COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017, Suppl. 5) 1202–1215

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12462

Relevance and Reason Relations

Niels Skovgaard-Olsen,[a,b] Henrik Singmann,[c] Karl Christoph Klauer[b]

[a]*Department of Philosophy, University of Konstanz*
[b]*Department of Psychology, Albert Ludwigs Universität Freiburg*
[c]*Department of Psychology, University of Zürich*

Received 24 May 2016; received in revised form 23 August 2016; accepted 5 October 2016

Abstract

This paper examines precursors and consequents of perceived relevance of a proposition A for a proposition C. In Experiment 1, we test Spohn's (2012) assumption that $\Delta P = P(C|A) - P(C|\sim A)$ is a good predictor of ratings of perceived relevance and reason relations, and we examine whether it is a better predictor than the difference measure ($P(C|A) - P(C)$). In Experiment 2, we examine the effects of relevance on probabilistic coherence in Cruz, Baratgin, Oaksford, and Over's (2015) uncertain "and-to-if" inferences. The results suggest that ΔP predicts perceived relevance and reason relations better than the difference measure and that participants are either less probabilistically coherent in "and-to-if" inferences than initially assumed or that they do not follow P(if A, then C) = P(C|A) ("the Equation"). Results are discussed in light of recent results suggesting that the Equation may not hold under conditions of irrelevance or negative relevance.

Keywords: Relevance; Reason relations; And-to-if inferences; Conditionals; Probabilistic coherence; the Equation

1. Introduction

Although the *reason relation* plays a central role in a number of philosophical discussions, a precise explication of this concept is usually absent (e.g., Brandom, 1994; Brewer, 2002; McDowell, 1994; Reisner & Steglich-Petersen, 2011). In Spohn (2012), a precise account has, however, been given in terms of the difference that one proposition, A, makes in the degree of belief of another proposition, C, which draws on the literature on confirmation measures[1]:

Correspondence should be sent to Niels Skovgaard-Olsen, Department of Philosophy, University of Konstanz, Konstanz 78457, Germany. E-mail: niels.skovgaard.olsen@psychologie.uni-freiburg.de, n.s.olsen@gmail.com

The supplemental materials including all data and analysis scripts are available at https://osf.io/fdbq2/.

$$\text{A is a reason for C} \quad \text{iff } P(C|A) > P(C|\sim A) \tag{1}$$

$$\text{A is a reason against C} \quad \text{iff } P(C|A) < P(C|\sim A) \tag{2}$$

At the same time, the notion of epistemic *relevance* is explicated by stating that A is *positively relevant to* C iff (1) holds, *negatively relevant* iff (2) holds, and *irrelevant* iff $P(C|A) = P(C|\sim A)$. One of the general advantages of having such a formal account is that it enables one to investigate the formal properties of reason relations and relevance and to formulate a taxonomy of reason relations (see Spohn, 2012; Section 6.2), which has repercussions for their application to philosophy and psychology (Skovgaard-Olsen, 2015; Spohn, 2013).

The first goal of this study is to test the following prediction attributed to Spohn (2012): There is both a high correlation between ΔP (i.e., $P(C|A) - P(C|\sim A)$) and perceived relevance and between ΔP and ratings of reason relations. As ΔP is only one among a whole family of confirmation measures, we contrast it with the difference measure ($P(C|A) - P(C)$), which is another popular alternative (Douven & Verbrugge, 2012; Tentori, Crupi, Bonini, & Osherson, 2007). Formally, $P(C|A) > P(C|\sim A)$ entails $P(C|A) > P(C)$.[13] However, the degree of relevance as measured by $P(C|A) - P(C|\sim A)$ need not match the degree of relevance as measured by $P(C|A) - P(C)$. This raises the empirical issue of which of the two best describes the degree of perceived relevance and the perceived strength of the reason relation of the participants.[2]

In Experiment 2, we turn to the effects of relevance on the probabilistic coherence of the participants in the uncertain and-to-if inference (i.e., inferring "if A then C" from "A and C").

2. Experiment 1

2.1. Method

2.1.1. Participants

A total of 725 people from the United States, United Kingdom, and Australia completed the experiment, which was launched over the Internet (via www.Crowdflower.com) to obtain a large and demographically diverse sample. Participants were paid a small amount of money for their participation.

