Conditionals, Individual Variation, and the Scorekeeping Task

Niels Skovgaard-Olsen (niels.skovgaard.olsen@psychologie.uni-freiburg.de)
Department of Psychology, Engelbergerstraße 41,
79085 Freiburg Germany

David Kellen (dkellen@syr.edu)
Department of Psychology, Huntington Hall 516
Syracuse, NY 13244 USA

Ulrike Hahn (u.hahn@bbk.ac.ukk)
Department of Psychological Sciences, Malet Street
London, WC1E 7HX UK

Karl Christoph Klauer (christoph.klauer@psychologie.uni-freiburg.de)
Department of Psychology, Engelbergerstraße 41,
79085 Freiburg Germany

Abstract

In this manuscript we study individual variation in the interpretation of conditionals by establishing individual profiles of the participants based on their behavioral responses and reflective attitudes. To investigate the participants' reflective attitudes we introduce a new experimental paradigm called the Scorekeeping Task, and a Bayesian mixture model tailored to analyze the data. The goal is thereby to identify the participants who follow the Suppositional Theory of conditionals and Inferentialism and to investigate their performance on the uncertain and-to-if inference task.

Keywords: conditionals; individual variation; and-to-if; norms; the Equation; inferentialism

Introduction

According to a popular theory in the psychology of reasoning (the *Suppositional Theory*, or 'ST'), the probability of an indicative conditional (e.g. 'If I forget to pay the rent, then my landlord will complaint') is evaluated by a mental algorithm known as the *Ramsey test* (Evans & Over, 2004; Oaksford & Chater, 2007; Baratgin, Over, and Politzer, 2013).

> THE RAMSEY TEST: to evaluate P(if A, then C) the participants add the antecedent to their background beliefs, make minimal adjustments to secure consistency, and evaluate the probability of the consequent on the basis of this temporarily augmented set of beliefs.

Quantitatively, this introduces the following prediction, which is known as "the Equation":

PRED_1: P(if A, then C) = P(C|A)

Given that P(C|A) ≥ P(A,C) follows from the axioms of probability theory (an inequality referred to as *probabilistic coherence;* PCh), ST also predicts that:

PRED_2: P(if A, then C) ≥ P(A,C)

Accordingly, the participants are predicted to conform to the following inequality in the so-called *uncertain and-to-if inference* (UAI), where they are presented with 'A and C' as a premise and 'if A, then C' as a conclusion and asked to assign probabilities to each:

PRED_{2A}: P(Conclusion) ≥ P(Premise)

Cruz, Baratgin, Oaksford, and Over (2015) found that the participants conformed to PRED_{2A} at above chance levels. This has been taken as indirect evidence in favor of ST.

There is presently a considerable interest in and-to-if inferences, because recently a theory known as '*inferentialism*' made its appearance into the psychology of reasoning, which posits that indicate conditionals express inferential relations. In the truth-conditional version of inferentialism, it rejects the validity of the and-to-if inference 'A∧C ⊨ if A, then C' (Douven, 2015). Truth-conditional inferentialism rejects the validity of this argument scheme, because the indicative conditional is viewed as expressing a *reason relation* and the mere truth of A and C does not ensure that they are inferentially connected. Rejecting the validity of the and-to-inference is a distinguishing feature of this approach that separates it from other popular semantics of conditionals like Stalnaker's possible worlds semantics or the de Finetti truth table endorsed by proponents of ST.

In Skovgaard-Olsen, Singmann, and Klauer (2016a) a weaker probabilistic implementation of inferentialism was given in the form of the Default and Penalty Hypothesis (DP), which employs the following explication of the reason relation:

> PO: A is positively relevant for C (and a reason *for* C) iff P(C|A) > P(C|~A)
> NE: A is negatively relevant for C (and a reason *against* C) iff P(C|A) < P(C|~A)
> IR: A is irrelevant for C iff P(C|A) = P(C|~A)

DP posits that the participants have the goal of evaluating whether a sufficient reason relation obtains when evaluating

P(if A, then C). According to Spohn's (2012: ch. 6) explication of the reason relation given above, this requires at least two things: (a) assessing whether A is positively relevant for C, and (b) assessing the sufficiency of A as a reason for C by means of P(C|A). DP moreover postulates that the participants follow the heuristic, when processing natural language conditionals, of making the default assumption that (a) is satisfied, which reduces their task of assessing P(if A, then C) to assessing P(C|A). However, once the participants are negatively surprised by a violation of this default assumption, such as when they are presented with stimulus materials implementing the NE or IR category, they apply a penalty to P(if A, then C) to express the conditional's failure to express that A is a reason for C. An example would be the conditional 'If Oxford is in England, then Napoleon is dead' which sounds defective to the extent that the antecedent is obviously irrelevant for the consequent.

