The Dialogical Entailment Task

Niels Skovgaard-Olsen

University of Göttingen

Author Note

Niels Skovgaard-Olsen, Department of Cognition and Decision Making, University of Göttingen, Germany.

Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).

The supplemental materials including all data and analysis scripts are available at: https://osf.io/npc69/

Abstract

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

Keywords: Entailment Judgments, Relevance, Conditionals, Negations, Then, Probabilities.

The Dialogical Entailment Task[1]

The empirical measurement of accepted entailments has been the subject of some recent controversy in the psychology of reasoning. In an influential paper, Evans (2002) criticizes five decades of reasoning research for following a deductive paradigm that has investigated participants' reasoning competence with a particular type of task that many participants find unnatural, based on a normative model of correct reasoning derived from classical logic. More specifically, participants were usually asked to reason with abstract stimulus materials (e.g. letters and numbers) in tasks, where they were asked to assess logical arguments, or produce logically valid conclusions, with little or no instructions on how to understand central semantic notions like *validity*, *soundness*, and *logical necessity*. Moreover, even when more naturalistic stimulus materials were employed, the tasks still required participants without logical instruction to set aside their background knowledge and evaluate conclusions in light of premises that they were supposed to just assume to be true. Yet, this is a type of processing that participants find unnatural as shown by well-documented context effects and belief bias effects (Klauer, Musch, and Naumer, 2000). Furthermore, in this paradigm, participants were assessed based on interpretations of natural language words like *some*, *if*, and *not* from first-order logic, which is something that subsequent research has shown to be particularly problematic for natural language conditionals (Evans & Over, 2004).

One type of response in the so-called New Paradigm in the psychology of reasoning has been to adopt a probabilistic task format, where participants are required to indicate their responses in terms of degrees of belief and are permitted to use their background knowledge (Elqayam & Over, 2013). This shift has been instrumental in investigating knowledge-rich

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

inferences closer to commonsense in individual reasoning and in opening up new lines of investigation into argumentation and social reasoning (Oaksford and Chater, 2019).

The replacement of response format, however, also raises questions about how well participants' performance under the New Paradigm compares with the decades of data collected under the old Deduction Paradigm (Singmann & Klauer, 2011). Moreover, an often overlooked feature of the probabilistic representations of degrees of belief within psychology is that they also require basic logical properties like freedom from inconsistency and logical closure, which remain requirements of rational belief even within the New Paradigm (Skovgaard-Olsen, 2017a).[2] Probability theory can either be formulated in terms of set theory or in the language of propositional logic. Either way, there are certain logical properties that degrees of beliefs represented by probabilities must satisfy, like the ones listed below (Peterson, 2017, Ch. 6). Consequently, participants tested in the New Paradigm should still exhibit deductive competence to count as rational, Bayesian agents. For instance, they should still be able to assign probability 1 to logical consequences when reasoning with premises that have probability 1 (Oaksford and Chater, 2009). And, more generally, degrees of belief of rational Bayesian agents are constrained by the properties of logical truth, logical consequence, consistency, and logical equivalence as follows (Adams, 1998, p. 21-24):

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

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

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

P-validity

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

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

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

The empirical question of whether participants are then better able to reason based on p-validity is, however, not entirely clear. For instance, Evans et al. (2015) obtained mixed results when investigating the four inferences of the conditional inference task: MP (If A, C; A, therefore C), MT (If A, C; ¬C, therefore ¬A), DA (If A, C; ¬A, therefore ¬C), AC (If A,

C; C, therefore A). When examining chance-corrected hit-rate levels according to p-validity, Evans et al. (2015) only found a reliable above chance performance for the valid MP and the invalid AC inference; for the valid inference MT the hit rate was below chance levels. As the authors note: "participants did not conform to p-validity on the inferences that are actually valid, MP and MT. Indeed there was a small trend in the opposite direction" (p. 9). Similarly, Singmann et al. (2014) found that participants only conformed to p-validity for MP inferences and not for MT inferences. Moreover, when Cruz et al. (2017) stipulate the premise probability to be 100%, mean estimates for the conclusion of the valid inferences considered were around 85%-92%, in violation of the uncertainty sum-rule.

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

In the abovementioned psychological studies advocating p-validity, it is common to investigate only a handful of premise probabilities (e.g. by stipulating that the premise probability is 60%, 80%, and 100%) and measure the probability assigned to the conclusion of valid and invalid inferences. Since, however, this type of task does not address the

universal quantifier, and the modal content of p-valid inferences, it would be more accurate to say that what these studies investigate is first and foremost participants' *probabilistic coherence*, or whether their probability assignments are *in agreement* with the uncertainty sum-rule. In contrast, these studies do not directly investigate participants' *acceptance of entailments* in p-valid inferences—since for this, the experimental tasks would have had to be designed in a way that is suited for the modal content of p-valid inferences. To draw an analogy: from a handful of (or even many) truth-value assignments to the premises and conclusions of MP inferences, one has not shown that participants accept *the entailment* from the premises to the conclusion. For this, one would have to show that participants accept that the conclusion *cannot* fail to be true, once the premises are true.

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

The Dialogical Entailment Task

For the reasons indicated above, the present paper seeks to present a more natural, dialogical paradigm for eliciting participants' acceptance of entailments.[3] The inspiration comes from various sources. First, from the observation that classical logic is best viewed as a competence model for adversarial reasoning when we attempt to disprove the arguments of our interlocutors (Stenning and van Lambalgen, 2008). Second, the idea is motivated by the observation that attributions of consequential commitments in argumentative contexts provide a natural setting for assessing participants' grasp of the logical consequences of their beliefs (Skovgaard-Olsen, 2017a). Finally, it is informed by linguistic work on empirical evidence for semantic theories (Tonhauser and Matthewson, 2015).

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

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

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

require participants to have a sophisticated grasp of semantic notions like soundness, validity, or logical necessity (Tonhauser and Matthewson, 2015).

Earlier studies have examined which inferences participants draw in dialogical settings (Stevenson and Over, 1995; Thompson and Byrne, 2002) and investigated their degree of belief in the conclusion of informal reasoning fallacies as well as their acceptance of such arguments (Oaksford and Hahn, 2004; Hahn and Oaksford, 2007). These studies were, however, not designed to elicit participants' entailment judgments (as opposed to their acceptance of other types of inferences like, say, inductive inferences or implicatures). In fact, much of the research on argumentation within the New Paradigm has been conducted with the explicit goal of showing how everyday informal arguments that have been set aside by classical logic can nevertheless be captured by rational Bayesian reconstructions (Hahn, Harris, and Oaksford, 2012). In contrast, in Eva & Hartmann (2018) it is argued that even on a Bayesian approach to argumentation, an interest should be taken in valid arguments. The reason they give is that valid arguments have the property of ensuring that increases to the probability of one of the premises will *guarantee* that the probability of the conclusion increases. This points in the same direction as Adams' (1975) work on p-validity reviewed above, but is shown to hold in a much more general framework based on minimizing the Kullback-Leibler distance between the prior and posterior probability distributions.[4] This goes to show that even within the New Paradigm there is a need to investigate participants' acceptance of entailments in argumentative contexts.

Entailment judgments

The following principles are much discussed in conditional logics:

[4] However, it should be noted that Eva & Hartmann's (2018) argument is based on conjectures generalizing from examining inferences like MP, MT, AC, and DA without presenting a proof for the general case. It is also unclear how far their conclusions generalize to other frameworks. For instance, Kleiter (2018) finds that while MT is p-valid it is not *n-increasing*, in the sense that if the probability of any of the *n* premises increases, the probability of the conclusion also increases.

The Negation Principle	¬(if A, C) ⇔ if A, ¬C
Conditional Excluded Middle	(if A, C) ∨ (if A, ¬C)

As Adams (1998) says:

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

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

This principle moreover follows on general grounds connecting conditionals, subjective probability, and betting that have been influential in the New Paradigm based on work by de Finetti and Ramsey (Baratgin, Over, & Politzer 2013; Baratgin, Politzer, Over, & Takahaschi, 2018). On such accounts, the indicative conditional is explicated by the de Finetti truth table, which assigns conditionals the value 'True' in the ⊤⊤ cell, 'False' in the ⊤⊥ cell, and 'void' in the false antecedent cells.[5] This assignment is in turn motivated by a betting analysis, according to which a conditional bet on "if A, C" is won if "A & C" turns out to be the case, lost if "A & ¬C" turns out to be the case and rendered void if "¬A" is the case. Since bets on [¬(if A, C)] and [if A, ¬C] have the same pattern of wins and losses, the Negation Principle follows for probabilistic accounts of conditionals that are based on these principles.