The following exclusion criteria were used: not having English as native language (33 participants), completing the experiment in less than 240 s or in more than 5,400 s (43 participants), failing to answer two simple SAT comprehension questions correctly in a warm-up phase (214 participants), providing answers outside the range of 0% to 100% (three participants), and answering "not serious at all" to the question of how serious they would take their participation at the beginning of the study (zero participants). Since some of these exclusion criteria were overlapping, the final sample consisted of 475

participants. Mean age was 38.91 years, ranging from 18 to 73, 55.8% indicated that the highest level of education that they had completed was an undergraduate degree or higher.

2.1.2. Design

The experiment implemented a mixed design with three factors that determined the content and relationship of the antecedent, A, and consequent, C, of a conditional "If A then C." There were two factors that varied within participants: relevance (with three levels: positive relevance (PO), negative relevance (NE), irrelevance (IR)), and priors (with four levels: HH, HL, LH, LL, meaning, e.g., that P(A) = low and P(C) = high for LH). One further factor varied between participants: type of irrelevance (with two levels labeled "same content" and "different content"). Participants were randomly assigned to one of the two irrelevance conditions for each scenario implementing a conceptual distinction between whether A is topically relevant or irrelevant for C. As this factor did not affect any of the results reported here, we do not discuss it any further (see supplementary materials S1, Section 3) and only use the different content irrelevance condition in Experiment 2.

2.1.3. Materials and procedure

We created 18 different scenarios (see supplemental materials for full list), for each of which we constructed 16 conditions according to our design (i.e., four conditions for PO [i.e., HH, HL, LH, LL], four conditions for NE, four conditions for IR-same content, and four conditions for IR-different content; note again that the two IR conditions were collapsed for the analysis as they did not differ). Each participant worked on one randomly selected (without replacement) scenario for each of the 12 within-subjects conditions such that each participant saw a different scenario for each condition.[3] Following the recommendations of Reips (2002) to reduce dropout rates, we presented two SAT comprehension questions as an initial high hurdle in a warm-up phase (in addition to using them for excluding participants).

The experiment was split into 12 blocks, one for each within-subjects condition. The order of the blocks was randomized anew for each participant and there were no breaks between the blocks. Within each block, participants were presented with two pages. The scenario text was placed at the top of each page. One participant might thus see the following scenario text:

> Julia has gained some weight during her holiday in Egypt, and now wishes to lose 5 kg. She is very determined to make lifestyle changes. She is not obese by any means. Yet it is unlikely that she will end up looking like a model—nor is it her goal. Most would characterize her as being within the normal range.

The idea was to use brief scenario texts concerning basic causal, functional, or behavioral information that uniformly activates stereotypical assumptions about the relevance and prior probabilities of the antecedent and the consequent of 12 conditionals that

implement our experimental conditions for each scenario. So, to introduce the 12 within-subjects conditions for the scenario text above, we, *inter alia*, exploited the fact that participants would assume that Julia's beginning to exercise would raise the probability of her losing weight (PO), lower the probability of her gaining weight (NE), and that a sentence describing the present weather conditions of the location where Julia spent her holiday would be irrelevant for whether or not "Julia will lose weight" by exercising after returning from the holiday (IR).

On the first page of each block, the scenario text was followed by two questions presented in random order that measured the prior probability of the two sentences:

Please rate the probability of the following statement on a scale from 0% to 100%:

[Julia has weight loss surgery/Julia will gain weight]

On the second page, the same scenario text was followed by four questions presented in random order. The first two questions measured the conditional probability of the consequent given the antecedent, P(C|A), and its negation, $P(C|\sim A)$. To illustrate using the NE-LL condition (=negative relevance, P(A) = low, P(C) = low) for the scenario above:

Suppose Julia has weight loss surgery.
Under this assumption, how probable is it that the following sentence is true on a scale from 0% to 100%:
Julia will gain weight.

The third question, the relevance rating, asked the participants to rate the extent to which the antecedent was relevant for the consequent on a five-point scale ranging from "irrelevant" to "highly relevant." The fourth question, the reason relation scale, asked the participants to rate the extent to which the antecedent was a reason for/against the consequent on a five-point scale ranging from "a strong reason against," "a reason against," "neutral," and "a reason for" to "a strong reason for." For each question, participants gave their response by entering a number into a specified field. The full list of scenarios, the raw data, the data preparation script, and the analysis script for both Experiment 1 and 2 can all be found at https://osf.io/fdbq2/.

