Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'

Niels Skovgaard-Olsen

University of Konstanz and Albert-Ludwigs-Universität Freiburg

David Kellen

Syracuse University

Hannes Krahl

University of Chemnitz

Karl Christoph Klauer

Albert-Ludwigs-Universität Freiburg

Author Note

Niels Skovgaard Olsen, Department of Philosophy, University of Konstanz, Konstanz, Germany, and Department of Psychology, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany. David Kellen, Department of Psychology, Syracuse University, USA. Hannes Krahl, Department of Psychology, University of Chemnitz. Karl Christoph Klauer, Department of Psychology, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

We are very grateful for discussions with Igor Douven, Wolfgang Spohn, Karolina Krzyżanowska, Peter Collins, Ulrike Hahn, Maria Biezma, Ulrike Hahn, and the audiences of talks at the annual meeting of New Frameworks of Rationality, and the What-If group in Konstanz. Moreover, careful comments by David Over, Valerie Thompson, Henry Markovits, and anonymous reviewers substantially improved the manuscript.

This work was supported by grants to Wolfgang Spohn and Karl Christoph Klauer from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

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

Abstract

In this study we investigate the influence of reason-relation readings of indicative conditionals and ‘and’/‘but’/‘therefore’ sentences on various cognitive assessments. According to the Frege-Grice tradition, a dissociation is expected. Specifically, differences in the reason-relation reading of these sentences should affect participants’ evaluations of their acceptability but not of their truth value. In two experiments we tested this assumption by introducing a relevance manipulation into the truth-table task as well as in other tasks assessing the participants’ acceptability and probability evaluations. Across the two experiments a strong dissociation was found. The reason-relation reading of all four sentences strongly affected their probability and acceptability evaluations, but hardly affected their respective truth evaluations. Implications of this result for recent work on indicative conditionals are discussed.

Keywords: Relevance, truth conditions, indicative conditionals, probability, acceptability, conjunctions

Relevance Effects on ‘And’, ‘But’, ‘Therefore’, and ‘If Then’

The goal of the present paper is to investigate experimentally which cognitive assessments are affected by salient reason-relation readings of two conjoined sentences. Following the tradition in linguistics, philosophy and psychology, we will focus on truth, probability, and acceptability evaluations (Garmut, 1991; McCawley, 1993; Nickerson, 2015). For our target sentences, we have chosen indicative conditionals (e.g. ‘if it rains, then Sally’s birthday party will be ruined’) and ‘but’, ‘and’, and ‘therefore’ sentences (henceforth referred to as 'ABT sentences'). As we will explain in detail below, a focused comparison between these sentences allows us to investigate which aspects of their meaning are influenced by their reason-relation reading.

The standard work of reference in this context is Grice’s (1989) seminal work. In that book, Grice was concerned with enforcing *Modified Occam’s Razor* as a methodological principle, according to which “*Senses are not to be multiplied beyond necessity*” (p. 47). Grice’s central point was that when modeling the meaning of natural language content, we can keep the logic simple by distinguishing between the truth-conditional content of sentences and the different *implicatures* that enrich the truth-conditional content of sentences. In particular, Grice (1989) was concerned about things such as preventing that:

1) we reject the truth table of the material implication (‘⊃’) as a model of the truth conditional content of indicative conditionals (see Table 1 below) based on the fact that natural language conditionals have a prominent reason-relation reading, according to which the antecedent is a reason *for* the consequent,

2) we reject the truth table of logical conjunction (‘∧’) based on the fact that natural language conjunctions have readings that indicate temporal succession and causal relations, and
3) we reject logical conjunction as a model of the truth-conditional content of sentences involving ‘but’ and ‘therefore’ based on the salient reason-relation reading of natural language sentences in which they occur.

Grice (1989) introduced a key distinction between *conversational* and *conventional implicatures.* These twin notions provided the means to explain how these various sentences could have a meaning that goes beyond their truth-conditional content. The notion of a conversational implicature was called upon to explain the pragmatic phenomenon that we often use words to convey a meaning that differs from what we literally say. For instance, when Julia responds that “I don’t like parties” to the question “Are you going to the party tonight?”, her interlocutor can infer that the message that Julia really intends to convey is that she won’t go to the party tonight (Blome-Tillmann, 2013). Grice’s (1989) project was to reconstruct such pragmatic inferences to the intended meaning rationally, based on maxims of informative, truthful, relevant, and clear communication that implement the goal of cooperative discourse.

In addition, Grice (1989) also introduced the category of conventional implicatures to account for the cases, where it is not a feature of the utterance of a sentence in a particular context that invites an interpretation of the intended meaning that differs from the truth-conditional content. In particular, Grice thought that it was part of the conventional meaning of ‘therefore’ to introduce a consequence relation (e.g. ‘He is an Englishman; he is, *therefore*, brave’), which goes beyond the sentence's truth conditions as modeled by logical conjunction.

Another example is that sentences such as ‘she was poor *but* honest’ have been thought to express a contrast between being poor and honest since Frege (1892), which is not to be found in its truth-conditional equivalent ‘she was poor *and* honest’. Often the treatment of ‘but’ is left with this observation in the philosophical literature. In fact, the analysis of ‘but’ is a rich topic in linguistics, where at least four different readings are dissected. However, attempts have been undertaken to subsume these various readings under the prominent denial of expectation reading, where an implicit or explicit assumption is denied by the second clause (Iten, 2000, Chap. 5; Blackmore, 2004, Chap. 4), which is closely related to a reason-relation reading.

Common to both conversational and conventional implicatures is the expectation that their content will not affect the truth-conditional content of the sentences in which they occur (Potts, 2015). Conversational implicatures and conventional implicatures differ in that only conversational implicatures can be cancelled (e.g. adding the qualification “I don’t like parties, but I would be happy to come tonight” cancels the conversational implicature that Julia won’t attend the party tonight). Furthermore, only conventional implicatures can be detached, or removed, by uttering a different sentence with the same truth-conditional content in the context of utterance (e.g. uttering “she was poor *and* honest” instead of “she was poor *but* honest”). Finally, only conversational implicatures can be reconstructed on the basis of Grice’s maxims of conversation (Blome-Tillmann, 2013).

It is not uncommon to find references to semantic and pragmatic modulation in the psychological literature (Johnson-Laird and Byrne, 2002), or to invoke Grice’s theory to explain pragmatic effects (Nickerson, 2015). Given the prominent role that Grice’s thought continues to have on theorizing about natural language, it is important that its central claims are subjected to empirical testing and do not just figure as *ad hoc* explanations that are invoked to explain divergent results when convenient.

One natural assumption is that meaning components classified as conversational or conventional implicatures influence the acceptability/assertability assessments of the participants, while not affecting their truth evaluations. This implies a dissociation between the influence of factors relating to conversational and conventional implicatures on acceptability/assertability[1] and truth-value assessments. Although probability assignments were not originally discussed in this context, we would expect them to vary in tandem with the acceptability judgments.

As noted above, the ABT sentences have been thought to be truth-conditionally equivalent in the Frege/Grice tradition. In addition, therefore-sentences have a salient reason relation reading, whereby φ is a reason *for* ψ in "Nick forgot about her birthday (φ), therefore he didn't buy Sally a present (ψ)". In contrast, φ appears to be a reason *against* ψ in "Nick dislikes Sally (φ), but he attended her birthday party (anyway) (ψ)". Finally, and-sentences seem to fall somewhere in between and can be used as a baseline for comparing these two extremes.

In light of these assumed differences in reason-relation readings despite truth-conditional equivalence, comparing the ABT sentences on their truth, probability, and acceptability evaluations will give us important clues about whether the abovementioned dissociation can be found. If the differences in their reason-relation reading are attributable to implicatures, then these differences should only show up in their acceptability evaluations and probability evaluations. Yet their truth evaluations should remain unaffected. If, on the contrary, the

[1] In this paper we will not distinguish the two.

assumption is wrong that the ABT sentences only differ in their implicatures, then we should see evidence of their different reason-relations readings affecting truth evaluations.

There is a long tradition in the psychology of reasoning of investigating truth conditions by presenting participants with the cells of truth tables in the *truth-table task* (for reviews, see Manktelow, 2012; Nickerson, 2015). However, in order to address the question of what range of cognitive evaluations the reason-relation reading affects, a measurement tool is needed to capture the presence and absence of specific reason relations and to be able to vary it orthogonally to the presence or absence of other psychological factors of interest.

Reason Relations

For the purpose of this paper, we will rely on Spohn's (2012, Chap. 6) explication of reason relations. According to Spohn, the reason relation and the notion of epistemic relevance can be explicated as follows:

THE ΔP RULE: $\Delta p = P(\psi \mid \varphi) - P(\psi \mid \neg\varphi)$

POSITIVE RELEVANCE/φ IS A REASON FOR ψ: $\Delta p > 0$

IRRELEVANCE: $\Delta p = 0$

NEGATIVE RELEVANCE/φ IS A REASON AGAINST ψ: $\Delta p < 0$

The underlying intuition is that relevant information changes the probability of the propositions that it concerns. The probabilistic change is here explicated through a comparison between conditioning on the information and conditioning on its negation. When such a comparison shows that φ increases the probability of ψ, then φ is said to be a *reason for* ψ. When it shows that φ decreases the probability of ψ, then φ is said to be a *reason against* ψ. Using these explications, we specify Grice's conventional-implicature hypothesis as follows: sentences containing 'φ therefore ψ' differ from sentences containing the connective 'φ and ψ' in suggesting that φ is a reason *for* ψ (Δp > 0), and sentences containing 'φ but ψ' differ from the latter in suggesting that φ is a reason *against* ψ (Δp < 0).

In Skovgaard-Olsen, Singmann, and Klauer (2016b), participants' perceived reason relations and relevance were investigated for a large range of everyday contexts and empirical support for the explication of the reason relation above could be obtained. The stimulus materials used in the present study contains 12 scenarios with 12 different conditions, implementing all permutations of positive relevance (PO), negative relevance (NE), irrelevance (IR) and high (H) and low (L) prior probability within each scenario. For Δp values ranging from -1 to 1, a pretest with 725 participants showed that the average Δp was .32 for the *positive relevance conditions*, -.27 for the *negative relevance conditions*, -.01 for the *irrelevance conditions*. Moreover, the pretest found that the average prior probability was ca. 70% for the high probability items and less than 30% for the low probability items.