In support of DP, it was found in Skovgaard-Olsen *et al.* (2016a) that $PRED_1$ only holds when A is positively relevant for C in virtue of raising its probability. When A is negatively relevant by lowering C's probability, and when A is irrelevant for C by leaving its probability unchanged, violations of $PRED_1$ occur. Consistent with these findings, it was found in Skovgaard-Olsen *et al.* (2016b) that the above-chance level of conformity to $PRED_{2A}$ reported in Cruz *et al*. (2015) only holds for PO. In NE and IR the participants are performing below chance levels. Further-more, this is a pattern that is not reflected in their conformity to the theorem $P(C|A) \geq P(A,C)$ across relevance levels, in spite of the fact the participants are supposed to conform to P(if A, then C) = P(C|A), according to ST.

It is presently unclear whether this finding of lack of conformity to $PRED_{2A}$ in the NE and IR conditions indicates that the participants are making a reasoning error (by following ST) or whether they are not making a reasoning error but simply basing their performance on a different interpretation of conditionals (by following DP). The goal of the present study is to address this question.

In the present experiment, we seek to establish individual profiles of the participants based on their behavioral responses and reflective attitudes. In order to study their reflective attitudes we implemented a novel experimental paradigm – the *Scorekeeping Task* – suggested in Skovgaard-Olsen (2015), as well as a Bayesian mixture model tailored to classify the data coming from it (both are discussed in detail below). Based on this novel task and the associated data-analytic method, we were able to investigate two key questions: First, whether participants classified as ST accord with ST's $PRED_{2A}$ prediction for the UAI across a relevance manipulation. Second, whether participants classified as DP accord with DP's prediction that $PRED_{2A}$ only holds in the PO condition. In the IR condition, DP participants are expected to apply a penalty to conditionals in the conclusion of the UAI, such that P(if A, then C) < P(C|A) can occur, effectively dismissing $PRED_{2A}$.

Experiment

Method

Participants

A total of 354 people from the USA, UK, Canada, and Australia completed the experiment, which was launched over the Internet (via Mechanical Turk) 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 (6 participants), completing the experiment in less than 300 seconds (2 participants), failing to answer two simple SAT comprehension questions correctly in a warm-up phase (89 participants), and answer-ring 'not serious at all' to the question 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 261 participants. Mean age was 36.53 years, ranging from 20 to 75, 66% were female, 66% indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment implemented a within-subject design with two factors varied within participants: relevance (with two levels: PO, IR) and priors (with four levels: HH, HL, LH, LL, meaning, for example, that P(A) = low and P(C) = high for LH).

Materials and Procedure

We used a slightly modified version of 12 of the scenarios presented in Skovgaard-Olsen *et al*. (2016b). For each scenario we had 8 conditions according to our design (i.e., 4 conditions for PO [i.e., HH, HL, LH, LL], 4 conditions for IR). Each participant worked on one randomly selected (without replacement) scenario for each of the 8 within-subjects conditions such that each participant saw a different scenario for each condition. 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 four phases and on average took ca. 23 minutes to complete. Here we focus on conveying the underlying conceptual ideas.

Phase 1, Behavioral Responses

The first phase contained eight blocks, one for each within-subjects condition. The order of the blocks was randomized anew for each participant and there were no breaks. Within each block, the participants were presented with four pages. On the first page, the participants were shown a scenario text like the following:

> *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.

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 8 conditionals that implement our experimental conditions for each scenario. So to introduce the 8 within-subjects conditions for the scenario above we, *inter alia*, exploited the fact that the participants would assume that Scott's turning on the warm water would raise the probability of Scott being warm soon (PO) and that Scott's friends being roughly the same age as Scott would be irrelevant for whether Scott will turn on the warm water (IR).

This scenario text was repeated on each of the following three pages, which measured P(A and C), P(C|A), and P(if A, then C) in random order. Throughout the experiment, the participants gave their probability assignments using sliders with values between 0 and 100%. To measure P(C|A), the participants might thus be presented with the following question in an IR condition:

Suppose Scott's friends are roughly the same age as Scott. Under this assumption, how probable is it that the following sentence is true on a scale from 0 to 100%:
Scott will turn on the warm water.