Concerning the Principle of Conditional Excluded Middle, there is a famous dispute between Lewis (1973) and Stalnaker (1980) about whether to accept it for subjunctive conditionals (e.g. 'If Oswald hadn't killed Kennedy, someone else would have'). Yet, Bacon

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

(2015, 2019) argues that the status of the Principle of Conditional Excluded Middle is much less controversial for indicative conditionals (e.g. 'If Oswald didn't kill Kennedy, someone else did') than for subjunctive conditionals. Both the Negation Principle and the Principle of Conditional Excluded Middle require the following inference to be valid (where '⊨' indicates semantic consequence):

Target Inference: ¬(if A, C) ⊨ if A, ¬C

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

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

Agree Baseline: if A, ¬C ⊨ ¬(if A, C)

Disagree Baseline: if A, C ⊨ if A, ¬C

The idea behind the use of these baselines is to have two inferences which most theories will treat as valid,[6] and invalid respectively, as a manipulation check for the Dialogical Entailment Task (described in further details below). The test then consists in assessing whether participants' performance concerning Target Inference is more like their performance with respect to the Agree or the Disagree Baseline.

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

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

The Relevance Effect

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

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

On the other hand, the Ramsey test has inspired the probabilistic semantics of indicative conditionals of Adams (1975), which in its original form denies that indicative conditionals have truth conditions, and subscribes to either P(if A, C) = P(C|A) or acc(if A, C) = acc(C|A), for 'if A, C' referring to simple conditionals (which exclude nestings of

conditionals). Here ‘acc(if A, C)’ stands for the acceptability of the conditional. Often this version of Adams’ thesis is preferred, because it is unclear whether P(if A, C) can still be interpreted as a probability in light of the so-called triviality results, which supply a reduction of the most obvious way of implementing this thesis (Bradley, 2007; Douven, 2015). Through the influence of the writings of Edgington (1995) and Bennett (2003), the psychological hypothesis that the probability of indicative conditionals is evaluated as the conditional probability, P(C|A), found its way into the psychological literature (Evans and Over, 2004), where it goes by the name “the Equation”.

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

It is sometimes suggested that the Relevance Effect should be interpreted in terms of causal readings of conditionals (e.g. van Rooij and Schulz, 2018; Oaksford and Chater, 2019), given that ΔP makes up the numerator in causal power (Cheng, 1997). But it is also possible

to consider causal relations as a specific instance of a more generic reason relation (Spohn, 2012), which then turns the Relevance Effect into a finding concerning the relationship between conditionals, reasons, and arguments. Possible explanations for the Relevance Effect are diverse and have been explored in several recent publications (Cruz, Over, Oaksford & Baratgin, 2016; Krzyżanowska, Collins, & Hahn, 2017; Skovgaard-Olsen, Collins, Krzyżanowska, Hahn, & Klauer, 2019a). In this paper, the goal is to investigate whether relevance effects extend to participants' reasoning with conditionals containing negation operators, in their probability assignments and entailment judgments. Experiment 1 starts out by applying the Dialogical Entailment Task to the three types of inferences introduced above.

Experiment 1

Methods

Participants

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

Design

The Experiment implemented a within-subjects design. Three factors were individually varied: Relevance (Positive Relevance vs. Irrelevance), Priors (HH, HL, LH, LL,

meaning, for example, that P(A) = low and P(C) = high for LH) and Inference Type. The Inference type factor had three levels: Agree Baseline, Disagree Baseline, and Target Inference (repeated below). Each participant thus completed 24 within-subject conditions in total.

Materials and Procedure

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

The participants were given the following task instructions:

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

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

The 12 scenarios used in this study were taken from Skovgaard-Olsen et al. (2016b). These scenarios were found to reliably induce assumptions about relevance and prior probabilities of the antecedent and the consequent in previous studies that implement our

experimental conditions. Table 1 displays sample items for the Mark scenario for Positive Relevance ($\Delta p > 0$), and Irrelevance ($\Delta p = 0$), for $\Delta p = P(C \mid A) - P(C \mid \bar{A})$.

Table 1. Stimulus Materials, Mark Scenario

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

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

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

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

The participants saw two control items and a practice item before the actual experiment started, where it was emphasized that attention was needed to notice subtle differences between the wordings (e.g. use of 'not', 'false', 'wrong', 'correct', and 'if') of the

various sentences presented in the experiment. For the control items, Samuel would either assert “Some of the employees are invited to the party” and deny that “not all of the employees were invited” (i.e. consistently deny a scalar implicature), or assert that “John is a bachelor” and deny that “John is unmarried” (i.e. inconsistently denying an analytical consequence of his first assertion).

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

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

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

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

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

Disagree Baseline: if A, C $\models$ if A, $\neg$C

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

In Table 2, Samuel's assertions with respect to these inferences are illustrated using the stimulus materials from Table 1 (however, without 'then' and 'will' in the consequents):[7]

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

Table 2. The Dialogical Entailment Task

Scenario
Mark has just arrived home from work and there will shortly be a great movie on television, which he has been looking forward to. Mark is quite excited because he recently bought a new TV with a large screen. He has a longing for popcorn, but his wife has probably eaten the last they had while he was gone.

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

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

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

Results

Control Items. The degree to which participants agreed with accusing Samuel of an inconsistency was found to be significantly higher in the entailment control item (*Mdn* = 4.00)

than in the scalar implicature control item (*Mdn* = 2.00), V = 86, $p < .01$, $r = -.29$, for the Wilcoxon signed-rank test. The experimental task was thereby found to pass a first manipulation check.

Entailment Judgments. To examine ratings of entailment for the three types of inferences, we relied on a set of mixed generalized linear models, which represent the acceptance of an entailment (a binary variable formed by answering "Agree" or "Strongly agree" to Louis' objection to Samuel) by a binominal likelihood function together with a logit link function. The models had crossed random effects for intercepts and slopes by participants and by items (Baayen, Davidson, and Bates, 2008) to control for the effect of replicates for each participant and item in the experimental design. The models were fitted in a Bayesian framework using the R-package `brms` (Bürkner, 2017) with weakly informative priors and featured the following predictors:

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

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

Table 3. Model Comparison

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

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

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

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

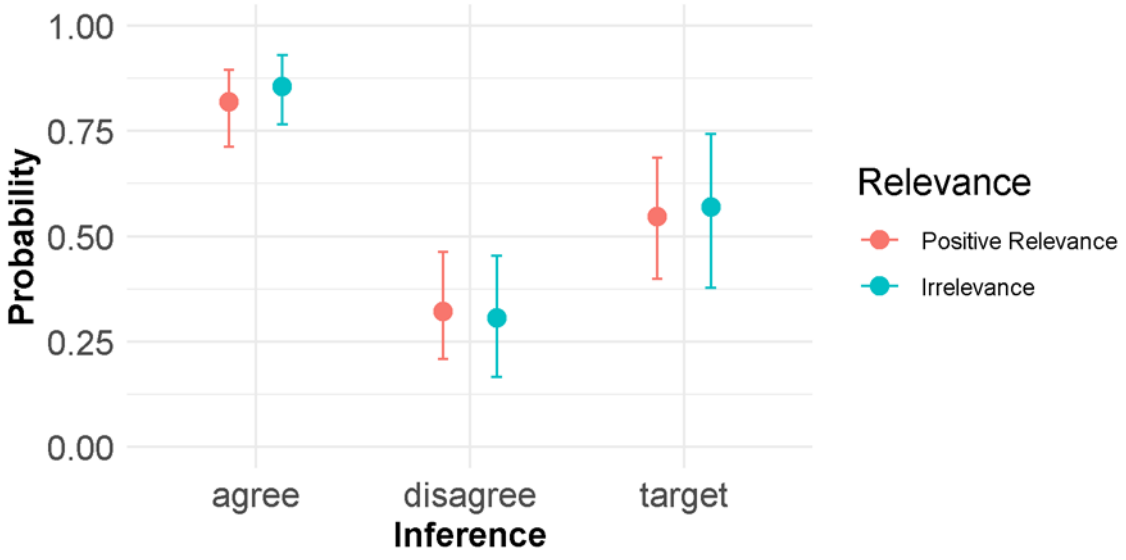


Figure 1. Weighted posterior predictive probability of acceptance of entailment. 'agree' = baseline for agreement; 'disagree' = baseline for disagreement; 'target' = inference to be compared with the baselines. 'Probability' on the y-axis indicates posterior probability of accepting the entailment for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

Discussion

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

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

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

Experiment 2

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

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

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

Session 1: Negations, Then, and Probabilities

On the Meaning Contribution of 'Then'

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

If John is dead or alive, (#then) Bill will find him.

Even if John is drunk, (#then) Bill will vote for him.

If I were the richest linguist on earth, (#then) I (still) wouldn't be able to afford this house.

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

In line with this, Biezma (2014) puts forward a general theory on the non-truth functional meaning of 'then'. The central claim is that 'then' operates at the level of discourse

structures by establishing an anaphoric relation between two discourse moves. As part of its felicity conditions, it is claimed that non-temporal uses of 'then' require that two propositions enter into a causal explanatory relationship, whereby the antecedent proposition provides a reason for the consequent proposition. In paraphrase, when 'then C' occurs alone, the meaning conveyed is 'C because of A', where A may remain an implicit part of the antecedent discourse.

