010). The Adams family. *Cognition*, *117*(3), 302–318.

Douven, I., & Verbrugge, S. (2012). Indicatives, concessives, and evidential support. *Thinking & Reasoning*, *18*(4), 480–499.

Douven, I., & Verbrugge, S. (2013). The probabilities of conditionals revisited. *Cognitive Science*, *37*(4), 711–730.

Douven, I., & Wenmackers, S. (2017). Inference to the best explanation versus Bayes's rule in a social setting. *British Journal for the Philosophy of Science*, *68*(2), 535–570.

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

Eva, B., & Hartmann, S. (2018). Bayesian argumentation and the value of logical validity. *Psychological review*, *125*(5), 806–821.

Evans, J. S. B. T. (2006). The heuristic–analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395.

Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove: Psychology Press.

Evans, J. S. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335.

Evans, J. S. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1327–1343.

Fugard, A. J., Pfeifer, N., Mayerhofer, B., & Kleiter, G. D. (2011). How people interpret conditionals: Shifts toward the conditional event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 635–648.

Gauffroy, C., & Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: An integrative developmental theory. *Developmental Review*, *29*(4), 249–282.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873.

George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology*, *86*(1), 93–111.

Glass, D. H. (2007). Coherence measures and inference to the best explanation. *Synthese*, *157*(3), 275–296.

Glass, D. H. (2012). Inference to the best explanation: Does it track truth? *Synthese*, *185*(3), 411–427.
Grice, H. P. (1989). Indicative conditionals. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 58–85). Cambridge MA: Harvard University Press.
Hadjichristidis, C., Stevenson, R. J., Over, D. E., Sloman, S. A., Evans, J. S. B. T., & Feeney, A. (2001). On the evaluation of *If p then q* conditionals. In *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 381–386).
Hahn, U., & Oaksford, M. (2007a). The burden of proof and its role in argumentation. *Argumentation*, *21*(1), 39–61.
Hahn, U., & Oaksford, M. (2007b). The rationality of informal argumentation: a bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704–732.
Hall, S., Ali, N., Chater, N., & Oaksford, M. (2016). Discounting and augmentation in causal conditional reasoning: Causal models or shallow encoding? *PLoS ONE*, *11*. Retrieved from e0167741.http://dx.doi.org/10.1371/journal.pone.0167741
Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive science*, *31*(5), 765–814.
Janveau-Brennan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology*, *35*(4), 904.
Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.
Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford: Oxford University Press.
Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd ed.). Boston: Academic Press.
Krzyżanowska, K. (2015). *Between "if" and "then"* (Unpublished doctoral dissertation). University of Groningen.
Krzyżanowska, K., Collins, P. J., & Hahn, U. (2017). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199–205.
Krzyżanowska, K., Wenmackers, S., & Douven, I. (2013). Inferential conditionals and evidentiality. *Journal of Logic, Language and Information*, *22*, 315–334.
Krzyżanowska, K., Wenmackers, S., & Douven, I. (2014). Rethinking Gibbard's riverboat argument. *Studia Logica*, *102*(4), 771–792.
Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.
Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Thinking & Reasoning*, *6*(4), 335–347.
Markovits, H., Fleury, M.-L., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child development*, *69*(3), 742–755.
Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition*, *29*(5), 736–744.
McGee, V. (1985). A counterexample to Modus Ponens. *Journal of Philosophy*, *82*(9), 462–471.
Mill, J. S. (1843). *A system of logic*. London: Longmans, Green, and Company.
Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind & Language*, *18*(4), 359–379.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2010). Causation and conditionals in the cognitive science of human reasoning. *The Open Psychology Journal*, *3*, 105–118.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*(3–4), 346–379.

Oaksford, M., & Chater, N. (2014). Probabilistic single function dual process theory and logic programming as approaches to non-monotonicity in human vs. artificial reasoning. *Thinking & Reasoning*, *20*(2), 269–295.

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 327–346). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2020a). Integrating causal Bayes nets and inferentialism in conditional inference. In S. Elqayam, I. Douven, J. S. B. T. Evans, & N. Cruz (Eds.), *Logic and uncertainty in the human mind.* London: Routledge.

Oaksford, M., & Chater, N. (2020b). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 1–26.

Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? the roles of covariation and causality. *Thinking & Reasoning*, *13*(4), 340–369.

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*, 431–438.

Over, D. E., Douven, I., & Verbrugge, S. (2013). Scope ambiguities and conditionals. *Thinking & Reasoning*, *19*(3–4), 284–307.

Over, D. E., & Evans, J. S. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind & Language*, *18*(4), 340–358.

Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 153–173). Oxford: Oxford University Press.

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking & Reasoning*, *16*(3), 172–197.

Ryle, G. (1950). "If", "so", and "because". In M. Black (Ed.), *Philosophical analysis: A collection of essays.* Ithaca NY: Cornell Press University.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, *48*, 257–279.

Skovgaard-Olsen, N. (2016). Motivating the relevance approach to conditionals. *Mind & Language*, *31*(5), 555–579.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. (2019). Norm conflicts and conditionals. *Psychological Review*, in press.

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

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, *37*(6), 1074–1106.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, *48*(3), 613–643.

Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking & Reasoning*, *7*(4), 367–390.

Sundholm, G., & van Atten, M. (2008). The proper explanation of intuitionistic logic: On Brouwer's demonstration of the bar theorem. In M. van Atten, P. Boldini, M. Bourdeau, & G. Heinzmann (Eds.), *One hundred years of intuitionism* (pp. 60–77). Basel: Birkhäuser.

Trpin, B., & Pellert, M. (2019). Inference to the best explanation in uncertain evidential situations. *British Journal for the Philosophy of Science*, in press.

van Atten, M. (2017). The development of intuitionistic logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/intuitionistic-logic-development/.

van Dalen, D. (2001). Intuitionistic logic. In L. Goble (Ed.), *The Blackwell guide to philosophical logic* (pp. 224–257). Oxford: Blackwell.

van Rooij, R., & Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vidal, M., & Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, *29*(6), 769–781.