Phase 2, the Scorekeeping Task

In this phase the participants were first presented with a new IRHH item to be rated in the same way as the items in phase one. Then the participants were presented with the following instruction:

When given the task you just completed, John and Robert responded very differently to some of the scenarios as outlined below.

And it was explained that John and Robert responded in the following way to the "if-then sentence" and the "suppose-sentence" (where the "suppose-sentence" had been identified for the participants as the type of question quoted above for measuring P(C|A)):

*John assigned **99%** to the suppose-sentence and **1%** to the if_then sentence.*
*Robert assigned **90%** to the suppose-sentence and **90%** to the if_then sentence.*

Note that although John and Robert are fictive participants, these values were based on actual data provided by other participants in response to the IRHH item in previous experiments. In order to reduce the processing demands, these values were repeated on each of the following four pages along with the IRHH item, which John and Robert allegedly had responded to. The conditional took the following form, and it was evaluated in the context of a dating scenario describing Stephen's preparations for a date with Sara: 'If Stephen's neighbour prefers to put milk on his cornflakes, then Stephen will wear some of his best clothes on the date'.

As part of the scorekeeping task, the participants were instructed to apply a sanction to John or Robert's response based on its adequacy. Given their large divergence, the participants were instructed that at most one of John or Robert's responses could be approved as adequate.

Since the experiment was run on Mechanical Turk we exploited the fact that an ecologically valid sanction for the participants would be not to have a task (a "HIT") approved. Since the approval of HITs on Mechanical Turk determines whether the participants are paid for a completed task (and moreover counts towards their reputation on Mechanical Turk, which determines whether they can participate in future HITs) it is our experience that the participants care a lot about the approval of their HITs. We therefore expected that applying the sanction of not approving either John or Robert's HIT based on its adequacy would be a contextually salient sanction, which the participants would be highly motivated to reason with.

Next the participants were asked to state the reasons that they could think of which could be given for or against John and Robert's responses in an open entry question, which was included in the experiment for exploratory purposes.

On the two pages that followed, the participants were presented with John's criticism of Robert and Robert's criticism of John in random order. Robert made the following complaint about John's response:

Robert's no difference justification: "There is no difference between the two questions. So why do you give a lower probability to:
'IF Stephen's neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date'
than you gave to: *'Stephen will wear some of his best clothes on the date'* under the assumption that *'Stephen's neighbour prefers to put milk on his cornflakes'*?
This makes no sense!"

John in turn made the following complaint about Robert's response:

John's irrelevance justification: "Whether *'Stephen's neighbour prefers to put milk on his cornflakes'* or not is irrelevant for whether *'Stephen will wear some of his best clothes on the date'*.
So why do you give such a high probability to: *'IF Stephen's neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date'?* This makes no sense!"

In each case, the participants were asked to indicate (yes/no) whether they agreed with the following statements:

John's irrelevance justification [/Robert's no difference justification] shows that Robert's [/John's] response is wrong.
Robert [/John] needs to come up with a very good response to John's [/Robert's] criticism, if his HIT is to be approved.

Finally, after having seen the justifications from both sides, the participants were asked which justification they found

most convincing by choosing between the following options, presented in random order:

The two justifications are equally convincing
John's irrelevance justification
Robert's no difference justification

The participants then had to indicate who's HIT deserved to be approved based on their justifications by selecting one of the options below, presented in random order:

None of their HITs should be approved
Robert's HIT should be approved
John's HIT should be approved

Phase 3, the Uncertain And-to-If Inference

This phase tested the participants' performance on the UAI under relevance manipulations. Phase 3 was used to measure whether the participants displayed a consistent behavior on the UAI with the interpretation of the conditional that they had been classified according based on their responses in phase 1 and phase 2.