The scenarios used were designed to trigger the participants' stereotypical assumptions about basic causal, functional, or behavioral information to implement the above relevance categories. For instance, one scenario text runs as follows:

> Scott was just out playing with his friends in the snow. He has now gone inside but is still freezing and takes a bath. As both he and his clothes are very dirty, he is likely to make a mess in the process, which he knows his mother dislikes.

As verified in Skovgaard-Olsen *et al*. (2016b), the two sentences 'Scott turns on the warm water' and 'Scott will be warm soon' both have a high prior probability, and the first sentence raises the probability of the second (PO HH). Moreover, the sentence 'Scott turns on the cold water' has a low prior probability and it lowers the probability of 'Scott will be warm soon' (NE LH). Finally, in the absence of further information it seems reasonable to assume that 'Scott's friends are roughly the same age as him' has a high prior probability. Yet this sentence leaves the probability of the previous sentences unchanged and is therefore irrelevant for them. Now if 'φ therefore ψ' expresses positive relevance, then the sentence 'Scott turns on the warm water *therefore* Scott will be warm soon' should sound fine, and if 'φ but ψ' expresses negative relevance then the sentence 'Scott turns on the warm water *but* Scott will be warm soon' should sound rather strange. Conversely, the negative relevance version should sound better with 'but' ('Scott turns on the cold water *but* Scott will be warm soon') yet strange with 'therefore' ('Scott turns on the cold water *therefore* Scott will be warm soon').

One area in which these explications have already been applied is in the empirical research on conditionals to which we now turn. As we shall see, indicative conditionals are another type of sentences which have a salient reason-relation reading, and there is currently a large theoretical interest in this field in diagnosing for which types of cognitive assessments the reason-relation reading plays a role. Accordingly, we will be interested in whether the same kind of dissociations predicted above with respect to the ABT sentences can be found for the cognitive evaluations of indicative conditionals.

Reason Relations and Indicative Conditionals

Throughout the last 10-15 years, a new paradigm has been introduced to the psychology of reasoning (Elqayam & Over, 2013) with researchers turning to probabilistic competence models of reasoning and drawing on Bayesian formal epistemology (Pfeifer & Douven, 2014). In the study of conditionals, this paradigm-shift is reflected in the widespread endorsement of "the Equation", the Ramsey Test, and the de Finetti truth table (see Table 1 below).

THE EQUATION: $P(\text{if } \varphi, \text{ then } \psi) = P(\psi \mid \varphi)$ (Bennett, 2003; Evans & Over, 2004; Oaksford & Chater, 2007).

THE RAMSEY TEST: Instead of calculating conditional probabilities by means of the ratio $P(\varphi \text{ and } \psi)/P(\varphi)$, the participants are conjectured to evaluate conditional probabilities on the basis of the Ramsey test. The Ramsey test is a mental algorithm that temporarily adds the antecedent to the participant's stock of beliefs, makes minimal changes to preserve consistency, and evaluates the consequent under its supposition (Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer, 2013).

Table 1. Truth Tables of the Indicative Conditional

Truth-Conditional Inferentialism				
	⊤⊤	⊤⊥	⊥⊤	⊥⊥
PO	⊤	⊤	⊤	⊤
NE	⊥	⊥	⊥	⊥
IR	⊥	⊥	⊥	⊥
Material Implication Account				
PO	⊤	⊥	⊤	⊤
NE	⊤	⊥	⊤	⊤
IR	⊤	⊥	⊤	⊤
De Finetti Table				
PO	⊤	⊥	void	void
NE	⊤	⊥	void	void
IR	⊤	⊥	void	void

Note. '⊤' = true; '⊥' = false; PO = positive relevance; NE = negative relevance; IR = irrelevance.

As discussed in Elqayam and Over (2013), there are two direct sources of evidence supporting this approach to conditionals. First, it has long been known that the participants tend to judge the false antecedent cells of the truth table (⊥⊤, ⊥⊥) to be 'irrelevant' to the truth or falsity of the indicative conditional in the truth-table task (an effect known as "the defective truth table"). This defective truth table effect has been interpreted as direct evidence in favor of the *de Finetti truth table* (see Table 1), because it is hard to reconcile with accounts based on the material implication, which would require that the conditional is treated as true in the false antecedent cells (Over & Evans, 2003). Second, direct investigations of the probability of indicative conditionals have repeatedly supported the Equation. For example, $P(\psi \mid \varphi)$ turns out to be a much better predictor of $P(\text{If } \varphi, \text{ then } \psi)$ than $P(\neg\varphi \vee \psi,)$, the probability under the material implication interpretation (Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Douven, 2015b, Chapters 3, 4).

Despite this empirical support for The Equation, recent results have shown, however, that the relationship between P(if φ, then ψ) and P(ψ | φ) is moderated by the relevance manipulation reviewed above. Employing the above-described stimulus material, Skovgaard-Olsen *et al.* (2016a) showed that the marginal means of assessments of P(if φ, then ψ) were judged to be substantially lower in NE and IR than in the PO condition and that the slopes of the regression lines employing P(ψ | φ) as a predictor of P(if φ, then ψ) were steeper in the PO condition. These results corroborate the idea that there is something defective about indicative conditionals that violate the default assumption of positive relevance. Thus whereas the sentence ‘If Scott turns on the warm water, then Scott will be warm soon’ (PO HH) has a high probability in the scenario above, sentences like ‘If Scott’s friends are roughly the same age as him, then Scott will turn on the warm water’ (IR HH) and ‘If Scott makes an effort to be tidy, then the bathroom will be dirtier than before he took his bath’ (NE HH) have a much lower probability than what would be expected based on P(ψ | φ) alone.

In earlier empirical work (Oberauer, Weidenfeld, & Fischer, 2007; Over, Hadjichristidis, Evans, Handley, and Sloman, 2007; Singmann, Klauer, & Over, 2014), the Δp rule was investigated in relation to probability evaluations of conditionals, with little or no evidence for a relationship. But the crucial difference from Skovgaard-Olsen *et al.*’s (2016a) original study consists in that these earlier studies did not systematically vary Δp to be negative, equal to zero and positive for realistic stimulus materials. In addition, Skovgaard-Olsen *et al.* (2016a) did not assume that P(if φ, then ψ) would be directly predicted by either Δp or P(ψ | ¬φ) as its proxy. Instead a heuristic for judging P(if φ, then ψ) based on a reason relation assessment was formulated according to which the presence of negative relevance or irrelevance would make the participants apply a penalty to P(if φ, then ψ) based on the conditionals' perceived defect in expressing a reason relation.

In philosophy, it has long been argued that indicative conditionals without connection between the antecedent and the consequent (e.g. 'If Copenhagen is in Denmark, then H. C. Anderson is dead') are defective (Spohn, 2013; Olsen, 2014; Douven, 2015a; Krzyżanowska, 2015). This intuition has not, however, been integrated into contemporary treatments of conditionals in the psychology of reasoning based on either the classical treatment of conditionals as the material implication ('⊃'), or by probabilistic accounts based on the de Finetti truth table and the Ramsey test (Skovgaard-Olsen *et al.*, 2016a).

As an alternative, Krzyżanowska, Wenmackers, and Douven (2014) and Krzyżanowska (2015) argued that there should be an inferential relation (or reason relation) between the antecedent and the consequent as part of the truth conditions of indicative conditionals. In Krzyżanowska (2015, p. 62), the presence of either an indicative, abductive, or deductive inferential relation between the antecedent and the consequent is made part of the truth conditions of indicative conditionals. Moreover, they allow for auxiliary assumptions coming from the participants' background knowledge to play a role in determining the inferential relation and add further qualifications that need not concern us here. Based on this line of reasoning, *truth-conditional inferentialism* predicts that the modal truth value assignments should follow the first rows in Table 1.

One thing to note about the 'True' prediction for the ⊤⊥ cell is the following:[2] according to Krzyżanowska (2015: Chap.. 3), the inferential relation between φ and ψ admits exceptions (i.e. the ⊤⊥ cells) via inductive or abductive consequence relations. Inductive consequence relations would be instances of probability-raising based on purely frequentist information. Abductive consequence relations would be instances of probability raising based on explanatory considerations (e.g. causal structures and theoretical assumptions). Most likely, the stimulus material reviewed above, and further specified in Table 2 below, instantiates the abductive consequence relation. But crucially, it is not based on the deductive consequence relation, which does not admit of exceptions (i.e. the ⊤⊥ cells).

Note that whereas the Frege-Grice tradition is based on ignoring relevance differences between sentences in truth-value evaluations, truth-conditional inferentialism stands out by making predictions that directly vary with the levels of the relevance factor. In contrast, both the *material implication* and the *de Finetti truth tables* follow the Frege-Grice tradition in assigning truth conditions to indicative conditionals that are invariant across the different levels of the relevance factor (see Table 1).

For both the ABT sentences and indicative conditionals, the main focus of our experiments is to investigate whether evidence for the hypothesized dissociation can be found that the reason relation readings of these sentences affect the acceptability and probability evaluations of these sentences, but not their truth evaluations. In order to investigate these issues, Experiment 1 sets out to introduce the relevance manipulation into the truth-table task and

[2] Both Karolina Krzyżanowska and Igor Douven have confirmed that Table 1 is a reasonable explication of the definition cited above (personal communication, February, 2016). However, Karolina Krzyżanowska did express some doubts about her earlier proposed definition and expressed concerns that it is less clear whether truth conditional inferentialism is really committed to predicting True for the ⊤⊥ cell in the positive relevance condition.

Experiment 2 introduces the relevance manipulation into tasks that elicit the probability and acceptability judgments of our four target sentences.

Experiment 1

As discussed above, it is part of the dissociation predicted by the Frege-Grice tradition that the truth table of logical conjunction '$\wedge$' fits the truth tables 'φ and ψ', 'φ but ψ', and 'φ therefore ψ', and that no effect of the relevance conditions can be found for these three sentences (a hypothesis we denote as 'H_{0_ABT}'). Similarly, the material implication and de Finetti truth tables predict that no effect of the relevance condition should be found for 'if φ, then ψ' (a hypothesis we denote as 'H_{0_IF}') and that the participants' truth evaluations fit their respective truth tables (see Table 1). In contrast, truth-conditional inferentialism predicts that a relevance effect on the truth evaluations of 'if φ, then ψ' can be found, and that its respective truth table accurately describes participants' modal responses.

Method

Participants

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 752 people completed the experiment. The participants were sampled through Mechanical Turk from USA, UK, and Australia and were paid a small amount of money for their participation.