2.2. Results and discussion

We performed a manipulation check (see supplementary materials, S1) and prepared the data for the analysis. Perceived relevance was initially measured in an undirected way, because it was assumed that participants would not be sensitive to the theoretical distinction between positive and negative relevance. To obtain a directed perceived relevance rating, we combined the directional information of the reason relation scale with the relevance rating to generate a directional relevance scale ranging from −4 (strongly negatively relevant) to +4 (strongly positively relevant). If participants indicated that A was a reason *against* C on the reason relation scale, their assessment of how relevant A was for C was interpreted as *negative* relevance. If the participants indicated that A was a reason

for C, their assessment of how relevant A was for C was interpreted as *positive* relevance. If the participants indicated that A was neutral in relation to C, their assessment was interpreted as that A was *irrelevant* for C. The reason relation scale was coded on a scale from −2 (strong reason against) to +2 (strong reason for). ΔP and the difference measure were calculated from the conditional probability questions and the prior of the consequent.

As the data had replicates both on the level of the participant (each participant provided one response for each of the 12 within-participant conditions) and on the level of the scenarios (each scenario could appear in each relevance condition across participants), we employed a linear mixed model (LMM) analysis with crossed random effects for participants and scenarios for the analysis (Baayen, Davidson, & Bates, 2008). We estimated one LMM with directional relevance scale as dependent variable and one LMM with reason relation as dependent variable. Each LMM had fixed effects for ΔP as well as the difference measure. The random effects structures were "maximal" (Barr, Levy, Scheepers, & Tily, 2013): random intercepts for participants and contents with by-participant and by-content random slopes for both fixed effects, and correlations among all by-participant and among all by-content random terms. Fixed effects were evaluated via the Kenward-Roger approximation (via afex; Singmann, Bolker, Westfall, & Aust, 2016). The LMMs did not include effects for the prior manipulations, which were primarily introduced to ensure that our results generalize to the whole spectrum of sentences describing likely and unlikely events.

Fig. 1 displays the estimated effects from both models which clearly show that the effects of ΔP are considerably stronger than the effects of the difference measure (dashed lines). Furthermore, for the model with relevance scale as dependent variable, ΔP was a significant predictor, $F(1, 20.94) = 269.07$, $p < .0001$, but the difference measure failed to be, $F(1, 18.21) = 1.30$, $p = .27$. This indicates that in contrast to the difference measure, only ΔP could explain unique variance. For the model with reason relation as dependent variable, both ΔP, $F(1, 18.99) = 232.35$, $p < .0001$, as well as the difference measure, $F(1, 17.26) = 7.72$, $p = .01$, were significant predictors.[4] Although initially purely philosophically motivated, it turns out that the explication of relevance and reason relations of Spohn (2012) in terms of ΔP is descriptive of the assessments of our participants. See the supplementary materials (S1, Section 2) for a comparison with a further confirmation measure.[5]

3. Experiment 2

The last 10–15 years of research on conditionals within the psychology of reasoning have been marked by the emergence of a *New Paradigm* characterized by a shift from models based on classical logic to probabilistic competence models (Elqayam & Over, 2013). Within the New Paradigm, there is a widespread endorsement of *the Equation*, P(if A, then C) = P(C|A) (Baratgin, Over, & Politzer, 2013; Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer, 2013). In addition to direct evidence stemming from investigations of the probability of the conditionals, and evidence from the truth table task

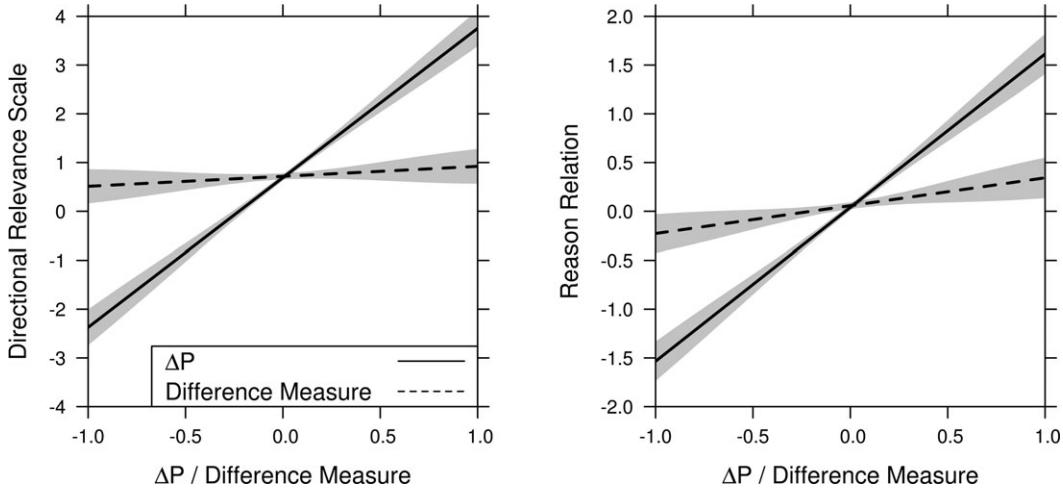

Fig. 1. LMM estimates of fixed effects for Experiment 1. In the left panel, the directional relevance scale is the dependent variable and in the right panel, reason relation is the dependent variable. Error bands show 95% confidence intervals from the LMM.