Phase 3 contained 8 blocks implementing the same within-subjects conditions as phase 1. For each participant, the same permutations of scenarios and within-subject conditions that had been randomly generated in phase 1 was displayed again in random order. First the participants were instructed that they would be presented with a scenario text as earlier and a short argument based on the scenario text. They were told that the premise and the conclusion of this argument could be uncertain and that it was their task to evaluate the probabilities of the premise and conclusion. Each block contained one page. On the top of the page the scenario text was placed as a reminder. Below the participants were instructed to read an argument containing the conjunction as a premise and the conditional as a conclusion, employing sentences that they assigned probabilities to in phase 1. Furthermore, the actual value of the probability that they had assigned to the premise in phase 1 was displayed to the participants in a salient blue color. We here illustrate it using the example from above from phase 1 of a POHH item:

Premise: *Scott's turns on the warm water AND Scott will be warm soon.*
Conclusion: *IF Scott's turns on the warm water, THEN Scott will be warm soon.*
You have estimated the probability of the premise as: ***90%****. Please rate the probability of the statement in the conclusion on a scale from 0 to 100%.*

In Phase 4, we tested the participants' interpretation of the probabilities (Hertwig & Gigerenzer, 1999). These results are beyond the scope of the present manuscript and therefore not reported here.

Bayesian Mixture Modeling

In order to investigate the participants' interpretation of the conditional, the probability judgments they produced in Phase 1 were classified as coming from one of two latent classes using an indicator variable *w*. This classification was achieved by means of a Bayesian Mixture model (for a similar approach, see Lee, 2016). In the PO condition, where both ST and DP make the same predictions (see the left panel of Figure 1), the mixture model assumed that responses from an individual i were generated by ST/DP ($w_i^{PO} = 1$), or by an unclassifiable response-generation mechanism ($w_i^{PO} = 0$), for an item-pair j:

$$P(if\ A, then\ C)_{i,j} = \begin{cases} \beta_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 0, \\ P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{PO} = 1, \end{cases}$$

where $0 \leq \beta_{i,j} \leq 100$.

When an individual follows ST/DP, the generated P(if A, then C) are expected to follow P(C|A) along with some truncated Gaussian noise term $\varepsilon_{i,j}$ with mean 0 and variance σ^2 (see the left panel of Figure 1). This noise captures the variability that is commonly observed in probability judgments across the [0%, 100%] interval (see Costello & Watts, 2016). When an individual follows an unclassified pattern, their responses were captured by a saturated model, which established a β parameter per data point (predicting the latter perfectly).[1]

In the IR condition, the model only considered participants that were classified as ST/DP in the PO condition (i.e., the PO condition served as a filter for the IR condition). Here, both ST ($w_i^{IR} = 0$) and DP ($w_i^{IR} = 1$) make distinct predictions:

$$P(if\ A, then\ C)_{i,j} = \begin{cases} P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{IR} = 0, \\ \theta_i P(C|\ A)_{i,j} + \varepsilon_{i,j}, & w_i^{IR} = 1, \end{cases}$$

with $0 \leq \theta_i \leq 1$.

When individuals follow ST, the generated P(if A, then C) are again expected to follow P(C|A). In contrast, when individuals follow DP, P(if A, then C) follows a penalized version of P(C|A) (with the penalty being determined by θ).

Note that when θ=1, the ST and DP models coincide, although the implied predictions are not really in accordance with the gist of DP. However, this point turns out not to be of practical import, because since ST is more parsimonious it will be preferred when θ=1 (see Lee, 2016).

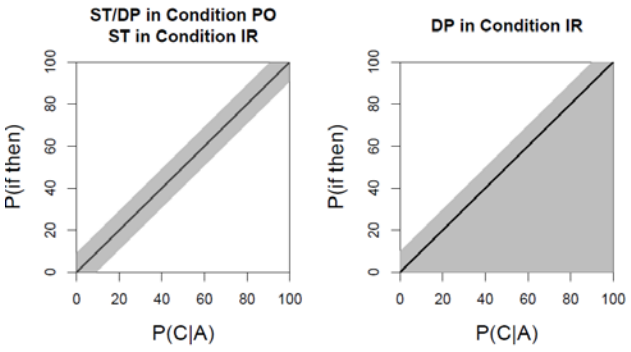

Figure 1. Predictions from both theoretical accounts (including some moderate degree of truncated noise).

[1] To make the saturated model identifiable, we constrained σ^2 to be the same for both latent classes.