One of the central advantages of the theories reviewed above is that apparently the Relevance Effect of conditionals reported in Skovgaard-Olsen et al. (2016a) can be explained by pointing to the occurrence of 'then' in the investigated stimulus materials ('if A, *then* C').[8] This in turn would allow us to adopt the Lewis (1973), Stalnaker (1968), and Kratzer (1986) framework to provide a semantics for 'If A, C' while predicting the influence of reason relations on the evaluation of the felicity conditions of 'if A, *then* C', which in turn should affect probability and acceptability evaluations. On this view, 'if A, C' merely provides a description of the worlds in the context set (to wit, that in the most similar A-worlds to the actual world, C is also true), whereas 'if A, *then* C' establishes a causal, explanatory claim whereby the antecedent provides causal information about the consequent.

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

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

---

8 I thank María Biezma, Maribel Romero, and Eva Csipak for discussion.

The Negation Task

As a test of whether Iatridou (1994), von Fintel (1994), and Biezma's (2014) theories are able to account for the Relevance Effect, the Negation Task from Skovgaard-Olsen et al. (2019a) was selected. In this task, participants are asked to assign probabilities to the following conditionals across manipulations of the antecedent's relevance for the consequent (see below):

AFFIRMATIVE CONDITIONAL: if A, C

WIDE-SCOPE NEGATION: ¬(if A, C)

NARROW-SCOPE NEGATION: if A, ¬C

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

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

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

THE NEGATION PRINCIPLE: $\neg(\text{if A, C}) \Leftrightarrow \text{if A, } \neg\text{C}$

Probabilistic version: $P(\neg(\text{if A, C})) = P(\text{if A, } \neg\text{C})$

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

Methods

Participants

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

Design

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

Materials and Procedure

First, participants were given a brief general introduction:

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

They were then presented with four practice items in random order. As practice items, participants were asked to assign a probability on a scale from 0 to 100% to a categorical sentence with an existential presupposition failure (e.g. “The queen of the USA is in her mid-thirties”, which falsely presupposes that there is a queen of the USA) and its wide and narrow scope negations. After this, participants were instructed to pay attention to subtle differences

in the wording of the sentences used for the rest of the experiment, such as whether they contain words like 'not', 'false', and 'if'.

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

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

Results

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

- Model M4 modelled the ratings as a function of the DV factor, encoding the three different types of conditionals (Affirm [if A, C], Wide [¬(if A, C)], Narrow [if A, ¬C)], and the Relevance factor, encoding the two relevance levels. The model also included the interaction of these two factors.

- Model M5 built upon M4 but did not include the two-way interaction.
- Model M6 built on M5 but did not include the Relevance factor.

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

Table 4. Model Comparison

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

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

The information criteria in Table 4 display a clear preference for M4. Consistent with this, very strong evidence for a main effect of Relevance (b_{IR} = -1.17, 95%-CI [-1.41, -0.94], BF_{H0H1} = -6.05 * 10^{-59}), the DV factor (b_{Wide} = -1.16, 95%-CI [-1.39, -0.92], BF_{H0H1} = -8.78 * 10^{-154}; b_{Narrow} = -1.10, 95%-CI [-1.34, -0.87], BF_{H0H1} = 2.97 * 10^{-22}), and the two-way interaction ($b_{IR:Wide}$ = 1.77, 95%-CI [1.41, 2.14], BF_{H0H1} = 2.51 * 10^{-16}; $b_{IR:Narrow}$ = 0.96, 95%-CI [0.65, 1.27], BF_{H0H1} = 4.04 * 10^{-15}) were found. The interaction is illustrated in Figure 2 with the characteristic cross-over of the lines representing Positive Relevance and Irrelevance, which makes the wide-scope negated conditionals the highest rated for the Irrelevance condition.

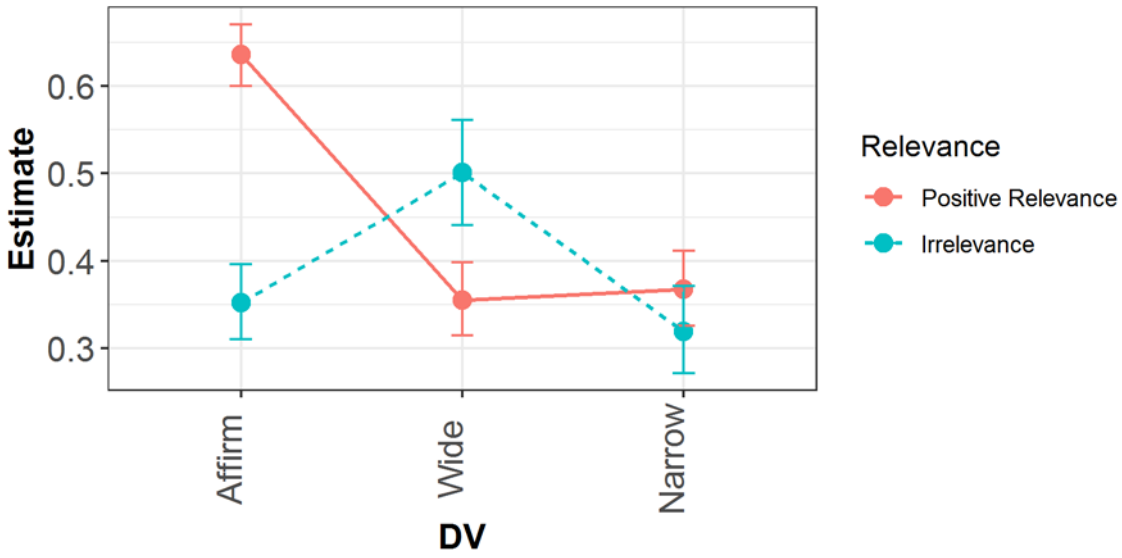


Figure 2. Posterior mean estimates of M4.

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

Discussion

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

differences between the present data set and the dataset reported in Skovgaard-Olsen et al. (2019a, Experiment 2). Participants thus appear to treat the difference between 'if A, C' and 'if A, *then* C' to be little more than a stylistic difference when assigning probabilities to [if A, C], [if A, ¬C], and [¬(if, A, C)] across relevance levels. This in turn agrees with the notion in Geis and Lycan (1993) that genuine conditionals can take the proform 'then' in their consequents without change in meaning, in contrast to superficially similar constructions that are not conditional in meaning, like so-called biscuit conditionals:[9]

If you want any, there are biscuits on the sideboard

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

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

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

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

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

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

...but if she is sleepy.

...since hunger keeps her quit.

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

One way of interpreting the interaction between the Relevance factor and the negation operator for probability assignments, which was raised by one of the reviewers, is that Positive Relevance and Irrelevance brings out this ambiguity in the scope of the negation operator and that Irrelevance forces the wide-scope interpretation (in which both 'if A, then C' and 'if A, then ¬C' are rejected) whereas Positive Relevance typically goes along with the narrow-scope interpretation (according to which 'if A, then not-C' and ‘¬(if A, then C)' are

equivalent). Further research will have to determine the merits of this interpretation. So far, possible scope ambiguities like this are an underexplored topic in the psychology of reasoning. However, their importance has recently been stressed by Over, Douven, and Verbrugge (2013).

Session 2: Negations and Entailments

The Dialogical Entailment Task

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

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

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

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

Table 5. Applying the Cross-Task Consistency Constraint

	Inference		License
Agree Baseline	if A, $\neg$C $\vDash$ $\neg$(if A, C)	only if	P($\neg$(if A, C)) $\geq$ P(if A, $\neg$C)
Disagree Baseline	if A, C $\vDash$ if A, $\neg$C	only if	P(if A, $\neg$C) $\geq$ P(if A, C)
Target	$\neg$(if A, C) $\vDash$ if A, $\neg$C	only if	P(if A, $\neg$C) $\geq$ P($\neg$(if A, C))

Based on the results from Session 1, it is very clear that the participants have acquired a license to accept the Agree Baseline inferences and that the participants do not have a license to accept the Disagree Baseline inferences. Matters are, however, less clear when it comes to the Target inference. The reason is the interaction with Relevance and the negation operator that was found, which lead to violations of the probabilistic version of the Negation

Principle for the Irrelevance condition. More specifically, in Session 1 it was found for the Positive Relevance condition that $P(\neg(\text{if } A, C)) \approx P(\text{if } A, \neg C)$. Yet, for the Irrelevance condition it was found that $P(\neg(\text{if } A, C)) > P(\text{if } A, \neg C)$, at the group level. According to Table 5, the participants are in other words only licensed to accept the Target Inference for the Positive Relevance condition. If, however, participants accept the Target Inference for the Irrelevance condition, then it would lead to violations of the above cross-task consistency constraint that ensures that probabilistic reasoning is consistent with deductive logic (Joyce, 2004; Oaksford, 2014). A central purpose of Session 2 is to investigate whether participants violate this cross-task consistency constraint for the Target Inferences.