The following exclusion criteria were used: 1) not having English as native language, 2) failing to answer two simple SAT comprehension questions correctly in a warm-up phase, 3) completing the task in less than 160 seconds or in more than 3600 seconds, and 4) answering 'not serious at all' to the question of how serious the participant would take his or her participation at the beginning of the study. The final sample consisted of 557 participants. Mean age was 38 years, ranging from 18 to 75 years; 36 % of the participants were male; 72 % indicated that the highest level of education that they had completed was an undergraduate degree or higher. The demographic measures of the participants differed only minimally before and after the exclusion.

Design

The experiment involved a within-subject design. Specifically, there were three factors that were varied within participants: 1) sentence, with four levels: '…and…', '…but…', '…therefore…', 'if…, then…'; 2) relevance, with three levels: PO, NE, and IR; and 3) priors, with four levels: HH, HL, LH, and LL (e.g. LH indicates that P(φ) = low and P(ψ) = high). The prior manipulation had the goal of ensuring that the participants' truth evaluations were representative across different combinations of prior probabilities of the sentences in question.

Materials

The twelve within-participants conditions crossing the factors relevance and priors were randomly assigned to twelve different scenarios for each participant. Within each relevance level, each participant saw all four sentence levels randomly distributed across the priors manipulation. One participant might thus see the sentences '…and…', '…but…', '…therefore…', 'if…, then…' in the PO level in the HH, LH, LL, HL prior levels, whereas the next would see them in a different permutation of the priors factor.

With minor adjustments, the twelve scenarios used in this study were obtained from a large pre-study (Skovgaard-Olsen *et al.*, 2016b).[3] From each of the twelve selected scenarios we

[3] These minor adjustments concern slight formulation changes to a few of the sentences and changing the temporal structure of all the sentences. Whereas the sentences in Skovgaard-Olsen *et al.* (2016b) had the temporal form of 'if φ occurs, then ψ will occur', their temporal form was 'if φ is now happening, then ψ will occur' (or: 'φ is now happening *and*/*but*/*therefore* ψ will occur'). The latter temporal form was introduced in the present study to allow for the

were able to construe all twelve within-participant conditions. Consequently, mapping of the condition to each possible scenario was completely randomized for each participant anew.

To better illustrate these differences, Table 2 contains all of the experimental conditions for the 'Scott scenario' presented in the Introduction here illustrated using the connective 'And'.

Table 2. Stimulus Materials, Scott Scenario illustrated with And-Sentences

	PO	NE	IR
HH	Scott is now turning on the warm water AND he will be warm soon.	Scott is now making an effort to be tidy AND the bathroom will be dirtier than before he took his bath.	Scott's friends are now also going home to take a bath AND Scott will turn on the warm water.
HL	Scott is now making an effort to be tidy AND the bathroom will be just as clean as before he took his bath.	Scott is now turning on the warm water AND he will soon start to freeze even more.	Scott's friends are now also going home to take a bath AND Scott will turn on the cold water.
LH	Scott is now bathing in a hot spring AND he will be warm soon.	Scott is now turning on the cold water AND he will be warm soon.	Scott's friends are now participating in the Winter Olympics AND Scott will turn on the hot water.
LL	Scott is now turning on the cold water AND he will soon start to freeze even more.	Scott is now bathing in a hot spring AND he will soon start to freeze even more.	Scott's friends are now participating in the Winter Olympics AND Scott will turn on the cold water.

Note. PO = positive relevance; NE = negative relevance; IR = irrelevance.

The complete list of stimulus materials, R-scripts, and raw data can be found on *the Open Science Framework*: https://osf.io/yder9/.

Modified Truth Table Task

For each of the twelve priors×relevance within-participants conditions stemming from our experimental design, the participants were presented with two pages. The first page featured a modified version of the truth-table task. In typical implementations of the truth-table task participants are asked to evaluate the truth value of a conditional statement on the basis of an outcome statement describing a cell in the truth table (TT, TF, FT, FF) with either binary or ternary response options (Schroyens, 2010; Nickerson, 2015, pp. 38). In our modified version, the participants were asked for each trial to evaluate the truth value of a randomly chosen

introduction of a reversal condition to test for violations of commutativity, which was later dropped prior to launching the experiment, however.

sentence from our four target sentences (‘φ and ψ’, ‘φ but ψ’, ‘φ therefore ψ’, ‘if φ, then ψ’) on the basis of two randomly chosen truth table cells.

Since none of the truth tables reviewed in the introduction holds that speaker intentions and the Gricean maxims should play a role for the truth evaluation of the target sentences, we decided to test the participants’ truth evaluations under the relevance manipulation in a situation in which they would not have to worry about the speaker intentions behind uttering the strange IR items. To achieve this the participants were instructed that they should consider the target sentences as output produced by a computer program in the development phase in response to the scenario texts as input. A computer program in the development phase does not have any communicative intentions when producing odd sentences. Hence, calibrating its output based on truth values should increase a focus on truth evaluations of the sentences produced based solely on their content. The combination of naturalistic stimulus materials with our computer calibration task meant that the participants were encouraged to set aside concerns about speaker intentions behind the presented assertions and encouraged to use their background knowledge, underlying our manipulation of relevance, in evaluating their content. To illustrate, one participant might have seen the following scenario text:

> *INPUT: Scott was just out playing with his friends in the snow. He has now gone inside but is still freezing and takes a bath. As both he and his clothes are very dirty, he is likely to make a mess in the process, which he knows his mother dislikes.*

with the following PO HH sentence presented as output produced by the computer:

```
Computer output: Scott is now turning on the warm water BUT he
will be warm soon.
```

To help the participants organize the information, the output sentences were distinguished by a different font, as illustrated above. Following the output sentences, two randomly chosen truth-table cells were presented as continuations of the scenarios which occurred after the computer produced its output:

Continuation: Scott turned on the warm water. He did become warm.

On the basis of this continuation, the computer output turned out to be:

True False Neither true nor false

As shown below, the task of the participants was then to help us calibrate the output sentences of the computer by evaluating, separately for each continuation, whether the output sentences were ‘True’ (⊤), ‘False’ (⊥), or ‘Neither true nor false’ (NN) given the continuations of the scenario. The exact wording of the instruction was as follows:

> On each of 13 pages, you will read, in order, a short text describing a scenario, a sentence, and two different continuations of the scenario. For each case, we ask you to imagine that a computer has been given the scenario text as input and that it produced the sentence as output. The computer is still in the development phase and we need you to help us calibrate its output sentences. For each of the two possible continuations of the scenario, your task is to evaluate whether the sentence produced by the computer turned out to be 'True', 'False', or 'Neither true nor false' by the way the scenario developed.

Before continuing to our 12 experimental conditions, the participants first saw a practice trial, which was later discarded in the analysis.

On the second page, the participants were instructed to evaluate how confident they were in their responses on a scale from 0 % to 100 %.

Procedures

To reduce the dropout rate once the proper experiment started, participants first went through three pages that: 1) stated our academic affiliations, 2) asked for personalized information (which was not paired with the participants' other responses, however), 3) posed two SAT comprehension questions in a warm-up phase, and 4) presented a seriousness check emphasizing the importance of careful responses for the scientific utility of the results (Reips, 2002). After a practice trial and a repetition of the instructions, the experiment itself began with the presentation of the twelve within-participants conditions. Their order was randomized anew for each participant.

Results

The observed response frequencies were analyzed with multinomial processing tree models (MPT; Riefer & Batchelder, 1988), a well-known class of models that provides a convenient testbed for hypothesis concerning categorical data. We will evaluate the MPT models' absolute performance via the G^2 statistic (Read & Cressie, 1988) and their relative performance with the Fisher Information Approximation (FIA; Grünwald, 2007). FIA is a model-selection statistic that penalizes models according to their functional flexibility and improves upon traditional statistics such as AIC and BIC. Further details on the models and the analyses are provided in the Appendix. But it is worth to highlight here that the models assumed that individual's responses are stochastic in the sense that they can fail to reflect their true judgments with some probability. When specifying the different hypotheses, we relied on the most lenient stochastic specification, which only imposes the constraint that the preferred response option should be the modal response. For example, in the case of the TT cell, the stochastic implementation of the material implication account then predicts that

$$P(\top) \geq P(\bot), P(\text{NN})$$

The reason behind the adoption of this specific stochastic specification is the diagnostic power associated with its failure, as any theory that fails to succeed under these minimal constraints faces a severe explanatory challenge.

As can be seen from Figure 1 (right upper panel), aside from the ⊥⊥ cell, there does not appear to be much of a relevance effect for the truth evaluations of the indicative conditional. Indeed, for the true antecedent cells (⊤⊤, ⊤⊥), there appears to be an absolute majority for ⊤ and ⊥ respectively across the different relevance levels, which is shared by all of the *ABT sentences* (‘φ and ψ’, ‘φ but ψ’, ‘φ therefore ψ’). It is only for the false antecedent cells (⊥⊤, ⊥⊥) that the indicative conditional seems to stand out from the ABT sentences. For the ABT sentences, there is an absolute majority of ⊥ responses for the false antecedent cases, whereas there is a mixed response for the indicative conditional with large differences across the relevance levels for the ⊥⊥ cell with the indicative conditional.