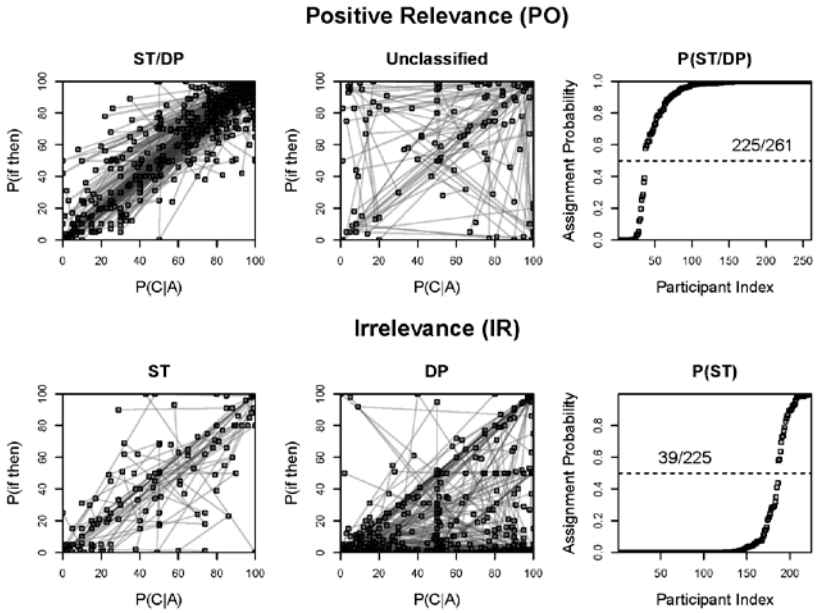

Figure 2. Individual associated to the different Phase 1 classifications, and their respective posterior individual-level classifications (note that in the IR condition, only participants classified as ST/DP in the PO condition were considered).

The key parameters of interest in this analysis are the posterior probabilities of $w_i = 1$ obtained in the PO and IR conditions. In the PO condition, when the mean of this posterior probability was estimated to be below or equal to .50, the individual was classified as following the saturated model. When the mean is estimated to be larger than .50, the individual was classified as following ST/DP. In the IR condition, these same ranges of values led to the ST and DP classifications, respectively.

The individual classifications jointly obtained for PO and IR were used to characterize the conformity of individuals' responses to theoretically-meaningful inequalities, namely UAI and PCh. For participant *i*, the probability that her response to a given item-pair *j* conformed to a given inequality is given by $\Phi(\Delta_i + K_{i,j})$, with $\Phi()$ being the probability function of the standard Normal distribution. Parameter $K_{i,j}$ is a correction term for participant *i* and item-pair *j* such that $\Phi(K_{i,j})$ corresponds to the probability that the responses to a given item-pair were inequality-conforming by chance alone (Singmann, Klauer, & Over, 2014). Parameter Δ_i corresponds to that individual's displacement from chance (i.e., when Δ_i is positive, that individual produces inequality-conforming responses at an above-chance rate). Using a hierarchical framework, these individual parameters were assumed to come from a Normal group-level distribution, with mean μ_Δ and standard deviation σ_Δ. If individuals in general conform to the UAI or PCh, then their respective μ_Δ should be consistently above 0 (i.e., the probability of μ_Δ being below 0 should be very small). These parameters were estimated separately for individuals classified as ST and DP in the IR condition.

A very similar hierarchical approach was used to model the relative probability of an individual judging the no-difference justification (in line with ST) as most convincing after having seen both sides, as well as the relative probability attributing the HIT to such justification.

Results

The posterior-parameter distributions of mixture model were estimated via Gibbs sampling using the general-purpose software JAGS (Plummer, 2003). Chain convergence was confirmed via the R-hat statistic and visual inspection. The individual-level classifications shown in Figure 2 show that the probabilities generated by the majority (225 out of 261) of individuals in the PO condition were in line with ST/DP. In contrast, only a very small group of individuals were in line with ST in the IR condition (39 out of 225); most followed the predictions of DP. The individual data shown in Figure 1 shows that the data classified as ST/DP in the PO condition as well as ST and DP in the IR condition were in line with the model predictions. To address the worry that participants belonging to ST were misclassified as DP, we visually inspected the responses of every participant individually.

The classifications lead to clear differences in both UAI and PCh, as well as in the probability of judging the no-difference justification as most convincing. As shown in Table 1, for UAI the posterior μ_Δ estimates in the IR condition for individuals classified as ST are systematically above 0, but systematically below 0 for individuals classified as DP. In the case of PCh, the posterior μ_Δ estimates were systematically above 0, as expected. The latter result was less clear for ST, but this is expected given the small number of participants classified as being in line with ST.

Finally, the relative probabilities of judging the no-difference justification (consistent with ST) as most convincing and attributing the HIT were drastically different for individuals classified as following ST and DP. These posterior probabilities were considerably larger for ST (see Table 1). Note that these were conditional probabilities of finding the ST justification most convincing, and accepting the ST HIT, given the participants expressed preferences for either ST or DP in phase 2.