Method

Participants

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

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

Design

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

Materials and Procedure

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

Results

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

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

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

Table 6. Model Comparison

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

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

The information criteria displayed in Table 6 favour M7 indicating that there was an interaction making the Target Inferences slightly higher rated in the Positive Relevance condition than in the Irrelevance condition (b = 0.96, 95%-CI [0.26, 1.66], BF_{H0H1} = 0.25). But no main effect of Relevance could be found (b = -0.25, 95%-CI [-0.78, 0.28], BF_{H0H1} = 7.25). In contrast, very strong evidence in favour of a main effect of the Inference factor could be obtained. Both the posterior probabilities of accepting the entailment in the Disagree Baseline (b = -2.80, 95%-CI [-3.63, -2.02], $BF_{H0H1} = 5.73 * 10^{-43}$) and for the Target Inference (b =
-2.20, 95%-CI [-2.86, -1.59], $BF_{H0H1} = 5.96 * 10^{-19}$) were substantially below the posterior probability of accepting the entailment in the Agree Baseline, as displayed in Figure 3.

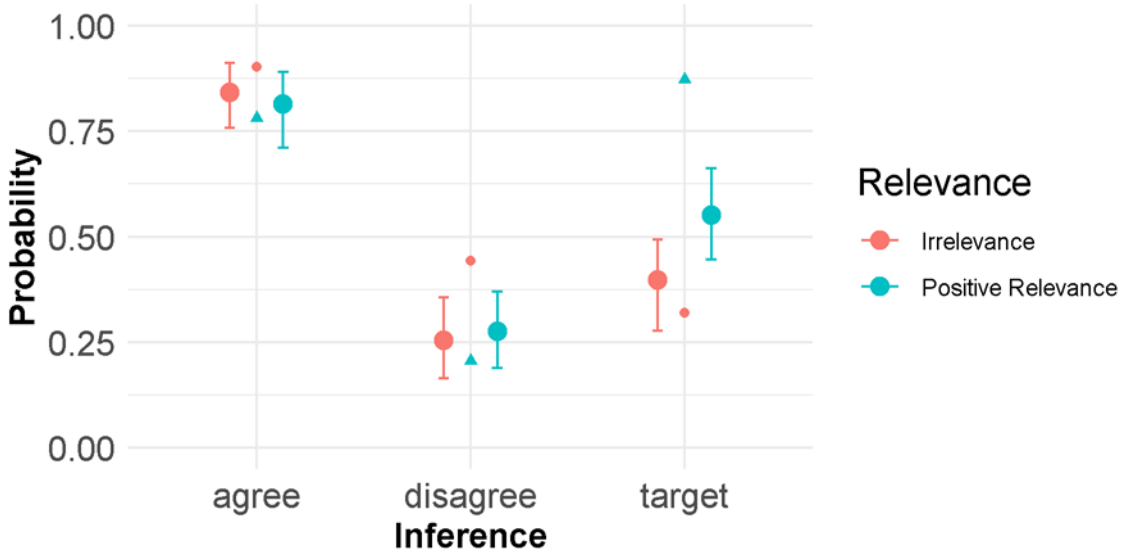


Figure 3. Weighted posterior predictive probability of acceptance of entailment. 'agree' = baseline for agreement; 'disagree' = baseline for disagreement; 'target' = inference to be compared with the baselines. The large dots indicate the posterior probability of accepting the entailment for a given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

As outlined in Appendix 1B, a Bayesian mixture model was applied to identify latent classes for whether the participants possessed a license to accept the entailments in session 2 based on the cross-task consistency constraint in Table 5 and their performance in Session 1. Figure 3 displays the predicted acceptance of entailment based on the assignments of latent classes of inference licenses in session 1 as little dots and triangles. The prediction assumes that P(*acceptance of entailment*) = 1 – P(*missing license*). It was found that the central tendency in the posterior probability of acceptance of entailment in session 2 was highly correlated with the predicted acceptance based on the majority assignment of latent classes in session 1, $r = 0.84$, $t(4) = 3.13$, $p = 0.035$.[10] The main exception was the unused license for accepting the Target Inference in the Positive Relevance condition. Here the majority response (n = 33) would predict an 87% posterior probability of acceptance of the entailment in session 2 (see Figure 3). In contrast, the participants' actual responses were more in line

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

with the minority response (n = 22) of having a posterior probability of 46% of acceptance of the entailment in this condition.

Discussion

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

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

As shown in Figure 3, the participants' conformity to the Negation Principle for positive relevance items in session 1 of Experiment 2 gave them a license to accept the Target Inference in the Positive Relevance condition in session 2. Yet, the participants displayed a similar lack of preference with respect to the Target Inference in the Positive Relevance condition as in the Irrelevance condition. On closer inspection, however, it would have appeared problematic, if the participants had selectively exploited this license by disagreeing

with Louis' objection for the Target Inference in the Irrelevance condition while agreeing with Louis' objection for the Positive Relevance condition. Doing so would have required that the participants agreed that Samuel's statements *cannot* both be true when seeing one type of item while accepting that they *can* both be true, when seeing a different type of item. In the first case, participants would have had to accept that there are no models satisfying the premise and the negation of the conclusion while agreeing, in the second case, that there are such models.

Experiment 3

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

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

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

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

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

Faced with such a variety of different ways of parsing the sentences, Experiment 3 sought to fix the parsing of the sentences through Louis' objections. This time, Louis' objection interprets the wide-scope negations in Samuel's statements as categorical rejections

of conditional statements. Accordingly, Louis' objection to the Target Inference now takes the form of that Samuel cannot both *reject* "if A, C" and *reject* "if A, ¬C" at the same time.

Another side-effect of this reformulation is that whereas the original formulation of the task concerns the more traditional semantic question of whether the premise and the negation of the conclusion of an inference can be true at the same time, the reformulated version concerns rational acceptability/assertability and whether warrant to assert the premise precludes a warrant for denying the conclusion.

Preservation of rational acceptability from the premises to the conclusion has traditionally been associated with the pragmatics of making assertions. However, there have also been attempts to replace classical notions of logical consequence with more use-oriented notions of inference based on rational assertability/acceptability (Tennant, 2002; Khlentzos, 2004). E.g. in Yalcin (2012), a consequence relation is defined based on that no information state that accepts the premises can fail to accept the conclusion, to model epistemic content.

Method

Participants

A total of 124 people from USA, UK, Canada, and Australia took part in the Online study, which was run on Mechanical Turk. The same exclusion criteria were used as in Experiment 1. Since some of these criteria were overlapping, the final sample consisted of 87 participants. Mean age was 41 years, ranging from 23 to 71, 56% of the participants were female; 79% indicated that the highest level of education that they had completed was an undergraduate degree or above. Applying the exclusion criteria had only slight effects on the demographic variables.

Design

Experiment 3 had the same experimental design as Experiment 1. In total, the participants were thus presented with the same 24 within-subject conditions.

Materials and Procedure

Experiment 3 followed the procedure of Experiment 1 and used the same materials. The only differences were that 1) the participants were instructed that Louis accuses Samuel of making two claims that he cannot *assert* at the same time, 2) Louis' objections were replaced by the objections in Table 8, 3) the participants were cautioned not to conflate agreeing/disagreeing with Samuel's statements and with Louis' objections, and 4) that the participants read Samuel's assertions on a separate page before processing Louis' objections. When presenting Louis' objections, Samuel's statements and the scenario texts were displayed as reminders in grey at the top of the page.

Table 8. Louis' Acceptability Objections

Target Inference	Disagree Baseline	Agree Baseline
Samuel: It is FALSE that IF A, C …But it would be WRONG to think that if A, not-C	Samuel: IF A, C …But it would be WRONG to think that IF A, not-C	Samuel: IF A, not-C …But it would be CORRECT to think that IF A, C
Louis: Wait, you cannot both **reject** that: "IF A, C" and **reject**: "IF A, not-C" at the same time!	Louis: Wait, you cannot both **accept** that: "IF A, C" and **reject**: "if A, not-C" at the same time!	Louis: Wait, you cannot both **accept** that: "IF A, not-C" and **accept**: "if A, C" at the same time!

Note. In the experiment, the words "accept" and "reject", which are marked in bold here, were made salient through a blue color to the participants. Here the structure of the objections is illustrated; in the actual experiment the propositional letters A and C were filled out with the same naturalistic scenarios as in Experiment 1.

Results

The same type of analysis was applied as in Experiment 1 with the following models:

- Model M10 modelled acceptance of entailment as a function of the Inference factor (Agree vs. Disagree vs. Target), the Relevance factor (Positive Relevance vs. Irrelevance), and their interaction.
- Model M10 built upon M11 but did not include the two-way interaction.
- Model M12 built on M11 but did not include the Relevance Factor.

Table 9 reports the performance of these models as quantified by WAIC and LOOIC.