(Over & Evans, 2003), it has been suggested that evidence from uncertain and-to-if inferences supports this view (Cruz et al., 2015). Cruz et al. (2015, p. 3) use their results from uncertain *and-to-if inferences* to make an argument in favor of the Equation based on the following line of thought: "If people's judgments are highly incoherent for one interpretation [of the conditional], and yet highly coherent for another, there is an argument in favor of the interpretation that renders their judgments coherent." Since it was found that the Equation was better able to make the participants' responses coherent than the material conditional,[6] they interpret their results as providing strong evidence in favor of the Equation.

In Skovgaard-Olsen, Singmann, and Klauer (2016), we found that the evidence for the Equation was qualified once P(if A, then C) was evaluated across three relevance levels, where relevance was defined as described above.[7] While there was an almost perfect relationship between P(if A, then C) and P(C|A) in the positive relevance (PO) condition, this relationship was markedly weaker in the negative relevance (NE) and even weaker in the irrelevance (IR) condition. Moreover, the results showed that P(C|A) is a much better predictor of P(Even if A, then still C) across relevance levels. The second goal of this study is therefore to test whether introducing the same relevance manipulation to and-to-if inferences leads participants to perceive a defect in the conditionals in the NE and IR conditions, which should make them more reluctant to infer the conclusion under these conditions.

Following our earlier findings (Skovgaard-Olsen et al., 2016), we hypothesize that the results of Cruz et al. (2015) are similarly affected by a relevance manipulation. More specifically, in line with Skovgaard-Olsen et al. (2016), we hypothesize that for indicative conditionals, we replicate their findings in the PO condition, but not in the NE or IR

conditions. In contrast, for the concessive (i.e., *even-if*) conditionals, the level of probabilistic coherence of the participants was not expected to drop in the NE/IR conditions as compared to the PO condition. Hence, we test whether the participant's degree of probabilistic consistency drops under manipulations of negative relevance and irrelevance for indicative conditionals, when P(if A, then C) and P(C|A) are equated.

3.1. Method

3.1.1. Participants

The present experiment was part of Skovgaard-Olsen et al. (2016); consequently, we analyzed the same 348 participants reported there.[8] However, the data for this specific task are reported here for the first time. Data were collected over the Internet.

3.1.2. Design

Experiment 2 implemented a mixed design with the same 12 within-participant conditions as Experiment 1. In addition, the type of conditional was varied between participants (with two levels: indicative ("if A, then C"), concessive ("Even if A, then still C")).

3.1.3. Materials and procedure

Prior to Experiment 2, we selected 12 scenarios from the set of 18 scenarios for which all within-subjects condition were most precisely realized (see supplementary materials, S1). The 12 within-participants conditions were randomly assigned to 12 different scenarios for each participant anew.[9] Moreover, in contrast to Experiment 1, the participants reported probabilities using sliders ranging from 0% to 100%. Aside from this, Experiment 2 was designed following the schema of Experiment 1.

Within each of the 12 within-participants conditions, the participants were presented with three pages, which had a randomly chosen scenario text at the top. On the first page of the experiment, the scenario text was followed by two questions presented in random order. The first measured the conditional probability of the consequent given the antecedent using the same question format as in Experiment 1. The second question measured the probability of the conjunction of the antecedent and the consequent, which was used to measure the probability of the premise of an inference task on the third page.

On the second page, the participants evaluated either the acceptability or the probability of conditionals in a task reported in Skovgaard-Olsen et al. (2016). On the third page, the participants were presented with a short argument, whose premise was the conjunction, and a conditional as the conclusion. The participants were here reminded of the probability that they had assigned to the conjunction on the first page and asked to assess the probability of the conditional on its basis. Thus, one participant might see the following question on page three:

In the following you will be presented with a short argument.

Premise Julia starts to exercise AND Julia will gain weight.
(You have estimated the probability of the premise as 13%)

Based on the premise and its probability, please indicate how much confidence you have in the following conclusion:

Conclusion Therefore, IF Julia starts to exercise, THEN Julia will gain weight.