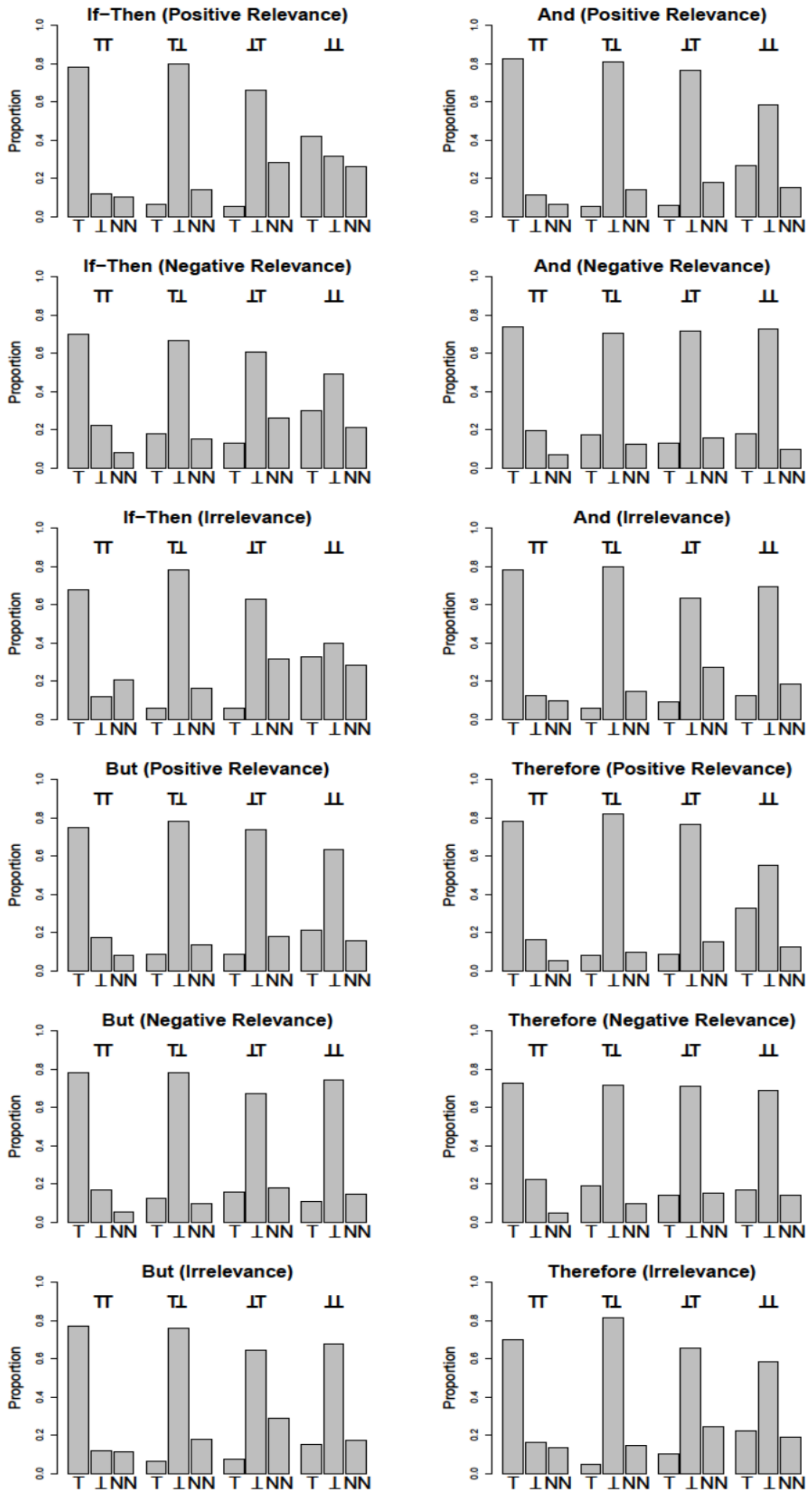


Figure 1. Truth evaluations across sentences, relevance condition, and truth-table cells. Participants could respond ‘True’ (⊤), ‘False’ (⊥), or ‘Neither True nor False’ (NN).

The ABT Sentences

We first tested whether response distributions differed across sentences and relevance levels. Our general approach for testing different hypothesis was the following: We fitted a set of constrained MPT models representing different hypotheses (e.g., response distributions do not differ across sentence levels), and compared their performance with an unconstrained model ($M_{\text{saturated}}$) that fits the data perfectly using one free parameter per degree of freedom provided by the data. An hypothesis (instantiated by a constrained model) is said to be rejected when it performs worse than the unconstrained model and/or any of the competing alternative hypotheses (even after taking differences in flexibility into account via FIA).

For the ABT sentences, we considered four models: 1) $M_{\text{saturated}}$, which imposes no constraints whatsoever, 2) M_{sentence}, which assumes that response probabilities are the same across sentences, but allows for differences across the different levels of the relevance factor, 3) $M_{\text{relevance}}$, which assumes no differences across the relevance manipulation, but allows responses to differ across sentences, and 4) M_{full}, which assumes no differences across both relevance and sentence levels. As shown in Table 3, the preferred model for the ABT sentences in terms of FIA was clearly $\text{M}_{\text{sentence}}$. Note that FIA differences larger than 3.40 already indicate a very strong preference for the winning model (see Kellen, Klauer, & Bröder, 2013). In absolute terms, all models were rejected under a significance level of $\alpha = .05$. However, it should be noted that when large samples are used, any minor deviation from model predictions can lead to a statistically-significant misfit, and having a *p*-value of .02 for $\text{M}_{\text{sentence}}$ can be considered satisfactory. In fact, in structural-equation modelling, which faces a similar problem of large samples in the interpretation of χ^2 tests, a ratio χ^2/df between 0 and 2 is considered to indicate a good fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003), and here G^2/df = 1.49.

Table 3. Model-Comparison Results: ABT

Model	G^2	df	*p*	ΔFIA
$M_{saturated}$	0	0	1	86.04
$M_{sentence}$	71.40	48	.02	0
$M_{relevance}$	237.51	48	.00	83.05
M_{full}	271.76	64	.00	50.84

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

ABT Sentences and Indicative Conditionals

When comparing the ABT sentences with the indicative conditionals ('if-then' sentences in Figure 1), we separated the true antecedent truth table cells from the false antecedent cells. In both cases, we considered three models: 1) $M_{\text{saturated}}$, 2) M_{same}, which assumes that responses to the ABT sentences are equal to the ones given to their indicative-conditional counterparts, and 3) $M_{\text{different}}$, which assumes the same responses rates across the ABT sentences, but allows them to differ from the responses to the indicative conditionals. In the case of the true antecedent cells, M_{same} was the preferred model (see Table 4). In contrast, $M_{\text{different}}$ was preferred in the case of the false antecedent cells (see Table 5).

Table 4. Model-Comparison Results: TT, TF, ABT + IF

Model	G^2	df	*p*	ΔFIA
$M_{saturated}$	0	0	1	64.12
M_{same}	57.30	36	.01	0
$M_{different}$	31.84	24	.13	19.16

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

Table 5. Model-Comparison Results: FT, FF, ABT + IF

Model	G^2	df	*p*	ΔFIA
$M_{saturated}$	0	0	1	40.98
M_{same}	243.73	36	.00	70.03
$M_{different}$	39.55	24	.02	0

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

Indicative Conditionals and Truth Tables

We now turn to an evaluation of the fit of truth tables in Table 1 for the indicative conditionals. As said, we will allow the models to predict that the expected responses constitute at least a relative majority (a very lenient requirement). The results reported in Table 6 show that none of the models was able to accurately characterize the individuals' responses. Overall, the modal responses indicate a slight tendency to judge indicative conditionals as true whenever the antecedent and the consequent have the same truth status (in accordance with the truth table of the material bi-conditional, which is true in the ⊤⊤ and ⊥⊥ cells and false otherwise). This pattern is corroborated by the set of studies gathered by Schroyens (2010), which involved abstract stimulus materials with explicit negations and the option to respond that the truth table cell is 'irrelevant' for the truth value of the conditional (see Table 7).

Table 6. Goodness-of-Fit Results

	Truth-Conditional Inferentialism			
	⊤⊤	⊤⊥	⊥⊤	⊥⊥
PO	0.00	193.88 (⊥)	171.94 (⊥)	0.00
NE	73.12 (⊤)	0.00	0.00	0.00
IR	107.41 (⊤)	0.00	0.00	0.00
	Material Implication Account			
PO	0.00	0.00	171.94 (⊥)	0.00
NE	0.00	0.00	96.58 (⊥)	12.64 (⊥)
IR	0.00	0.00	147.56 (⊥)	2.09 (⊥)
	De Finetti Table			
PO	0.00	0.00	43.44 (⊥)	10.70 (⊤)
NE	0.00	0.00	40.47 (⊥)	30.92 (⊥)
IR	0.00	0.00	28.31 (⊥)	5.87 (⊥)

Note. Values correspond to the G^2 statistic. The symbols in parentheses (⊤, ⊥, N) indicate the observed modal response in case the model prediction failed. All values above 2.71 are rejected ($p < .05$) according to the most conservative $\bar{\chi}^2$ distribution (Self & Liang, 1987).

Table 7. Data from Schroyens' (2010) Meta-Analysis

	⊤⊤	⊤⊥	⊥⊤	⊥⊥
⊤	100%	0%	5%	64%
⊥	0%	100%	77%	0%
Irrelevant	0%	0%	18%	36%

Note. Values correspond to percentage of studies (out of a subset of 22 studies with abstract stimulus material and 'irrelevant' as third response option) in Schroyens' (2010) meta-analysis in which a specific response was modal.

Confidence Ratings

All the confidence ratings were in the interval [76%, 81%]. For the 'therefore'-sentences, the participants were more confident in the PO (mean = 80.28, SD = 19.54) condition than in the NE (mean = 78.27, SD = 21.85), V = 60142, $p_H < .05$, $r = -.11$,[4] and IR (mean = 75.97, SD = 22.60) conditions, V = 66528, $p_H < .0001$, $r = -.21$. But these were small effects and the participants continued to remain highly confident in the truth values they provided even when these conflicted with the reason-relation readings of the 'therefore'-sentences. For the 'but'-sentences, the participants were no more confident in the NE (mean = 78.48, SD = 21.60) condition than they were in the PO (mean = 78.14, SD = 21.89) condition, V = 54333, $p_H = .53$, $r = -.027$. For the indicative conditionals, the participants were more confident in the PO (mean = 78.74, SD = 20.65) condition than they were in the IR (mean = 76.01, SD = 22.53) condition, V = 61801, $p_H < .05$, $r = -.12$, but no more confident than they were in the NE (mean = 76.42, SD = 21.90) condition, V = 63538, $p_H = .074$, $r = -.089$. Again, these were small effects and the participants continued to remain highly confident in the truth values they provided even when these conflicted with the reason-relation readings of indicative conditionals.

[4] We controlled for the family-wise error rate using the Bonferroni-Holm correction (indicated by the index "H").

Discussion

The results of Experiment 1 indicate that there are significant effects of relevance on response probabilities in truth evaluations for each of the ABT sentences. But these effects are minor judging from Figure 1. It is possible that some of the small differences observed are due to heterogeneity at the individual level as well as at the level of the stimulus material (e.g., Rouder *et al.*, 2008). Moreover, the model comparison in Table 3 shows that the preferred model neither assumes a difference between these three sentence types, nor in terms of how they are affected by the relevance manipulation. Accordingly, the results do not support the idea that conventional implicatures ("therefore" conveying PO; "but" conveying NE) affect truth-table evaluations. Instead a common truth-table semantics of the ABT emerged in line with H_{0_ABT}, as derived from the Frege-Grice tradition, and effects of relevance on their shared truth table were small.