Table 9. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M10	2130.94	0	--	2122.3	0.452
M11	2132.63	1.70	3.72	2124.2	0.194
M12	2131.42	0.49	3.76	2123.1	0.354

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

As the information criteria suggest, the full model, M10, was the winning model, but the edge given to this model was very slight as witnessed by the intermediary Akaike weights given to all models. In line with this, no main effect of relevance could be found (b = 0.09, 95%-CI [-0.43, 0.61], BF_{H0H1} = 10.5), and the Relevance factor was also not involved in an interaction ($b_{Disagree:Irrelevance}$ = 0.13, 95%-CI [-0.50, 0.74], BF_{H0H1} = 8.42; $b_{Target:Irrelevance}$ = -0.46, 95%-CI [-1.09, 0.16], BF_{H0H1} = 3.2). Like in the previous studies, strong evidence could be obtained that the posterior probability of accepting the entailment in the Disagree Baseline was below the Agree Baseline (b = -2.54, 95%-CI [-3.28, -1.84], $BF_{H0H1} = -2.36 * 10^{-23}$). In contrast, there was now only anecdotal evidence for a difference between the Target Inference and the Agree Baseline (b = -0.57, 95%-CI [-1.06, -0.07], BF_{H0H1} = 0.92). These findings are illustrated in Figure 5, which displays the weighted predictive posterior probabilities of all three models, when collapsing across the Relevance factor.

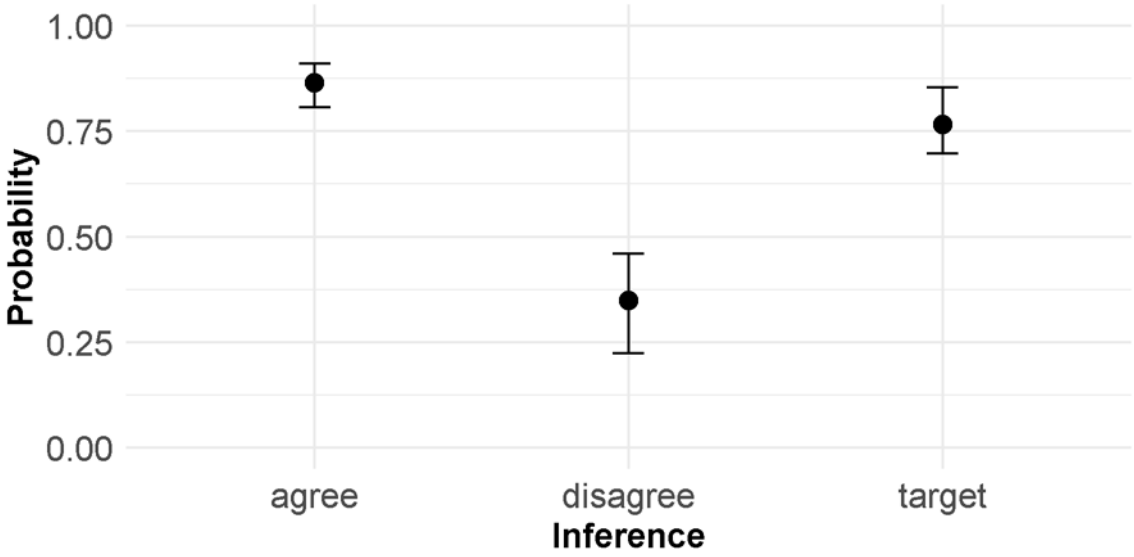


Figure 5. Weighted posterior predictive probability of acceptance of entailment. 'agree' = baseline for agreement; 'disagree' = baseline for disagreement; 'target' = inference to be compared with the baselines. The posterior probabilities of M10, M11, and M12, were weighted by the Akaike Weights

from Table 9 and collapsed across the Relevance factor to produce this plot.

Discussion

It is striking that only anecdotal evidence could be obtained for a difference between the Target Inference and the Agree Baseline in Experiment 3. This indicates that participants accept the entailment of the Target Inference when Louis' objection is presented as in Experiment 3. Experiment 3 thereby documents a *facilitation effect* compared to Experiments 1 and 2, where only a lack of preference with respect to the Target Inference could be found (with posterior probability of accepting the entailment around 50%). Apparently, fixing the parsing of the negation operator (as a wide-scope rejection of the whole statement), and changing the task to judging preservation of rational acceptability, has the effect of rendering the Target Inference acceptable to the participants.

General Discussion

In this paper, evidence was found against an unrestricted adoption of the Negation Principle both in its probabilistic version – with and without 'then' and 'will' in the examined conditionals – as well as against its truth-conditional version in an entailment task.

THE NEGATION PRINCIPLE: $\neg$(if A, then C) $\Leftrightarrow$ if A, then $\neg$C

Probabilistic version: P($\neg$(if A, then C)) = P(if A, then $\neg$C)

This principle has, however, played a prominent role in the psychological literature, where it has been cited by proponents of the Suppositional Theory of conditionals as a litmus test of their theory (Handley et al, 2006). In addition, the principle has played a role in the possible-worlds account of conditionals that is popular in linguistics (Stalnaker, 2011). The Negation Principle is moreover accepted by Adams (1998, p. 270) and certain three-valued logics of conditionals in philosophy (e.g. Cantwell, 2008a). Moreover, it follows from accounts emphasizing the connections between conditionals, subjective probability, and

conditional bets based on de Finetti truth tables (Baratgin, Over, & Politzer 2013; Baratgin, Politzer, Over, & Takahaschi, 2018).

The probabilistic version of the Negation Principle is violated due to an interaction between the reason relation of indicative conditionals and the negation operator, which strongly affect their probabilities. The evidence suggests that participants only conform to this principle for Positive Relevance conditions; for Irrelevance it is systematically violated.

The significance of the violation of the Negation Principle for its probabilistic version both with and without 'then' and 'will' is that it rules out an explanation of the Relevance Effect in Skovgaard-Olsen et al. (2016a) as based on the non-truth conditional contribution of the discourse marker 'then', along the lines of Iatridou (1994), von Fintel (1994), and Biezma's (2014). Had the Relevance Effect been due to the influence of 'then' in the investigated materials, we would expect the effects on probability evaluations of the contrast Positive Relevance ($\Delta P > 0$) vs. Irrelevance ($\Delta P = 0$) to go away once conditionals without 'then' in the consequent were investigated. But this turned out not to be the case; in fact, it was found that the results on the Negation task in Skovgaard-Olsen et al. (2019a) could be exactly replicated without the occurrence of 'then' (and 'will') in the consequent (see Appendix 1A). This suggests that as far as probabilistic relevance effects are concerned, there is no difference between 'if A, C' and 'if A, then C will be the case'.

In Experiments 1 and 2, it was found that the participants do not have strong preferences concerning the Target Inference [$\neg$(if A, C) $\vDash$ if A, $\neg$C], which is also required by the Principle of Conditional Excluded Middle (CEM).

Conditional Excluded Middle (if A, C) $\vee$ (if A, $\neg$C)

In a famous dispute between Lewis (1973) and Stalnaker (1980), Stalnaker defended this principle while making the concession that in practice issues of vagueness introduce ties in which possible worlds are most similar to the actual world. As a result, situations may arise where neither [if A, C] nor [if A, $\neg$C] can be treated as true for practical purposes, although

the inference principle of Conditional Excluded Middle continues to remain valid on the idealized theory. Bacon (2015, 2019) argues that while there is a dispute among Stalnaker and Lewis about the status of the principle of Conditional Excluded Middle for subjunctive conditionals, the principle is self-evident for indicative conditionals. Indeed, Bacon (2019, p. 20) proposes to treat the validity of the principle of Conditional Excluded Middle as: "a piece of data that any account of indicatives ought to be able to accommodate, not a controversial principle like its subjunctive cousin".[11]

In contrast, Khemlani, Orenes, and Johnson-Laird (2014) hold that [if A, then C] and [if A, then ¬C] make contrary but not contradictory assertions, because it is possible for both of them to be false. Interestingly, the data in Experiments 1 and 2 suggest that the participants do not treat the Target Inference [¬(if A, C) ⊨ if A, ¬C] as a valid entailment in relation to indicative conditionals. Yet, both the Negation Principle and the principle of Conditional Excluded Middle require the Target Inference to be valid.

At the same time, a facilitation effect was found in Experiment 3 indicating that the participants *do* accept the Target Inference when the parsing of the negation operator is fixed (as a wide-scope rejection of the whole statement) and the task is changed to judging preservation of rational acceptability, instead of preservation of truth. The implication appears to be that while our results are not supportive of the entailment of the Target Inference when validity is judged by classical logic, the Target Inference would fare better on consequence relations based on preservation of acceptance, like the one expounded in Yalcin (2012).