3.2. Results and discussion

We estimated probabilistic coherence across relevance manipulations in the and-to-if inference following Cruz et al. (2015).[10] This entailed comparing the observed coherence rates against a chance coherence rate.[11] As shown in Table 1, the descriptive data seemed to confirm our predictions. For indicative conditionals, participants' probabilistic coherence was above chance levels only for PO, while for the concessive conditionals, participants' probabilistic coherence was above chance levels for all relevance conditions. Table 1 also shows whether participants' probability evaluations conformed to P(C|A) ≥ P(A,C) independent of the uncertain and-to-if inference task (i.e., both responses from the first page of each within-subject condition). Participants reliably conform to this inequality in ≈78% of the cases (≈19% above chance) across relevance levels with 77% in PO, 81% in NE, and 76% in IR. In contrast, the participants' conformity to P(Conclusion) ≥ P(A,C) varied markedly across relevance levels with 87% in PO, 66% in NE, and 54% in IR. Given this apparent discrepancy between the effects of our relevance factor on the conformity to these two inequalities, we decided to analyze the effect of relevance on the conformity to both inequalities together while correcting for chance.

The statistical analysis of these data followed Singmann, Klauer, and Over's (2014; see also Evans, Thompson, & Over, 2015). We first coded coherent/conforming responses

Table 1
Frequency of probabilistically coherent and-to-if inference (and corresponding percentages)

	P(conclusion) ≥ P(A,C)			P(C|A) ≥ P(A,C)	
	True	Chance	Δ	True	Δ
P(conclusion) = P(If A, C)	1511 (69%)	59%	9%	1713 (78%)	19%
PO	634 (87%)	45%	41%	565 (77%)	32%
NE	481 (66%)	71%	−5%	591 (81%)	10%
IR	396 (54%)	62%	−8%	557 (76%)	14%
P(conclusion) = P(Even if)	1516 (77%)	59%	18%	1557 (79%)	20%
PO	515 (78%)	47%	31%	508 (77%)	30%
NE	482 (73%)	70%	3%	518 (79%)	9%
IR	519 (79%)	60%	19%	531 (81%)	20%

Note. "True" gives raw probabilistic coherence, "Chance" gives probabilistic coherence based on uniform responses, and "Δ," their difference. Value for conformity to P(C|A) ≥ P(A,C) is given in the two rightmost columns.

in which either P(Conclusion) or P(C|A) was at least as large as that of the premise with "1" and incoherent responses/non-conforming with "0." To implement the chance baseline, we subtracted 1 minus the probability of the premise from this value. We then estimated a LMM with this chance-corrected violation score (in which values above 0 indicate coherent/conforming responding above chance) as dependent variable and relevance condition as well as type of probabilistic measure (coherence vs. conformity) and their interaction as independent variables separately for the indicative and concessive conditional groups. We thus had two LMMs in total; one for each type of conditional group (i.e., indicative and concessive). We again estimated crossed random effects for participants and scenarios with maximum random slopes (i.e., by-participant and by-scenario random slopes for all fixed effects plus correlations among the slopes).

For the indicative conditionals, the statistical analysis confirmed our prediction that relevance affects probabilistic coherence. It also affected probabilistic conformity, but to a lesser degree. All effects of the LMM were significant (including the intercept), most important, the interaction of relevance condition and type of probabilistic measure, $F(2, 15.15) = 25.15$, $p < .0001$. It indicated that coherence was only above chance for PO ($\beta = 0.42$, 95% CI [0.33, 0.50]), but not for NE ($\beta = -0.05$, 95% CI [−0.11, 0.00]) and IR ($\beta = -0.07$, 95% CI [−0.14, −0.00]). In contrast, conformity was above chance for all three conditions (smallest β for NE = 0.09, 95% CI [0.05, 0.14]). Note also that for both types, PO was larger than NE and IR (all ps $< .0001$), while the latter two did not differ from each other (ps $> .43$).

For the concessive conditionals, we found both a significant intercept indicating general above chance responses, $\beta = 0.19$, 95% CI [0.15, 0.23], $F(1, 21.37) = 78.39$, $p < .0001$, and an effect of relevance condition, $F(2, 13.05) = 18.75$, $p = .0001$, but no further effects (all remaining $p > .22$), indicating that type of probabilistic measure had no effect for the concessive conditionals. For the main effect of relevance, all three relevance conditions differed significantly from each other, all $p < .004$, but coherence and conformity were significantly above chance in each case ($\beta_{NE} = 0.06$, 95% CI [0.01, 0.12], $\beta_{IR} = 0.19$, 95% CI [0.14, 0.25], and $\beta_{PO} = 0.30$, 95% CI [0.23, 0.37]).