This finding was also supported by the high confidence ratings across conditions. Even in conditions where conflicts between truth evaluations based on logical conjunction and the reason-relation reading of the ABT sentences were induced, the participants indicated that they had a confidence of 76% or higher on average in their truth-value judgments.

In addition, the results indicate that none of the truth tables for indicative conditionals outlined in Table 1 are able to capture the patterns in the data for indicative conditionals, even when applying the most lenient test (i.e. that the deterministic truth tables only have to predict the relative, rather than absolute, majority responses). As shown in Table 6, the de Finetti table and the material implication are much better suited to capture the modal responses in the true antecedent cases than truth-conditional inferentialism.[5] In comparison, truth-conditional

[5] As noted in Footnote 2, Karolina Krzyżanowska has in discussion expressed doubts about whether the truth conditional inferentialism is really committed to predicting 'True' for the ⊤⊥ cell in the positive relevance condition—not least due to the context-sensitive interpretation of

inferentialism does a much better job in the false antecedent cases, where the de Finetti table and the material implication encountered difficulties. Indeed, although the defective truth table effect is often cited in the literature as strong evidence in favor of the de Finetti table (Over & Evans, 2003), neither the results reported in Table 6 with realistic stimulus material nor the results reported in Table 7 with abstract stimulus materials support the assumption that ‘Neither true nor false’ is the modal response in the false antecedent cases.

Interestingly, Figure 1 does indicate the presence of a relevance effect on the truth evaluations of the indicative conditional in the $\bot\bot$ cell, which is compatible with the truth table for truth-conditional inferentialism (see Table 1).

However, the gross failure of the predictions in Table 1 to account for the data of Experiment 1 suggests the possibility of some kind of within-subject variation with respect to the truth tables the participants rely on. Indeed, it is possible that a mixture of truth tables (*inter alia*, the bi-conditional table and the de Finetti table) would have to be invoked to account for our results. However, the present data suggest that it would have to be a small proportion of the individuals that follow the de Finetti table. The proportion of ‘NN’ responses in the false-antecedent cells for ‘if-then’ sentences hardly exceeds 33%, suggesting that no more than a third of the individuals in these cells of the experimental design responded in accordance with the de Finetti table. Moreover, such a mixture account is not able to account for the relevance effect found for the $\bot\bot$ cell, captured by truth-conditional inferentialism. Follow-up studies better suited for testing individual variation that have the participants fill out all four truth table cells

the indicative conditional voiced in Krzyżanowska *et al.* (2014). However, even when taking this point into account, it is not clear to how the theory could be adjusted to successfully accommodate an absolute majority of ‘True’ responses in the $\top\top$ cell for the NE and IR conditions.

for each sentence (perhaps multiple times in order to estimate response-error probabilities) would be needed to investigate this possibility.

Experiment 2

Given that no difference was found among the ABT sentences in Experiment 1, we wanted to see in Experiment 2 whether a dissociation between these sentences occurs when they are evaluated in the context of a probability-judgment task and an acceptability-ranking task. According to the reason-relation reading, 'φ but ψ' expresses that φ lowers the probability of ψ ($\Delta p < 0$), and 'φ therefore ψ' expresses that φ raises the probability of ψ ($\Delta p > 0$). In contrast, 'φ and ψ' can suggest that φ raises the probability of ψ, but according to its reading as logical conjunction 'φ & ψ', φ need not affect the probability of ψ at all.

Hence, we expected that when presented with the TT cell, the acceptability ratings would accord with the following pattern, where 'φ and ψ' acts as a baseline:

(NE)	φ but $\psi \succ \varphi$ and $\psi \succ \varphi$ therefore ψ	($b \succ a \succ t$)
(PO)	φ therefore $\psi \succ \varphi$ and $\psi \succ \varphi$ but ψ	($t \succ a \succ b$)
(IR)	φ and ψ / φ but $\psi \succ \varphi$ therefore ψ	($b/a \succ t$)

Moreover, on the assumption that 'φ but ψ' expresses that φ is assumed to be a sufficient reason *against* ψ, and that 'φ therefore ψ' expresses that φ is assumed to be a sufficient reason *for* ψ, we would expect $P(\psi \mid \varphi)$ = high/low to act as a moderator variable. That is to say, we expect the pattern $b \succ a \succ t$ to be more frequent in NE when $P(\psi \mid \varphi)$ = low compared to $P(\psi \mid \varphi)$ = high, and $t \succ a \succ b$ to be more frequent in PO when $P(\psi \mid \varphi)$ = high as compared to $P(\psi \mid \varphi)$ = low.

As a manipulation check, we tested whether the effect of the relevance manipulation on $P(\psi \mid \varphi)$ as a predictor of $P(\text{if } \varphi, \text{then } \psi)$ from Skovgaard-Olsen *et al*. (2016a) replicates despite the procedural change that our conditionals had the form of ‘if φ is *now happening*, then ψ will occur’ as opposed to ‘if φ *occurs*, then ψ will occur’ (see Footnote 3). In addition, we tested whether a similar moderation of P(φ & ψ) as a predictor of P(φ but ψ), P(φ and ψ), and P(φ therefore ψ) could be found for our ABT sentences with the expectation that the marginal means would be higher in the NE condition compared to the PO condition for P(φ but ψ), and that the marginal means of P(φ therefore ψ) would be higher in the PO condition compared to the NE and IR conditions.

Method

Participants

Like Experiment 1, the experiment was conducted over the Internet. A total of 805 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk from USA, UK, and Australia and paid a small amount of money for their participation. The same exclusion criteria were applied as in Experiment 1. The final sample thus consisted of 593 participants. Mean age was 39 years, ranging from 18 to 80 years; 32% of the participants were male; 73% indicated that the highest level of education that they had completed was an undergraduate degree or higher. The demographic measures of the participants differed only minimally before and after the exclusion.

Design

Experiment 2 had the same experimental design as Experiment 1 for the probability task. In contrast, the acceptability task only differed by presenting the participants with three levels of the sentence factor (‘…and…’, ‘…but…’, ‘…therefore…’).

Materials and Procedure

The procedures were the same as in Experiment 1 unless otherwise stated. For each of the twelve priors×relevance within-participants conditions, the participants were presented with four pages. The first page featured only the scenario text. The participants were instructed that the scenario text had been supplied as input to a computer program in the development phase (following the instructions from Experiment 1). The second page asked the participants both to evaluate the probability of the antecedent (e.g. ‘Scott is now turning on the warm water’) and of the consequent (e.g. ‘Scott will be warm soon’) conditional on the antecedent on a slider with a scale from 0 to 100%. The instruction for evaluating the conditional probability was as follows:

Suppose Scott is now turning on the warm water.

Under this assumption, how probable is it that the following sentence is true on a scale from 0 to 100%:

Scott will be warm soon.

The third page asked the participants to evaluate the probability of a randomly chosen member of our four target sentences (‘φ and ψ’, ‘φ but ψ’, ‘φ therefore ψ’, ‘if φ, then ψ’). Continuing with our example of a PO HH condition from the Scott scenario, one participant might be asked to evaluate the probability of the ‘but’ sentence as follows:

Could you please rate the probability that the following sentence is true on a scale from 0 to 100 %:

Scott is now turning on the warm water, BUT Scott will be warm soon.

On page four the participants were presented with the acceptability task. Inspired by the task of evaluating the categorical acceptability of conditionals in Douven & Verbrugge (2012), we introduced the novel task of rank-ordering the acceptability of the ABT sentences given the TT cell with the computer program calibration instructions from Experiment 1. That is to say, the participants were presented with the scenario text, which they had been instructed to regard as input to a computer program. They were then presented with the TT cell as a continuation of the scenario, which took place after the computer program had produced its output sentences. The task was to evaluate which of the three ABT sentences was most acceptable in light of the continuation of the scenario. Continuing with the example from above:

> *Continuation: Scott turned on the warm water. Scott did get warm.*
>
> *Please order the OUTPUT in terms of how acceptable they are in light of the continuations of the scenarios. Click on the most acceptable output for rank 1, the second most acceptable for rank 2, and the third most acceptable for rank 3. Note that the responses can be deselected.*

```
Output: Scott is now turning on the warm water, BUT Scott will be warm soon.
Output: Scott is now turning on the warm water AND Scott will be warm soon.
Output: Scott is now turning on the warm water THEREFORE Scott will be warm soon.
```

As in Experiment 1, the computer output sentences were distinguished by a different font to help the participants organize the information. Finally, the participants were asked to indicate whether they agree with the statement that at least one of the output sentences was acceptable given the continuation.

The instructions presented after the practice trial followed Skovgaard-Olsen *et al.* (2016a) in giving the following explication of how ‘acceptable’ was meant to be understood:

Please note that when we ask – here and throughout the study – how ‘acceptable’ a statement is, we are not interested in whether the statement is grammatically correct, unsurprising, or whether it would offend anybody. Rather we ask you to make a judgment about the adequacy of the information conveyed by the statement. More specifically, we ask you to judge whether the statement would be a reasonable thing to say in the context provided by the scenarios and their continuations.

Results

Acceptability

We excluded rank orders for which participants found none of the output sentences to be categorically acceptable (24%).[6] Table 8 reports the rank order of the sentences from the acceptability task. Overall, the results matched our predictions, with the rank order $t \succ a \succ b$ occurring most often in PO, $b \succ a \succ t$ occurring most often in NE, and $a \succ b \succ t$ occurring most often in IR. Another prediction corroborated by the data was that in PO, $t \succ a \succ b$ occurred less often when the participants judged $P(\psi \mid \varphi)$ to be low. Indeed, the proportion of $t \succ a \succ b$ ranks was larger (66%) when $P(\psi \mid \varphi) \geq .50$ than when $P(\psi \mid \varphi) < .50$ (53%), a difference that was found to be statistically significant ($\Delta G^2 = 25.51$, $p < .001$, $\Delta\text{FIA} = 10.29$). Conversely, we predicted that in NE, $b \succ a \succ t$ would occur more often when $P(\psi \mid \varphi)$ was judged to be low rather than high, a difference that was also found in the data (74% versus 66%; $\Delta G^2 = 9.85$, $p = .001$, $\Delta\text{FIA} = 2.44$).

[6] Note that the ranking distributions did not differ qualitatively, when including rankings for which participants found none of the sentences to be acceptable. However, the increase of $b \succ a \succ t$ in the NE condition when $\mathrm{P}(\psi \mid \varphi) < .50$ was no longer significant ($p = .06$).