Grice (1989, p. 80-83) discusses the possibility of using a denial of conditional as a refusal to assert the conditional in question, but not because it does not represent the facts. To

[11] Part of Bacon's (2019) theoretical argument for the Principle of Conditional Excluded Middle for indicative conditionals relies on Adams' thesis (P(if A, C) = P(C|A) for simple conditionals). However, Adams' thesis has already been shown to break down for missing-link conditionals in Skovgaard-Olsen et al. (2016a), which is a result that the data from Session 1 (Experiment 2) replicated for bare indicative conditionals without 'then' and 'will' in the consequent.

illustrate: “to say “It is not the case that if X is given penicillin, he will get better” might be a way of suggesting that the drug might have no effect on X at all” (p. 81). Similarly, Adams (1998, p. 270) points out that “to assert “It is not the case that if φ, then ψ” *can* mean that “If φ, then ψ” isn’t probable enough to be asserted”. Accordingly, the fact that P(¬(if A, C)) received the highest value in the Irrelevance condition in Session 1 of Experiment 2 could be taken as an indicator that the participants treat both [if A, C] and [if A, ¬C] as unassertable. From this perspective, it is, however, strange that the participants would not permit Samuel to deny both [if A, C] and [if A, ¬C] in Experiment 3, where the facilitation effect was found. One possibility is that the participants were reacting to the oddity of why Samuel would connect unrelated sentences such as “Mark is wearing socks” and “Mark’s TV is working” out of the blue in sentences, if *he did not presuppose* that they were supposed to be connected. In retrospect, it might have been better to let a neutral interlocutor assert the missing-link conditionals, and have Samuel react to these assertions by denials, instead of making Samuel the originator of the missing-link items. Future research will have to determine whether the facilitation effect is robust with respect to such variations.

Finally, the participants’ cross-task consistency was examined in Experiment 2 by investigating whether the participants accepted entailments for which they had no license based on their probability assignments one week earlier. It was found that this was not the case, but that the participants did have an unused license to endorse the Target Inference for the Positive Relevance condition. On closer inspection, it was found, however, that by using this license, participants would have had to adopt the doubtful cognitive state of, on the one hand, accepting that there are no models satisfying the premise and the negation of the conclusion (when responding to the positive relevance items) while agreeing, in the second case, that there are such models (when responding to irrelevance items).

A further contribution of the present paper consists in the introduction of a novel experimental task for investigating participants’ acceptance of entailments, which avoids the

pitfalls of previous research into deductive reasoning identified in Evans (2002). In line with work by Stenning and van Lambalgen (2008) on deductive logic being most suited for adversial contexts, and with work on the argumentative nature of logical norms for rational beliefs in Skovgaard-Olsen (2017a), the Dialogical Entailment Task proposes to investigate participants' acceptance of entailments in argumentative contexts.

In this paper, the Dialogical Entailment Task was put to use to investigate the participants' acceptance of a Target Inference [¬(if A, C) ⊨ if A, ¬C] across relevance levels. While relevance did play a role on some of the open-ended responses in Experiment 1 of why the participants had agreed/disagreed with Louis (which were used here only for exploratory purposes), in general strong effects of relevance were not found in the entailment task (as opposed to the probabilistic Negation Task). Similarly, no relevance effects on the examined and-to-if entailment judgments were found in Skovgaard-Olsen et al. (2019b), echoing the lack of relevance effects for truth-value judgments in Skovgaard-Olsen et al. (2017).

There is room for improvements of the Dialogical Entailment Task in future studies. One obvious way of improving it would be to elicit the counterexamples produced by participants who do not accept a given inference principle. Furthermore, alternative entailment relations to the classical notion of logical validity could be tested. In Experiment 3 one such variant was investigated (i.e. preservation of rational acceptability), but many further kinds exist. For instance, versions of the Dialogical Entailment Task implementing p-validity could be investigated (e.g. by having Samuel assign high probabilities to the premises of an inference and a low probability to its conclusion). Furthermore, Cantwell (2008b) recommends using preservation of non-falsity as a notion of validity for three-valued logic. Finally, Chemla, Egré, and Spector (2017) and Chemla and Egré (2018) have investigated an even more general family of entailment relations for many-valued logics by, inter alia, exploiting the possibility of exhaustively investigating all possible truth tables through computer-aided search.

These developments indicate the importance of extending the Dialogical Entailment Task to further types of entailment relations, in particularly when three-valued truth tables of indicative conditionals are investigated, such as in Baratgin et al. (2018).

Conclusion

Given that intuitive entailment judgments arguably make up one of the primary sources of data for semantic theories, it would be desirable to have a substantive body of empirical data surveying the entailment judgments of ordinary people. In this paper, a novel Dialogical Entailment Task was developed to obtain data of participants' intuitive entailment judgments in the aftermath of the methodological criticism in Evans (2002) of a previous deductive paradigm in the psychology of reasoning.

Combining this task with participants probability assignments across relevance conditions, evidence was reported against the Negation Principle [¬(if A, then C) ⇔ if A, then ¬C] both in its probabilistic version – with and without 'then' and 'will' – as well against its truth-conditional version. In its probabilistic version, it was found that the Negation Principle was only conformed to for positive relevance items; for irrelevance items it was systematically violated. As an inference principle concerning truth-preservation from the premises to the conclusion, it was found that the participants did not have strong preferences in either direction (Experiments 1 and 2). Yet, when the entailment task was posed using preservation of rational acceptability, while disambiguating potential scope ambiguities, a facilitation effect was found (Experiment 3).

The Relevance Effect reported in Skovgaard-Olsen et al. (2016a) was found using indicative conditionals containing neither 'then' and 'will' in the consequent as stimulus materials. Consequently, it is possible that these results could be completely accounted for based on the meaning contribution of 'then' advanced in Iatridou (1994), von Fintel (1994), and Biezma (2014). Against such an account, it was found that the strong interaction for

probability evaluations between relevance and the negation operator reported in Skovgaard-Olsen et al. (2019a) could be completely replicated using indicative conditionals without 'then' (and 'will') in the consequents. We can therefore conclude that it is not the presence of 'then' in the investigated stimulus materials that is driving the Relevance Effect.

References

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

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

Arlo-Costa, Horacio (2007). The Logic of Conditionals. In E. N. Zalta (eds.), *The Stanford Encyclopedia of Philosophy* (spring 2016 Edition). Retrieved from: <http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/>.

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

Bacon, A. J. (2015). Stalnaker's thesis in context. *The Review of Symbolic Logic*, *8*(1), 131-163.

Bacon, A. J. (2019). On the Semantics of Indicatives. Ms, University of Southern California. Retrieved from: http://yalcin.work/workshop

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking & Reasoning*, *19*, 308-328.

Baratgin, J., Politzer, G., Over, D. E., and Takahashi, T. (2018). The Psychology of Uncertainty and Three-Valued Truth Tables. *Frontiers in Psychology*, *9* (1479), 1-17.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 159-219.

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

Bhatt, R. and Pancheva, P. (2006). Conditionals. In Everaert, M. and van Riemsdijk, H. (Eds.), *The Blackwell companion to syntax* 1 (pp. 638–687). Oxford: Blackwell.

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

Bradley, R. (2007). A Defence of the Ramsey Test. *Mind*, *116*(461), 1-21.

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

Bürkner, P., & Vuorre, M. (2018, February 28). Ordinal Regression Models in Psychological Research: A Tutorial. Retrieved from http://doi.org/10.17605/OSF.IO/X8SWP

Cann, R. (1993). *Formal Semantics*. New York: Cambridge University Press.

Cantwell, J. (2008a). The logic of conditional negation. *Notre Dame Journal of Formal Logic, 49(3),* 245-260.

Cantwell, J. (2008b). Indicative conditionals: Factual or Epistemic? *Studia Logica*, *88*(1), 157-194.

Chemla, E. and Egré, P. (2018). From Many-Valued Consequence to Many-Valued Connectives. Retrieved from https://arxiv.org/abs/1809.01066

Chemla, E., Egré, P., and Spector, B. (2017). Characterizing logical consequence in many-valued logic. *Journal of Logic and Computation*, *27*(7), 2193-2226.

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

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6*, 192.

Cruz, N., Over, D., Oaksford, M., & Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1104–1109). Austin, TX: Cognitive Science Society.

Cruz, N., Over, D. E., & Oaksford, M. (2017). The elusive oddness of *or*-introduction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 663-668). Austin, TX: Cognitive Science Society.

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

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

Égré, P. and Cozic, M. (2016). Conditionals. In: Aloni, M. and Dekker, P. (eds.), *The Cambridge Handbook of Formal Semantics*. Cambridge: Cambridge University Press, 490-524.

Elqayam, S. & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction. In S. Elqayam, J. Bonnefon, and D. E. Over (eds.), *Thinking & Reasoning*, 19:3-4, 249-265.

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

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

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

Evans, J. St. B. T., Thompson, V., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology, 6,* 398.

Geis, M. L. and Lycan, W. G. (1993). Nonconditional Conditionals. *Philosophical Topics, 21*(2), 35-56.

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

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

Hahn, U., Harris, A. J. L., and Oaksford, M. (2012). Rational argument, rational inference. *Argument and Computation*, *4*(1), 21-35.

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

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Blackwell Publishing.