As Table 1 indicates, the proportions of coherent or conforming responses were around 78% across relevance conditions in all cases except for probabilistic coherence for indicative conditionals. Table 1 indicates that all further differences in the statistical analysis are solely driven by different sizes of the chance intervals. This finding, thus, corroborates the result from Skovgaard-Olsen et al. (2016) that P(if A, then C) ≠ P(C|A) for negative relevance or irrelevance but that P(Even if A, then still C) = P(C|A) across all relevance levels.

Our results extend Cruz et al. (2015), who found that participants were probabilistically coherent above chance levels overall. However, they employed stimulus material inspired by the Linda problem from Tversky and Kahneman's (1983) work on the conjunction fallacy, which implements only the PO condition for the indicative conditional at one specific priors level ("If Linda votes in the municipal elections, then she votes for the Socialist Party").[12] In contrast, our results are based on all permutations of the relevance and priors levels for both indicative and concessive conditionals.

Interestingly, Tentori, Crupi, and Russo (2013) showed that the prevalence of the conjunction fallacy also depends on whether or not the information the participants are given in the scenario is positively relevant for the second conjunct. Tentori et al. (2013) interpret their results as showing that the participants committing the conjunction fallacy tend to substitute a sound estimation of confirmation relations for the intended probability assignment. This introduces the possibility that a similar cognitive mechanism may be implemented in the participants' lack of conformity to P(conclusion) ≥ P(premise) above chance levels in the NE and IR conditions.

4. General discussion

In this study, we manipulated relevance and prior probabilities using a new cluster of scenarios (see supplementary materials). Experiment 1 presented evidence for high agreement between ΔP and ratings of perceived relevance and reason relations and suggests that ΔP is a better predictor than the difference measure. Follow-up studies might contrast ΔP with further confirmation measures (see Douven & Verbrugge, 2012; Tentori et al., 2007). Interestingly, when removing extreme ($-\infty$) and undefined values, ΔP correlates to a very high degree, $r = .96$, with the following log odds ratio for our data set:

$$\tau(\mathrm{C} \mid \mathrm{A}) - \tau(\mathrm{C} \mid \sim \mathrm{A}) \approx \ln\left(\frac{\frac{P(Y=1|X=1)}{P(Y=0|X=1)}}{\frac{P(Y=1|X=0)}{P(Y=0|X=0)}}\right)$$

The logged odds ratio measure thus accounts for pretty much the same variance as ΔP. The log odds ratio measure is a more direct approximation of Spohn's (2012: ch. 6) ranking-theoretic explication of the reason relation (i.e., $\tau(\mathrm{C} \mid \mathrm{A}) - \tau(\mathrm{C} \mid \sim \mathrm{A}) > 0$) in probability theory, and it was therefore used as a relevance parameter in the logistic regression model of the conditional inference task put forward in Skovgaard-Olsen (2015).

In their relevance theory, Wilson and Sperber (2004; see also Sperber, Cara, & Girotto, 1995) propose that maximization of relevance is a general principle structuring both cognition and communication. Their account introduces an economic aspect to assessments of relevance; the cost of processing information decreases the perceived relevance, whereas the gain in cognitive effects increases the perceived relevance. In principle, it is possible to combine this idea with Spohn's (2012) theory as the latter gives us a precise notion of cognitive effect in terms of difference-making in degrees of belief, and the precise formal principles guiding belief revision, whereas Sperber and Wilson's theory introduces a focus on processing costs.

In Experiment 2, we examined the role of relevance for the uncertain and-to-if inference task presented in Cruz et al. (2015). It was found that the participants perform above chance levels for PO, but below chance levels for NE and IR, thus qualifying Cruz et al.'s (2015) results. This result is hard to reconcile with probabilistic approaches to the semantics of conditionals that equate P(if A, then C) with P(C|A) (Evans & Over, 2004;

Oaksford & Chater, 2007; Pfeifer, 2013). In fact, it presents these theories with a dilemma: Either P(if A, then C) = P(C|A) does not hold across relevance manipulations, or the participants are less probabilistically coherent than initially seemed to be the case.