Table 8. Percentage of rank orders in the acceptability task

	φ/ψ Priors	$a \succ b \succ t$	$a \succ t \succ b$	$b \succ a \succ t$	$t \succ a \succ b$	$b \succ t \succ a$	$t \succ b \succ a$
PO	HH	3%	31%	1%	**62%**	1%	2%
	HL	5%	28%	5%	**54%**	4%	4%
	LH	1%	23%	1%	**71%**	2%	3%
	LL	4%	19%	5%	**64%**	3%	4%
NE	HH	8%	4%	**73%**	5%	9%	1%
	HL	6%	4%	**74%**	5%	8%	3%
	LH	10%	10%	**66%**	7%	6%	1%
	LL	7%	3%	**74%**	4%	10%	3%
IR	HH	**49%**	35%	6%	8%	1%	1%
	HL	**52%**	19%	23%	3%	1%	1%
	LH	**57%**	19%	15%	4%	2%	2%
	LL	**53%**	22%	14%	7%	2%	2%

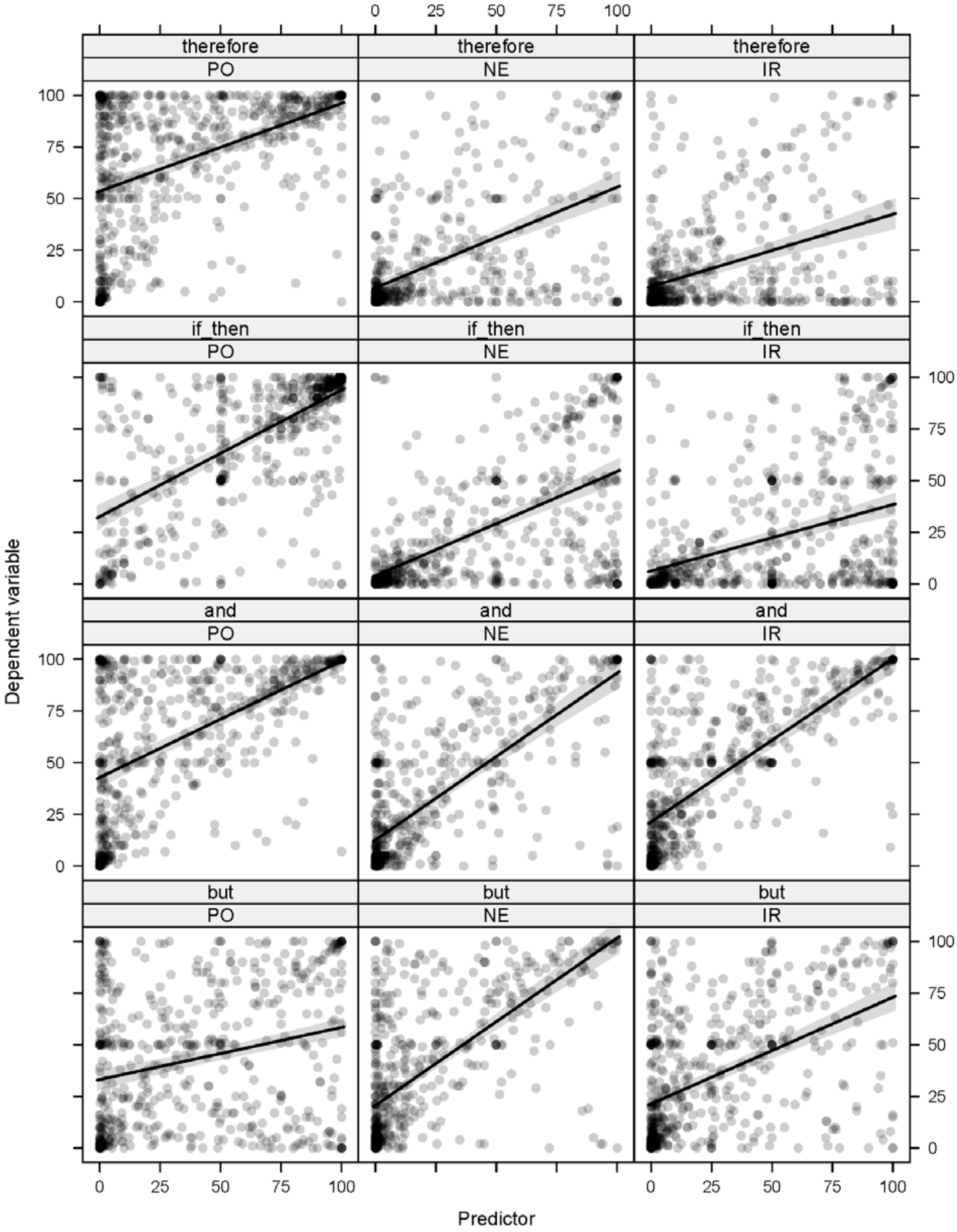
Note. φ = the antecedent (or first conjunct); ψ = the consequent (or second conjunct). The operator ≻ denotes "more acceptable than". 'a' = 'and', 'b' = 'but', and 't' = 'therefore'.

The interactions between sentence levels and relevance levels in determining the preferred rank orders shown in bold in Table 8 are of such magnitude that a statistical analysis is superfluous; their prominence is too severe to leave any doubt.

Probability Judgments

According to Figure 2 (see panels in the second row from the top), the results from Skovgaard-Olsen *et al.* (2016a) for P(ψ | φ) as a predictor of P(if φ, then ψ) across relevance levels appear to be replicated. In comparison to Figure 1 in Skovgaard-Olsen *et al.* (2016a), the intercept is slightly larger and less points fall on the diagonal in the present Figure 2. But the overall tendency is the same: There is a stronger relationship between P(ψ | φ) and P(if φ, then ψ) for the PO condition than for the IR condition in particular and the marginal mean (i.e. the overall level of the set of points on the y axis) of P(if φ, then ψ) is higher in the PO condition than in the NE and IR conditions, across the scale of P(ψ | φ). For the IR and NE conditions there is a substantial portion of data points assigning probabilities almost equal to zero to 'if φ, then ψ' across variations in P(ψ | φ). This same tendency is visible in the probability evaluations in Figure 2 (top row) for 'φ therefore ψ' with P(φ & ψ) as a predictor; only this time the differences in intercept and marginal means between PO and the IR and NE conditions appear to be even more pronounced. In Figure 2 (third row from top), it moreover seems that P(φ & ψ) is a good predictor of P(φ and ψ) across relevance levels, with a higher intercept in the PO condition. Finally, P(φ & ψ) appears to best predict P(φ but ψ) for the NE condition. In contrast, in the PO condition there appears to be only a very weak relationship between P(φ & ψ) and P(φ but ψ).

Figure 2. Probability evaluations of the sentence types across relevance levels



Note. P(φ & ψ) is plotted as a predictor of P(φ therefore/and/but ψ) across relevance levels (rows 1, 3, and 4). P(ψ | φ) is plotted as a predictor of P(if φ, then ψ) across relevance levels (row 2). Raw data values (plotted with 80% transparency) and LMM linear effect of the predictors on P(φ therefore ψ) (row 1), P(if φ, then ψ) (row 2), P(φ and ψ) (row 3), and P(φ but ψ) (row 4) across relevance manipulations (PO = left column, NE = center column, IR = right column). The confidence bands show the 95% confidence region of the effect of the two independent variables, P(φ & ψ) and P(ψ | φ).

This pattern was confirmed in an LMM analysis (see Appendix for details). Main effects of P(ψ | φ) as predictor of P(if φ, then ψ) and P(φ & ψ) as predictor of the ABT sentences were found, F(1, 18.4) = 1095.21, $p < .0001$. Also, a main effect of the relevance conditions was found, F(2, 12.3) = 99.24, $p < .0001$, as well as an interaction between the P(ψ | φ)/P(φ & ψ) predictors and the relevance conditions, F(2, 13) = 10.20, $p < .01$. Main effects for the sentence type, and interactions between sentence type and the other predictors, were also found.

Follow-up analyses for P(ψ | φ) as a predictor of P(if φ, then ψ) showed that the slope found in the PO condition ($b = 0.61$, 95%-CI = [0.54, 0.69]), was significantly larger than the slope in the IR condition ($b = 0.32$, 95%-CI = [0.26, 0.39]), z = 5.99, $p_T < .001$, but was not significantly larger than the slope in the NE condition ($b = 0.51$, 95%-CI = [0.43, 0.58]), z = 2.02, $p_T = .68$. The subscript ‘T’ indicates p-value correction by the Tukey method for comparing a family of 12 estimates for all follow-up analyses.

Follow-up analysis for P(φ & ψ) as a predictor of P(φ therefore ψ) showed that the slope in the PO condition ($b = 0.43$, 95%-CI = [0.36, 0.49]) was neither significantly larger than the slope in the IR condition ($b = 0.35$, 95%-CI = [0.27, 0.43]), z = 1.40, $p_T = .97$, nor significantly smaller than the slope in the NE condition ($b = 0.49$, 95%-CI = [0.40, 0.58]), z = -1.11, $p_T = .99$.

Furthermore, follow-up analyses for P(φ & ψ) as a predictor of P(φ and ψ) showed that the slope in the PO condition ($b = 0.56$, 95%-CI = [0.49, 0.64]) was both significantly smaller than the slope in the IR condition (*b* = 0.79, 95%-CI = [0.71, 0.87]), z = -4.30, p_T < .01, and significantly smaller than the slope in the NE condition (*b* = 0.81, 95%-CI = [0.72, 0.90]), z = -4.15, p_T < .01.

Finally, a follow-up analysis for P(φ & ψ) as a predictor of P(φ but ψ) showed that the slope in the PO condition (*b* = 0.25, 95%-CI = [0.18, 0.32]) was both significantly smaller than the slope in the IR condition (*b* = 0.52, 95%-CI = [0.44, 0.59]), z = -5.07, p_T < .0001, and significantly smaller than the slope in the NE condition (*b* = 0.81, 95%-CI = [0.72, 0.91), z = -9.24, p_T < .0001.

Differences in estimated marginal means for the four target sentences were tested for their statistical significance at the three scale points of the independent variables (P(ψ | φ) for P(if φ, then ψ) and P(φ & ψ) for P(φ and/but/therefore ψ)) in Table 9.