Iatridou, S. (1994). On the contribution of then. *Natural Language Semantics*, 2(3), 171-199.

Johnson-Laird, P. N. and Byrne, R. M. J. (2002). Conditionals: A Theory of Meaning, Pragmatics, and Inference. *Psychological Review, 109*(4), 646-678.

Johnson-Laird, P. N., Girotto, V., and Legrenzi, P. (2004). Reasoning From Inconsistency to Consistency. *Psychological Review, 111*(3), 640-661.

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

Joyce, J. M. (2004). Bayesianism. In Miele, A. R. and Rawling, P. (Ed.), *The Oxford Handbook of Rationality* (pp. 132-155). Oxford: Oxford University Press.

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

Khemlani, Orenes, and Johnson-Laird (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, *151*, 1-7.

Khlentzos, D. (2004). *Naturalistic Realism and the Antirealist Challenge.* Cambridge, MA: the MIT Press.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*, 852–884.

Kleiter, G. D. (2018). Adams' p-validity in the research on human reasoning. *Journal of Applied Logics*, *5*(4), 775-825.

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

Kratzer, A. (2012). *Modals and Conditionals*. Oxford: Oxford University Press.

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

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

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lewis, D. (1975). Adverbs of Quantification. In E. Keenan (eds.), *Formal Semantics of Natural Language*, 3-15. Cambridge: Cambridge University Press.

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

Mares, E.D. (2007). *Relevant Logic: A Philosophical Interpretation*. Cambridge: Cambridge University Press.

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

Oaksford, M., & Chater, N. (2009). The uncertain reasoner: Bayes, logic and rationality. *Behavioral and Brain Sciences*, *32*, 105–120.

Oaksford, M. and Chater, N. (2019). New paradigms in the psychology of reasoning. *Annual Review of Psychology* , ISSN 0066-4308. (In Press)

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

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

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

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206-217.

Ramsey, F.P. (1929). General Propositions and Causality. In: H. A. Mellor (eds.), *F. Ramsey: Philosophical Papers*. Cambridge: Cambridge University Press, 1990.

Raidl, E., & Skovgaard-Olsen, N. (2017). Bridging Ranking Theory and the Stability Theory of Belief. *Journal of Philosophical Logic*, *46*(6), 577–609. https://doi.org/10.1007/s10992-016-9411-0

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49* (4), 243-256.

Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, *17*(3), 247-281.

Singmann, H., Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology, 5*, article 316.

Skovgaard-Olsen, N. (2017a). The problem of logical omniscience, the preface paradox, and doxastic commitments. *Synthese*, *194*(3), 917-939.

Skovgaard-Olsen, N. (2017b). *Putting Inferentialism and the Suppositional Theory of Conditionals to the Test.* (Psychology Dissertation, University of Freiburg)

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

Skovgaard-Olsen, N., Kellen, D., Hahn, U., and Klauer, K. C. (2019b). Norm Conflicts and Conditionals. *Psychological Review*. http://dx.doi.org/10.1037/rev0000150

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

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

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

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

Stalnaker, R. (1968). A theory of conditionals. *Studies in Logical Theory*, *2*, 98–112.

Stalnaker, R. (1980). A defense of conditional excluded middle. InW. L. Harper, G. Pearce, & R. Stalnaker (eds.), *Ifs* (pp. 97–104). Dordrecht: Reidel.

Stalnaker, R. (2011). Conditional propositions and conditional assertions. In A. Egan & B. Weatherson (eds.), *Epistemic modality* (pp. 227–248). Oxford: Oxford University Press.

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

Stenning, K., and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT University Press.

Tennant, N. (2002). *The Taming of the True*. Oxford: Oxford University Press.

Thompson, V. A. and Byrne, R. M. J. (2002). Reasoning Counterfactually: Making Inferences About Things That Didn't Happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1154-1170.

Tonhauser, J. and Matthewson, L. (2015). Empirical evidence in research on meaning. Retrieved from http://ling.auf.net/lingbuzz/002595.

van Rooij, R. and Schulz, K. (2018). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, 1-17.

von Fintel, K. (1994). *Restrictions on quantifier domains*. (University of Massachusetts Amherst dissertation)

von Fintel, K. (2011). Conditionals. In K. von Heusinger, C. Maienborn & P. Portner (eds.), *Semantics: An international handbook of meaning*, vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 33.2), 1515–1538. Berlin/Boston: de Gruyter.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part 1: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57.

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

Yalcin, S. (2012). A counterexample to modus tollens. *Journal of Philosophical Logic, 41*, 1001-1024.

Zakkou, J. (2017). Biscuit Conditionals and Prohibited ‘Then’. *Thought, 6*(2), 84-92.

Appendix 1A: Comparison with Skovgaard-Olsen et al. (2019a)

As part of the analysis of Experiment 2, the data from its participants were compared to the data from Skovgaard-Olsen et al. (2019a, Experiment 2), which is publicly accessible at the *Open Science Framework*: https://osf.io/hz4k6/.

Like Experiment 2 of this paper, Skovgaard-Olsen et al. (2019a) conducted their experiment over the Internet using Mechanical Turk and sampling from USA, UK, Canada, and Australia. 105 people participated in the experiment in exchange for a small payment. The exclusion criteria were the same as in Experiment 2 of this paper. The final sample consisted of 67 participants. Mean age was 41.3 years, ranging from 23 to 71 years; 41.8 % of the participants were male; 68.7 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The sample differed only minimally on the demographic variables above before and after applying the exclusion criteria.

Results

Experiment 1 and 2. To investigate whether the findings from Skovgaard-Olsen et al. (2019a, Experiment 2) could be replicated with conditionals without ‘then’ and ‘will’ in the consequent, a set of mixed linear models were fitted to the data. The models had crossed random effects for intercepts and slopes by participants and by scenarios (Baayen, Davidson, and Bates, 2008) to control for the effect of replicates for each participant and item in the experimental design. To investigate whether a replication of the previous results was possible without 'then' and 'will', the models included an ‘Experiment’ factor that indicated whether the data originated from Skovgaard-Olsen et al. (2019a) and included ‘then’ and ‘will’ (Exp 1), or whether the data came from the present replication without ‘then’ and ‘will’ (Exp 2). The models featured the following predictors:

- Model M1A modelled the ratings as a function of the DV factor, encoding the three different types of conditionals (Affirm [if A, C], Wide [¬(if A, C)], Narrow

[if A, ¬C)], the Relevance factor, encoding the two different relevance levels, and of the Experiment factor (Exp1 vs. Exp2). The model also included all the interactions between these three factors.

- Model M2A built upon M1A but did not include the three-way interaction between DV, Relevance, and Experiment.
- Model M3A built upon M2A but did not include the two-way interaction between DV and Experiment.
- Model M4A built upon M3A but did not include the two-way interaction between Relevance and Experiment.
- Model M5A built upon M4A but did not include a main effect of the Experiment factor. M5A thus effectively eliminated the Experiment factor from the model of the two data sets.

In line with the previous studies, these models were implemented in a Bayesian framework with weakly informative priors, using R package `brms` (Bürkner, 2017). One advantage of the Bayesian framework is that it allows us to quantify the evidence in favour of the null-hypothesis in terms of Bayes factors, whereas classical statistics would only have allowed us to conclude that H_0 could not be rejected (Wagenmakers et al. 2018). Since the dependent variable consisted of continuous proportions containing zeros and ones, the values were first transformed to be within the interval [0,1] and a beta-likelihood function was used.[12] Table 1A reports the performance of these models as quantified by WAIC and LOOIC.

[12] Note that in Skovgaard-Olsen et al. (2019a), a zero-or-one inflated beta likelihood function was used to report a similar qualitative pattern as in Figure 1A below. Both are compromise solutions when modelling continuous proportions containing zeros and ones.

Table 1A. Model Comparison

	LOOIC	ΔLOOIC	SE	WAIC	Weight
M1A	-8564.17	3.74	3.70	-8496.1	0.078
M2A	-8564.06	3.85	2.74	-8497.4	0.074
M3A	-8565.73	2.18	1.68	-8498.4	0.170
M4A	-8565.78	2.13	1.55	-8500.2	0.175
M5A	-8567.91	0	--	-8501.2	0.505

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

Table 1A indicates that M5A was the winning model. Consistent with this, evidence of varying degrees could be obtained in favour of the null-hypotheses which set the coefficients of these fixed effects equal to zero for all effects involving the Experiment factor, reflecting the fact that the 95% credible interval in all cases crossed zero. For the three-way interaction, strong evidence in favour of the null-hypothesis was found ($b_{\_IR:Narrow:Exp2}$ = -0.29, 95%-CI [-0.71, 0.13], BF_{H0H1} = 19.0; $b_{\_IR:Wide:Exp2}$ = -0.22, 95%-CI [-0.76, 0.32], BF_{H0H1} = 26.61). For the two-way interaction between DV and Experiment, strong evidence in favour of the null-hypothesis was found ($b_{\_Narrow:Exp2}$ = 0.25, 95%-CI [-0.05, 0.55], BF_{H0H1} = 16.23; $b_{\_Wide:Exp2}$ = 0.05, 95%-CI [-0.26, 0.37], BF_{H0H1} = 58.32). For the two-way interaction between Relevance and Experiment, strong evidence in favour of the null-hypothesis was found ($b_{\_IR:Exp2}$ = 0.10, 95%-CI [-0.23, 0.44], BF_{H0H1} = 47.46). For the main effect of Experiment, strong evidence in favour of the null-hypothesis was found ($b_{\_Exp2}$ = -0.10, 95%-CI [-0.31, 0.11], BF_{H0H1} = 57.38).