The Equation (P(if A, then C) = P(C|A)) here acts as an auxiliary assumption that implies P(conclusion) ≥ P(premise), if the participants are to be probabilistically coherent. Throughout the last 10–15 years, the Equation has been supported time after time (see Douven, 2015; ch. 3–4). However, relevance levels dramatically moderate this relationship (Skovgaard-Olsen et al., 2016). Accordingly, we suggested that conditionals that violate the default assumption of positive relevance (e.g., "If the sun is shining in Egypt, then Julia will lose weight") are viewed as defective and penalized in their probability ratings. Based on these results, we suspect that the culprit that makes it appear that the participants are probabilistically incoherent in the NE and IR condition is the Equation. On the alternative outlined in Skovgaard-Olsen et al. (2016), the participants rely on a heuristic for assessing reason relations when evaluating P(if A, then C), which introduces no normative requirement that P(conclusion) ≥ P(premise) for the NE and IR conditions, where there is no strong relationship between P(if A, then C) and P(C|A). The participants' lack of conformity to P(conclusion) ≥ P(premise) above chance levels in the NE and IR conditions is, in other words, explained by the perceived defect of these conditionals owing to their violation of the expectation that A is a reason for C.

This interpretation is supported by the observation that participants conformed to P(C|A) ≥ P(A,C) in their probability evaluations independently of the uncertain and-to-if inference task in ≈78% of the cases, both in the group with indicative and with concessive conditionals across relevance conditions. Hence, the below chance level performance in the uncertain and-to-if inference task for the NE and IR conditions does not appear to reflect a general failure to conform to P(C|A) ≥ P(A,C) across relevance levels.

Aside from the Equation, which introduces the normative requirement that P(conclusion) ≥ P(premise), because P(Conclusion) is treated as P(C|A) and it is a requirement of probability theory that P(C|A) ≥ P(A,C), other semantics of conditionals are also committed to P(conclusion) ≥ P(premise) in the and-to-if inference. The reason is that several conditional logics treat "A ∧ C ⊨ if A, then C" as a valid argument schema (Arlo-Costa, 2007), and given P(B) ≥ P(A), whenever A ⊨ B, it holds that P(if A, then C) ≥ P(A ∧ C).

One example is Lewis's semantics of counterfactuals, and there is already discussion on whether a weakening of the system should be allowed, which avoids treating "A ∧ C ⊨ if A were, then C would have been" as a theorem (Kutschera, 1974). It remains to be seen, whether there are differences in the participants' conformity to P(Conclusion) ≥ P(A ∧ C) for indicative and counterfactual conditionals across relevance conditions.

However, it is clear that other theories aside from those endorsing the Equation are faced by explanatory challenges by our results. As Douven (2015: Section 2.1–2.2) points out, "A ∧ C ⊨ if A, then C" is valid for semantics of indicative conditionals such as the material conditional, Stalnaker's possible worlds semantics, and three-valued truth tables like the de Finetti table. Yet it is rejected by inferentialism, which holds that it is part of the truth conditions of indicative conditionals that there is an inferential relation connecting A and C.

Acknowledgments

We are very grateful for discussions with David Over, Igor Douven, Vincenzo Crupi, Nicole Cruz, Wolfgang Spohn, Karolina Krzyżanowska, Peter Collins, Ulrike Hahn, and the audiences of talks at the annual meeting of New Frameworks of Rationality, the What-If group in Konstanz, the Department of Social Psychology and Methodology in Freiburg, and the "Working Group in the History and Philosophy of Logic, Mathematics, and Science" in UC Berkeley. Moreover, careful comments by the 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).

Notes

1. Yet it should be noted that there is the large problem of the unification of theoretical reasons and practical reasons raised by the contributions in Reisner and Steglich-Petersen (2011), which Spohn's account does not yet tackle, and that there are predecessors for analyzing epistemic relevance in the way Spohn does in the literature (see Falk & Bar-Hillel, 1983; Walton, 2004, ch. 4).
2. Actually, Spohn (2012; ch. 6)'s preference for the delta-p measure over the difference measure is grounded in the different behavior of their ranking-theoretic analogs. Although it would indeed be attractive to investigate psychological applications of ranking theory, this study takes the more conservative, probabilistic route.
3. The supplementary materials contain details on how 12 scenarios were selected for future experimentation on the basis of the 18 scenarios we created.
4. Note that the zero-order correlations of the difference measure with both dependent variables were highly significant ($r > .39$). But its effect was reduced in the joint LMM due to the high covariance with ΔP ($r = .73$), which was itself more strongly correlated with both dependent variables ($r > .53$).
5. We tested a further confirmation measure, Keynes and Horwich's logged-ratio measure (Tentori et al., 2007). Unfortunately, this measure introduces the problem of extreme ($-\infty$) or undefined values for 24% of our observations (e.g., when the denominator is 0). When analyzing the reduced sample, no unique variance was accounted for by the logged-ratio measure and again only ΔP was a significant predictor (see Section 2, supplementary materials). Other proposed confirmation measures contain variables not collected in this study (such as the likelihood) and could, therefore, not be applied to our data.
6. The material conditional ("$\supset$") has a truth table that is logically equivalent to '$\neg A \vee C$' and for this reason, Cruz et al. (2015) attribute the prediction that P(if A, then C) = P($\neg A \vee C$) to this theory.