	ST Followers (N=39)	DP Followers (N=186)
	μ_Δ	μ_Δ
UAI	0.61 [0.16, 1.11] (72%)	-0.46 [-0.65 -0.28] (47%)
PCh	0.21 [-0.07, 0.51] (68%)	0.14 [0.02, 0.27] (66%)
P(ST mc)	.94 [.78, 1]	.15 [.09, .22]
P(ST HIT)	.92[.77, 1]	.21 [.15, .28]

Table 1. Median group-level posterior parameter estimates (and their respective 95% credibility intervals) obtained in the IR condition. Percentages of responses conforming to UAI and PCh are given in parentheses. The estimates associated to μ_Δ in the PO condition (where participants were classified as ST/DP) were 1.66 [1.14, 2.24] and 1.19 [0.82, 1.61] for UAI and PCh, respectively. 'P(ST mc)' = P(ST most convincing | ST or DP most convincing). 'P(ST HIT)' = P(ST receive HIT | ST or DP receive HIT).

Discussion

In this paper we have presented a novel experimental design to study the reflective attitudes of the participants and an accompanying Bayesian mixture model to study individual variation. We have seen that it is possible to classify the participants according to whether they follow the Suppositional Theory of Conditionals or the Default and Penalty Hypothesis. We then used these classifications to study the participants' performance on the uncertain and-to-if inference task to examine whether the participants consistently followed the assigned interpretation of the conditional in an inference task.

This experimental design gives us a very rich data set that we have not exhausted in this brief note. Nevertheless, the data we did analyze show a very clear pattern. In the PO condition of phase 1, 86% of the participants followed the Equation (PRED_1), whereas only 39 of these participants followed the Equation in the IR condition. The remaining 186 participants showed a clear tendency in the IR condition to assign lower probabilities than if they had treated the P(if A, then C) as a conditional probability. For the 39 ST participants from phase 1 there was a .94 probability that they find the ST character to be most convincing one, conditional on the fact that they had a preference. Of the 186 DP participants in phase 1, this conditional probability was .85, this time in favor of the DP character.

Finally, the participants' performance on the uncertain and-to-if inference task in phase 3 indicated that the participants acted consistently with their assigned interpretation of the conditional. As a theorem of probability theory, the PCh inequality ($P(C|A) \geq P(A,C)$) remains valid for both groups, so they should conform to it at above chance levels irrespectively of the relevance condition. In contrast, whether the participants should conform to the UAI inequality (P(Conclusion) ≥ P(Premise)) in the IR condition, depends on whether they interpret the conditional in the conclusion as a conditional probability.

In the PO condition both groups were above chance levels for conformity to both the UAI and PCh inequalities. For the ST participants, a tendency was found to continue to conform to the UAI and PCh inequalities in the IR condition at above chance levels. (However, the estimates were connected with uncertainty given the modest size of the ST group.) In contrast, for the DP participants an interaction was revealed between relevance and type of inequality in that these participants continued to display conformity to PCh at above chance levels in the IR condition while ceasing to conform to the UAI inequality at above chance levels. The results thus indicate that it was possible to separate two individual profiles in the participants' interpretation of the conditional. For each profile, the participants were shown to behave consistently with their interpretation of the conditional in the uncertain and-to-if inference.

In Skovgaard-Olsen *et al.* (2016b), it was found that the above-chance level conformity to UAI, which Cruz *et al.* (2015) did not generalize to the IR condition. However, since these results were analyzed at the group level, it was hard to tell whether they indicated that the participants were incoherent or whether they followed DP instead. With the present results we have a first indicator that two groups can be identified at the individual level that consistently follow their assigned interpretation of the conditional in the uncertain and-to-if inference.

References

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning, 19*, 308-28.

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology, 89*, 106-133.

Cruz, N., Baratgin, J., Oaksford, M. and Over, D. E. (2015). Bayesian reasoning with 'if's and 'and's and 'or's. *Frontiers in Psychology, 6*, 192.

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

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

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*, 275-305.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods, 48*, 29-41.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003).*

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

Skovgaard-Olsen, N. (2015). The problem of logical omniscience, the preface paradox, and doxastic commitments. *Synthese*.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016a). The Relevance Effect and Conditionals. *Cognition, 150*, 26-36.

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

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