Table 9. Differences in Estimated Marginal Means

Sentence	Relevance	IV = 0%		IV = 50%		IV = 100%	
P(if φ, then ψ)	PO	32.6%		63.3%		93.9%	
	NE	4.0%	***	29.2%	***	54.4%	***
	IR	6.3%	***	22.4%	***	38.4%	***
P(φ therefore ψ)	PO	53.6%		79.9%		96.1%	
	NE	6.8%	***	31.2%	***	55.6%	***
	IR	7.4%	***	24.9%	***	42.4%	***
P(φ and ψ)	PO	42.9%		71.1%		99.3%	
	NE	12.6%	***	53.0%	***	93.5%	
	IR	21.3%	***	60.8%	*	100%	
P(φ but ψ)	PO	33.2%		45.8%		58.4%	
	NE	20.6%	*	61.2%	***	100%	***
	IR	21.4%	.	47.2%		73.0%	

Note. Differences for the Estimated Marginal Means for the four sentence levels were tested for their statistical significance across the following scale points of the independent variables (IV = P(ψ | φ) for P(if φ, then ψ) and IV = P(φ & ψ) for P(φ and/but/therefore ψ)): 0%, 50%, and 100%. The pairwise contrasts indicate whether the NE or IR conditions differed significantly from the PO condition using z-ratios and adjusted *p*-values through Tukey's method for comparing a family of 36 estimates.
Signif. codes: '***' .001, '**' .01, '*' .05, '.' .1.

As Table 9 shows, the estimated marginal means of P(if φ, then ψ) and P(φ therefore ψ) were consistently statistically higher in the PO condition than in the NE and IR conditions across the probability scales of their respective independent variables. In contrast, the estimated marginal means of P(φ and ψ) was only significantly higher in the PO condition than in the NE and IR conditions at P(φ & ψ) = 0% and P(φ & ψ) = 50%. And finally, the estimated marginal means of P(φ but ψ) was only significantly higher in the PO condition than in the NE condition at P(φ & ψ) = 0%. For P(φ & ψ) = 50% and P(φ & ψ) = 100%, the estimated marginal means of P(φ but ψ) were significantly higher in the NE condition than in the PO condition.

Discussion

The results obtained in Experiment 2 show an unmistakable pattern. Participants rank ordered the acceptability of the ABT sentences given the continuation of the scenario in the TT cell. We predicted that 'φ but ψ' differs from the 'φ and ψ' baseline in being more preferable in NE, and that 'φ therefore ψ' differs from the 'φ and ψ' baseline in being more preferable in PO. The predicted rank orders clearly dominate the other possible rank orders in the data. Furthermore, P(ψ | φ) moderates the relationship between the rank order acceptabilities and the relevance levels in the manner one would expect, if 'φ but ψ' expresses that φ is a sufficient reason against ψ and 'φ therefore ψ' expresses that φ is a sufficient reason for ψ.

This basic finding is corroborated by the results from the probability evaluation task outlined in Table 9. Consistent with the reading of 'if φ, then ψ' and 'φ therefore ψ' as indicating that φ is a reason *for* ψ, the estimated marginal means for P(if φ, then ψ) and P(φ therefore ψ) in the PO condition were invariantly higher than the estimated marginal means in the NE and IR conditions, across the scale of their respective independent variables. Moreover, consistent with the reading of 'φ but ψ' as indicating that φ is a reason *against* ψ, the estimated marginal means for P(φ but ψ) was higher in the NE condition than in the PO condition at P(φ & ψ) = 50% and P(φ & ψ) = 100%.

General Discussion

Given the large sample sizes and high power for detecting even small differences, the finding from Experiment 1 that the (relative) response frequencies of the ABT sentences can be set equal across sentence levels, and that relevance does not interact with the sentence factor, is a strong and in fact surprising result. It suggests that the ABT sentences have exactly the same truth conditions. Taken together, Experiment 1 and 2 show a clear dissociation between evaluations of truth, probability, and acceptability. Relevance interacts with the sentence factor (and, but, therefore) in the rank-order acceptability task and the probability evaluation task as expected, but does not interact with the sentence factor in the truth-evaluation task. In fact, truth evaluations of the ABT sentences could be set equal across sentences in that task with only minor losses in goodness of fit. The fact that Table 8 and 9 display such strong interaction effects, when the participants are asked to rank order the acceptability and provide probability evaluations, makes the absence of an interaction of relevance and sentence in the truth-table task even more surprising. It suggests that there is a deeply entrenched modularization and little cross-talk between the processes and/or representations tapped by the tasks in Experiments 1 and 2.

These findings suggest that the Frege-Grice tradition was right in its assumption that the difference in reason-relation readings of these sentences does not affect their truth evaluations. Instead, Grice (1989) conjectured that the differences of the ABT sentences would be part of their conventional implicatures. In support of this, Experiment 2 found the signature effects of the reason-relation readings in the orderings of the ABT sentences according to their acceptability, as also corroborated in the probability judgments.

Finally, the fact that we were able to find strong effects of the relevance manipulation in the expected directions for the ABT sentences in Experiment 2 suggests that the absence of such a difference in Experiment 1 is not an artifact of the stimulus material (see Footnote 3) nor of the instructions of the computer-calibration task.

Turning to indicative conditionals, the results from Experiment 1 on the truth evaluations of the indicative conditional present an explanatory challenge, as none of the investigated truth tables were able to account for the patterns found. We found that there is a marked relevance effect on the truth evaluations of the indicative conditional in the ⊥⊥ cell. This finding challenges the H_{0_IF} assumption underlying the material implication account and the de Finetti truth table, which holds that truth evaluations of the indicative conditional are not affected by relevance manipulations. Moreover, it was shown that in spite of its recent popularity in psychology of reasoning (Baratgin, Politzer, and Over, 2013; Pfeifer, 2013; Elqayam and Over, 2013), the de Finetti truth table lacks support for its distinguishing feature: The prediction of 'Neither true nor false' evaluations in the false antecedent cases neither found support in Experiment 1 nor in the 22 studies with abstract stimulus material from the Schroyens (2010) meta-analysis presented in Table 7.

In contrast, truth-conditional inferentialism correctly rejects H_{0_IF}, but there is little support for its alternative predictions. In particular, truth-conditional inferentialism faced problems accounting for our results in the true antecedent cells, yet it was compatible with the relevance effect found in the ⊥⊥ cell.

It is remarkable that none of the investigated theories was able to fit the data even under the most lenient stochastic interpretation of their deterministic truth tables, whereby they only had to predict the relative-majority responses of the participants (see the Appendix for further discussion). A stricter interpretation would have imposed the requirement that the absolute majority responses were predicted by the theories, or by admitting relatively small "error" rates (e.g., 10%). As noted, it is possible that the data pattern of truth evaluations of the indicative conditional across relevance manipulations is best accounted for by a mixture of truth tables. It is up to future experiments better suited to testing for individual variation to explore this possibility.

However, a different perspective on our results is also possible. A non-truth functional conditional is a conditional whose truth value cannot be determined by the truth values of its parts. As Rescher (2007, p. 43) points out, for any non-truth functional conditional that is logically stronger than '⊃', it holds that "we can say nothing about the truth status of $p \rightarrow q$ without a deeper look at the specifics of the matter". It is therefore possible to interpret our results as indicating that indicative conditionals are non-truth functional with further parameters determining their truth values in the false antecedent cases. If the participants set these parameters differently, then mixed results, like those found in Figure 1, may be the outcome.

Interestingly, David Over (personal communication, 2017) suggested one such possibility. According to the Jeffrey truth table, the value in the ⊥⊥ cells is not NN but rather P(ψ | φ) (Jeffrey, 1991; Over and Baratgin, 2017). Conjoined with the auxiliary hypothesis that some participants interpret 'true' pleonastically as simply expressing endorsement (Edgington, 2003; Over *et al.*, 2007), these participants might respond with ⊤ when P(ψ | φ) = high, NN when P(ψ | φ) = middle and with ⊥ when P(ψ | φ) = low. In addition, Evans & Over (2004) have suggested that there may be conditions in which the participants supplement ‘if φ, then ψ’ with ‘if ψ, then φ’ yielding a bi-conditional event interpretation, which in turn could also be interpreted according to the Jeffrey table (Over and Baratgin, 2017). Further work is needed to test these intriguing possibilities.

Potential Limitations

When evaluating our results it is important to keep the following points in mind. First, when asking the participants to calibrate the output of a computer program, the aim was to create a situation in which the participants would not naturally begin interpreting communicative intentions behind producing assertions with irrelevant components and reconstruct speaker meaning. It turns out that such an approach has a track-record in the literature. For instance, Schwarz, Strack, Hilton, and Naderer (1991) employ a conceptually similar manipulation to set aside Gricean conversational norms (for discussion see Lee, 2006). Moreover, Doran, Ward, Larson, McNabb, and Baker (2012) achieved a similar effect by asking the participants to provide truth-value judgments based on adopting the perspective of someone, who is only able to understand the literal content of what has been said. Finally, in Wright and Wells (1988) an attempt was undertaken to control for demand characteristics relating to the Gricean ideal of cooperative discourse in the attitude-attribution paradigm by instructing the participants that the set of questionnaire items they were presented with had been randomly selected from a larger pool.

In Doran *et al.* (2012) confidence ratings are moreover used as a measure of task complexity, because confidence ratings have been found to be inversely correlated with perceived task complexity. In the truth-table task in Experiment 1, all mean confidence ratings were in the interval [76%, 81%], which indicates that the participants were highly confident of judgments across all conditions. Indeed, the participants continued to remain highly confident of their responses even when judging truth values based on logical connectives and reason relation constraints would have led to conflicting responses.

For the rank-ordering of acceptability and probability evaluations in Experiment 2, we retained the computer program instruction to discourage participants from looking for some hidden intention for why an irrelevance item had been asserted and to encourage a focus on rating the sentences for their acceptability and probability under different conditions.

Note, finally, that the differences between the findings in Experiment 1 and Experiment 2 cannot merely be attributed to the presence of a forced choice format or dependencies in the rank ordering task, since the probability evaluation task in Experiment 2 produced similar results without having a forced-choice format. Moreover, even if one restricts the attention to the most preferred sentence in each of the relevance conditions, because one is worried that the choices are not independent, a very clear pattern emerges: in the PO condition, the therefore-sentences are by far the most acceptable sentences, in the NE condition, the but-sentences are by far the most acceptable sentences, and in the IR condition the and-sentences are by far the most acceptable sentences. This pattern is in agreement with the reason-relation reading of these sentences and it is also one that is mirrored in the probability evaluations.