7. For results on introducing the relevance manipulation into the truth table task, see Skovgaard-Olsen, Kellen, Krahl, and Klauer (unpublished data).
8. In contrast to the other task in Skovgaard-Olsen et al. (2016), the present task did not involve differentiating between assessing the probability and acceptability of the respective sentences as two modes of evaluation. For this reason, the two groups evaluating probabilities and acceptabilities separated in Skovgaard-Olsen et al. (2016) are analyzed together below.
9. See also https://osf.io/j4swp/.
10. Cruz et al. (2015) argued that—given the truth of the Equation, where P(if A, then C) is interpreted as P(C|A)—participants have to respond with P(if A, then C) $\geq$ P(A&C) to be probabilistically coherent. From P(A&C) = P(C|A)*P(A) and $0 \leq P(A) \leq 1$, it follows that P(C|A) $\geq$ P(A&C).
11. Assuming that a response produced by any other process or a random response has an equal chance of falling on any point of the response scale, the probability of selecting a response greater than P(A,C) amounts to $1 - P(A,C)$.
12. Nicole Cruz (personal communication, 20.05.2016).
13. In the special case where P(A) = 1, P(C|A) = P(C) but $P(C|\sim A)$ is undefined.

References

Arlo-Costa, H. (2007). The logic of conditionals. In E. N. Zalta(Eds.), *The stanford encyclopedia of philosophy*. (Spring 2016 edition; accessed May 1, 2016). Available at http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005.

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning*, *19*(3), 308–28. doi:10.1080/13546783.2013.809018

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

Brandom, B. (1994). *Making it explicit*. Cambridge, MA: Harvard University Press.

Brewer, B. (2002). *Perception and reason*. Oxford, UK: Oxford University Press.

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6*, 192. doi:10.3389/fpsyg.2015.00192

Douven, I. (2015). *The epistemology of indicative conditionals. Formal and empirical approaches*. Cambridge, UK: Cambridge University Press.

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

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

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

Evans, J. S. B. T., Thompson, V. A., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology*, *6*, 398. doi:10.3389/fpsyg.2015.00398

Falk, R., & Bar-Hillel, M. (1983). Probabilistic dependency between events. *Two-Year College Mahematics Journal*, *14*, 240–7.

Kutschera, F. (1974). Indicative conditionals. *Theoretical Linguistics*, *1*, 257–69.
McDowell, J. (1994). *Mind and world*. Cambridge, MA: Harvard University Press.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
Over, D., & 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
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
Reisner, A. & Steglich-Petersen, A. (eds.) (2011). *Reasons for belief*. Cambridge, UK: Cambridge University Press.
Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. http://dx.doi.org/10.1027//1618-3169.49.4.243.
R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.
Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). afex: Analysis of factorial experiments. R package version 0.16-1 [accessed August 1, 2016]. Available at https://CRAN.R-project.org/package=afex
Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, *5*, 316. doi:10.3389/fpsyg.2014.00316
Skovgaard-Olsen, N. (2015). Ranking theory and conditional reasoning. *Cognitive Science*, *40* (4), 848–880, doi:10.1111/cogs.12267
Skovgaard-Olsen, N., Kellen, D., Krahl, H., & Klauer, K. C. (in review). Relevance differently affects the truth, acceptability, and probability evaluations of 'and', 'but', 'therefore', and 'if then'.
Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, *150*, 26–36. doi:10.1016/j.cognition.2015.12.017
Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*, 31–95.
Spohn, W. (2012). *The laws of beliefs*. Oxford, UK: Oxford University Press.
Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, *37*, 1074–1106. doi:10.1111/cogs.12057
Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107–119.
Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235–255. doi:10.1037/a0028770.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315. doi:10.1037/0033-295X.90.4.293
Walton, D. (2004). *Relevance in argumentation*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wilson, D., & Sperber, D. (2004). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 607–632). Oxford, UK: Blackwell.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Data S1. Manipulation Check and Selection of Scenarios.

The supplement materials including all data and analysis scripts are available at: https://osf.io/fdbq2/