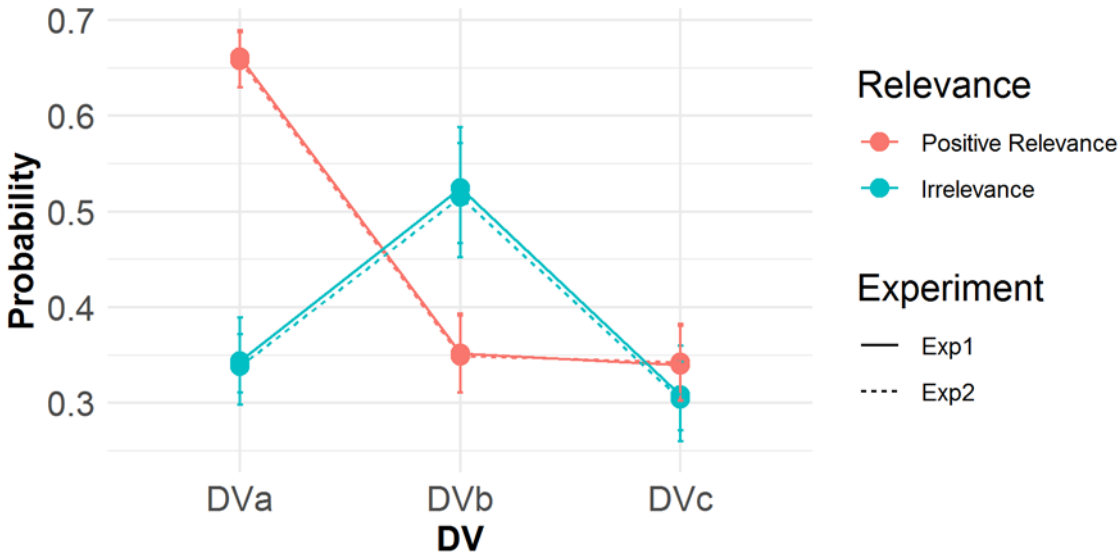


Figure 1A. Predictive posterior means from M1A, M2A, M3A, M4A, M5A weighted by the Akaike weights from Table 1A. 'Exp1' = conditionals with 'then' and 'will'; 'Exp2' = conditionals without 'then' and 'will'. 'DVa' = affirmative conditional; 'DVb' = wide scope negation; 'DVc' = narrow scope.

As Figure 1A indicates, the estimated marginal mean posterior probabilities across experiments were almost identical for all six measures.

Appendix 1B: Bayesian Mixture Model

It was assumed that participants' responses came from a mixture distribution consisting of a group of participants, who had a license to accept a given entailment in session 2 of Experiment 2 based on their probability assignments in session 1 (e.g. accepting the entailment "¬(if A, C) ⊨ if A, ¬C" after conforming to the inequality "P(if A, ¬C) ≥ P(¬(if A, C))" in session 1), and a group of participants who lacked such a license (e.g. conforming to "P(if A, ¬C) < P(¬(if A, C))" in the first session 1). The upper half of Table 1B below displays the license and inference pairs:

Table 1B. Applying the Cross-Task Consistency Constraint

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

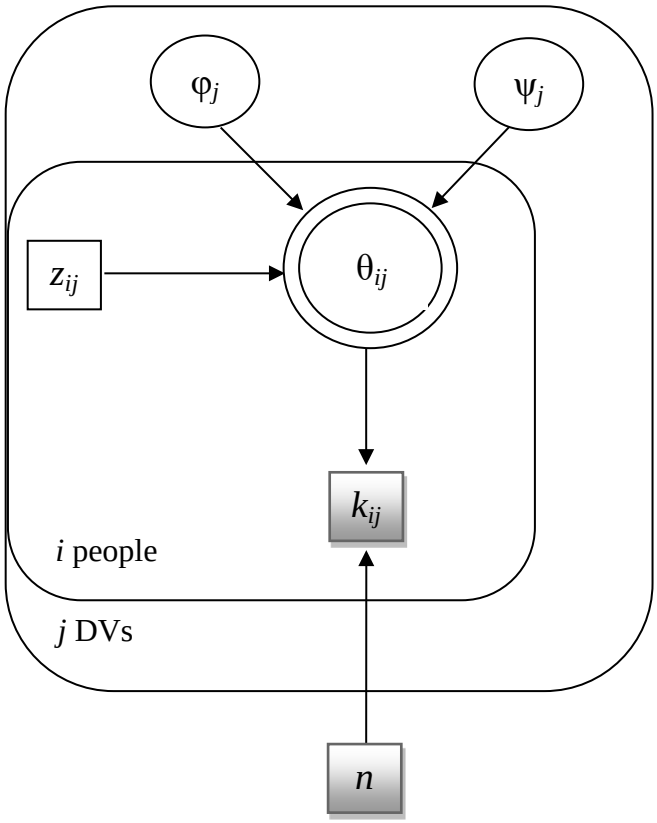
	P(missing license)			
	Positive Relevance		*Irrelevance*	
Agree Baseline	φ = *0.54* [0.50, 0.61]	ψ = *0.22* [0.13, 0.32]	φ = *0.53* [0.50, 0.59]	ψ = *0.097* [0.04, 0.17]
Disagree Baseline	φ = *0.80* [0.71, 0.87]	ψ = *0.48* [0.43, 0.50]	φ = *0.56* [0.50, 0.66]	ψ = *0.34* [0.22, 0.46]
Target Inference	φ = *0.54* [0.50, 0.61]	ψ = *0.13* [0.05, 0.22]	φ = *0.68* [0.55, 0.80]	ψ = *0.31* [0.19, 0.43]
	P(acceptance of entailment) = 1 – P(missing license)			
Agree Baseline	*0.46*	*0.78*	*0.47*	*0.90*
Disagree Baseline	*0.20*	*0.52*	*0.44*	*0.66*
Target Inference	*0.46*	*0.87*	*0.32*	*0.69*

Note. The 95%-credible intervals for the parameter estimates are listed in square brackets. The bottom row indicates the predicted posterior probabilities of acceptance of the entailments based on the latent classes in session 1. The grey boxes in the bottom row indicate the modal session 1 classification (n = 33).

To classify participants into two latent classes, the prior recommendations and Bayesian mixture models in Lee and Wagenmarkers (2014) were followed. Essentially, information or ignorance regarding the model parameters is represented by *prior distributions*. The observed data is then used to update our knowledge about the parameters, resulting in *posterior parameter distributions* (Kruschke, 2014; Lee & Wagenmakers, 2014; Skovgaard-Olsen et al. 2019b). As shown in Table 2B, participants' conformity to/violation of a given inequality (e.g. the Target Inference license) was modelled as produced by binominal rate parameters that come from two distributions (the φ_j distribution that was constrained to be above 0.5 or the ψ_j distribution, which was constrained to be below 0.5). An uninformed indicator variable (z_{ij}) classified which distribution a given participant belonged to in a given experimental condition. Based on the posterior probabilities of the indicator variables z_{ij}, each individual was classified per condition as possessing or lacking an inference license. Since Positive Relevance and Irrelevance were modelled separately, and there were three types of inference licenses (see Table 1B), six binominal rate parameters were assigned to a given participant

based on four trial replications (HH, HL, LH, LL). Identifiability was ensured by applying the constraint that the two binominal rate parameters were identical across participants for the two latent classes for a given DV. The lower half of Table 1B lists the estimated parameters.

Table 2B. Bayesian Mixture Model



z_{ij} ~ Bernoulli(0.5)

φ_j ~ beta(1,1)T(0.5,1)

ψ_j ~ beta(1,1)T(0,0.5)

$$\theta_{ij} \leftarrow \begin{cases} \varphi_j \; if \; z_{ij} = 1 \\ \psi_j \; if \; z_{ij} = 0 \end{cases}$$

k_{ij} ~ Binominal(θ_{ij},n)

Note. beta(1,1)T(0.5, 1) indicates that the beta-distribution with the shape-parameters $\alpha = 1$ and $\beta = 1$ is truncated to only take values from the interval [0.5, 1]. DV ∈ {Agree license$_{PO}$, Agree license$_{IR}$, Disagree license$_{PO}$, Disagree license$_{IR}$, Target license$_{PO}$, Target license$_{IR}$}.