Conclusion

In the Frege-Grice tradition of applying logic as a model of natural language connectives, it is assumed that the difference in reason-relation readings of the sentences ‘φ and ψ’, ‘φ but ψ’, and ‘φ therefore ψ’ does not affect their truth conditions. Support for this assumption was found in Experiment 1. In contrast, Experiment 2 indicated a dissociation between the effects of relevance on ‘φ and ψ’, ‘φ but ψ’, and ‘φ therefore ψ’ in truth evaluations and in evaluations of their acceptability and probabilities. For, when the participants are asked to rank order the acceptability of ‘φ and ψ’, ‘φ but ψ’, and ‘φ therefore ψ’ on the basis of the TT cell, the different rank orders predicted by a reason-relation reading of each sentence are strongly preferred. Moreover, their probability evaluations accord with the reason-relation reading. Turning to the indicative conditionals, a relevance effect on truth evaluations was found, and neither the truth tables supplied by the material implication, the popular de Finetti table, nor truth-conditional inferentialism were able to account for these results of Experiment 1, even under the most lenient stochastic interpretation of their predictions. Accordingly, the results for the truth evaluations of indicative conditionals across relevance levels from Experiment 1 present an explanatory challenge for further theorizing and empirical work to solve.

References

Baratgin, J., Politzer, G., & Over, D. E. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning*, 19(3), 308–328. http://dx.doi.org/10.1080/13546783.2013.809018.

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

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*. URL = http://arxiv.org/abs/1406.5823

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

Birnbaum, M. H. (2013). True-and-error models violate independence and yet they are testable. *Judgment and Decision Making, 8,* 717-737.

Blackmore, D. (2004), *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York: Springer.

Doran, R., Ward, G., Larson, M., McNabb, Y., and Baker, R. E. (2012). A Novel Experimental Paradigm for Distinguishing between What is Said and What is Implicated. *Language*, 88(1), 124-54. doi: 10.1353/lan.2012.0008

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

Douven, I. (2015b). How to account for the oddness of missing-link conditionals. *Synthese*, 1-14. doi:10.1007/s11229-015-0756-7

Douven, I. and Verbrugge, S. (2012). Indicatives, concessives, and evidential support. *Thinking and Reasoning* 18 (4), 480-99. doi: 10.1080/13546783.2012.716009

Edgington, D. (2003). What if? Questions about conditionals. *Mind & Language*, 18, 380-401.

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

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift fur Psychologie / Journal of Psychology, 217*, 108–124.

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

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

Garmut, L. T. F. (1991). *Logic, Language, and Meaning, Vol 1*. Chicago: The University of Chicago Press.

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

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, Mass: MIT Press.

Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, 21, 1431-1443. doi: 10.3758/s13423-014-0643-0

Iten, C. B. (2000). *'Non-Truth-Conditional' Meaning, Relevance and Concessives*. Doctoral thesis, University of London. URL = < http://discovery.ucl.ac.uk/1348747/1/324676.pdf>

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

Johnson-Laird, P. N. and Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678. http://dx.doi.org/10.1037//0033-295X.109.4.646

Karabatsos, G. (2005). The exchangeable multinominal model as an approach for testing axioms of choice and measurement. *Journal of Mathematical Psychology, 49*, 51-69.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition Memory Models and Binary Response ROCs: A Comparison by Minimum Description Length. *Psychonomic Bulletin & Review*, 20, 693-719.

Klauer, K. C. & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, 55, 430-50.

Klauer, K. C. & Kellen, D. (2015). The Flexibility of Models of Recognition Memory: The Case of Confidence Ratings. *Journal of Mathematical Psychology*, 67, 8-25.

Klauer, K. C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in Multinominal Processing Tree Models: An extension of Knapp & Batchelder (2004). *Journal of Mathematical Psychology*, 64, 1-5.

Krzyżanowska, K., Wenmackers, S. and Douven, I. (2014). Rethinking Gibbard's Riverboat Argument. *Studia Logica*, 102 (4), 771-92.

doi: 10.1007/s11225-013-9507-2.

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

URL = http://karolinakrzyzanowska.com/pdfs/krzyzanowska-phd-final.pdf

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

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46*, 1–26.

Luce, R. D. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology, 41*, 79–87.

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

McCawley, J: (1993): *Everything that Linguists have Always Wanted to Know About Logic*. Second Edition. Chicago, University of Chicago Press.

Nickerson, R. S. (2015). *Conditionals and Reasoning*. Oxford: Oxford University Press.

Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. Thinking & Reasoning, 13(4), 340–369. http://dx.doi.org/10.1080/13546780601035794.

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

Over, D. E., & Baratgin, J. (2017). The "defective" truth table: Its past, present, and future. In N. Galbraith, E. Lucas, & D. E. Over (Eds.), *The Thinking Mind: A Festschrift for Ken Manktelow* (pp. 15-28). Abingdon, UK: Routledge.

Over, D. and Evans, J. St. B. T. (2003). The Probability of Conditionals: The Psychological Evidence. *Mind and Language*, 18 (4), 340-58. doi: 10.1111/1468-0017.00231

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

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning* 19(3-4), 329-45. doi: 10.1080/13546783.2013.838189

Pfeifer, N. and Douven, I. (2014). Formal epistemology and the new paradigm psychology of reasoning. *The Review of Philosophy and Psychology*, 5(2), 199-221. doi: 10.1007/s13164-013-0165-0

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

Read, T., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology,* 49 (4), 243-256. doi: 10.1027//1618-3169.49.4.243

Rescher, N. (2007). *Conditionals*. Cambridge, MA: The MIT Press.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339.

Rouder, J. N., Lu J., Morey R. D., Sun D., & Speckman P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137,* 370-389.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8. Available from Methods of Psychological Research Online, http://www.mpronline.de

Schroyens, W. (2010). A meta-analytic review of thinking about what is true, possible, and irrelevant in reasoning from or reasoning about conditional propositions. *European Journal of Cognitive Psychology*, 22 (6), 897-921. doi: 10.1080/09541440902928915

Schwarz, N., Strack, F., Hilton, D. and Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of “irrelevant” information. *Social Cognition,* 9 (1): 67-84

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605-610. doi : 10.1080/01621459.1987.10478472

Singmann, H., B. Bolker, J. Westfall, S. Højsgaard, J. Fox, M. Lawrence, et al. (2016). afex: Analysis of Factorial Experiments. R package version 0.13-145, Available via http://cran.rproject.org/package=afex.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of Multinominal Processing Tree models with R. *Behavior Research Methods*, 45, 560-575.

Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. Frontiers in Psychology, 5, 316. http://dx.doi.org/10.3389/fpsyg.2014.00316.

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

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016b). Relevance and Reason Relations. *Cognitive Science*. doi: 10.1111/cogs.12462

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

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, 37, 1074–1106. doi: 10.1111/cogs.12057

Wright, E. F. and Wells, G. L. (1988). Is the Attitude-Attribution Paradigm Suitable for Investigating the Dispositional Bias? *Personality and Social Psychology Bulletin*, 14 (1), 183-190.

Appendix: Statistical Analyses

Experiment 1: MPT Analysis

The observed response frequencies in Experiment 1 were analyzed with multinomial processing tree models (MPT; Riefer & Batchelder, 1988). The Models were fitted with R package `MPTinR` (Singmann & Kellen, 2013). The MPT framework is typically used to characterize the mixtures of processes and cognitive states that underlie individuals' categorical responses (for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009). However, they can also be used to directly test hypotheses at the level of the observed response distributions through goodness of fit and model-selection statistics (e.g., Birnbaum, 2013; Hilbig & Moshagen, 2014; Karabatsos, 2005; Klauer & Kellen, 2011, 2015; Klauer, Singmann, & Kellen, 2015). We will evaluate the models' absolute performance via the G^2 statistic (Read & Cressie, 1988) and their relative performance with the Fisher Information Approximation (FIA; Grünwald, 2007). Where traditional model-selection statistics such as Akaike and Bayesian information criteria (Burnham & Anderson, 2002) rely on the number of parameters as a proxy for model complexity, FIA penalizes models according to their functional flexibility.

In the present study, the different theories establish distinct predictions for the truth evaluations. For example, according to the material-implication account, individuals should consider the ⊤⊤, ⊥⊤, and ⊥⊥ cells of the truth table as true (⊤), but deem the ⊤⊥ cell false (⊥). These predictions are deterministic in the sense that no other response is considered possible. The deterministic nature of axiomatic accounts represents a long-standing problem in psychology due to the need to recast these accounts in order to accommodate the stochastic nature of responses with respect to an experiment's empirical sample space (e.g., Luce, 1995, 1997; Regenwetter, Dana, & Davis-Stober, 2011). For example, the predictions of the material-implication account for the ⊤⊤ cell of the truth table would then be relaxed in order to allow false or NN responses with some probability. The exact manner of relaxation is not entirely clear though. For example, one could assume that individuals almost invariably respond true (e.g., 90% of the times), or alternatively that true responses constitute an absolute or a relative majority, among other possibilities. This issue has been thoroughly explored in studies focusing on whether preferences are transitive (when an individual prefers A to B, and B to C, then A is expected to be preferred to C), in which different stochastic implementations have been considered (e.g. weak, moderate, and strong stochastic transitivity, the triangle inequality; see Regenwetter et al., 2011). In the present case we adopted what we view as the most lenient stochastic implementation; to wit, that the predicted response should occur at last as often as each of the other responses (i.e., that it should enjoy at least a relative majority). The reason is the diagnostic power associated to its failure, as any theory that fails to succeed under these minimal constraints should be seriously questioned. Note that this stochastic specification is completely agnostic regarding the nature of the deviations from the predictions: Individuals might commit "errors" due to a misreading of the sentences, a failure in their evaluation, or a motor-response error, among other possibilities (see Birnbaum, 2013).

Experiment 2: LMM Analysis

For Experiment 2, a mixed linear model was fitted to the data with fully crossed fixed effects for the predictor (IV: P(ψ | φ) and P(φ & ψ)), relevance condition (PO, NE, and IR), and sentence type (P(if φ, then ψ) and P(φ and/but/therefore ψ)) and crossed random effects for participants and scenarios. In `lme4` syntax (Bates, Maechler, Bolker, & Walker, 2015), the LMM used took the following form:

```
DV ~ IV * relevance * sentence_type      + (IV * relevance | participant)
                                         + (IV * relevance | scenario)
```

The R package `afex` (Singmann *et al.*, 2016) was used to obtain the statistical significance of the fixed effects while controlling for variability due to participants and scenarios.