Supplemental Material to:

Relevance and Reason Relations -

Manipulation Check and Selection of Scenarios

Niels Skovgaard-Olsen

University of Konstanz and Albert-Ludwigs-Universität Freiburg

Henrik Singmann

University of Zürich

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. Henrik Singmann, Department of Psychology, University of Zürich, Zürich, Switzerland. Karl Christoph Klauer, Department of Psychology, Albert Ludwigs Universität Freiburg, Freiburg, Germany.

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).

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

1. Experiment 1: Manipulation Check

We first performed a manipulation check to ensure that the numbers the participants provided could be interpreted as probabilities satisfying the axioms of the probability calculus. To this end, the law of total probability, $P(C) = \sum_{i=1}^{n} P(C|A_i)P(A_i)$, was applied to the measurements of P(A), P(C|A), and P(C|¬A) to calculate an ideal value that P(C) should take if the participants were probabilistically consistent. This calculated value for P(C) was then subtracted from the actual value of P(C) supplied by the participants to form a probabilistic consistency scale using the following formula: $1 - |P(C) - [P(C|A) \cdot P(A) + P(C|\bar{A}) \cdot (1 - P(A))]|$. This measure takes on values smaller or equal to one, where a value of one indicates perfect probabilistic consistency. Fig. 1 shows the distribution of mean consistency values for the participants in a boxplot and reveals that participants are surprisingly probabilistically consistent with 75% of the distribution having probabilistic consistency rates of almost .9. Given these results we were confident that participants' responses could be interpreted as probabilities.

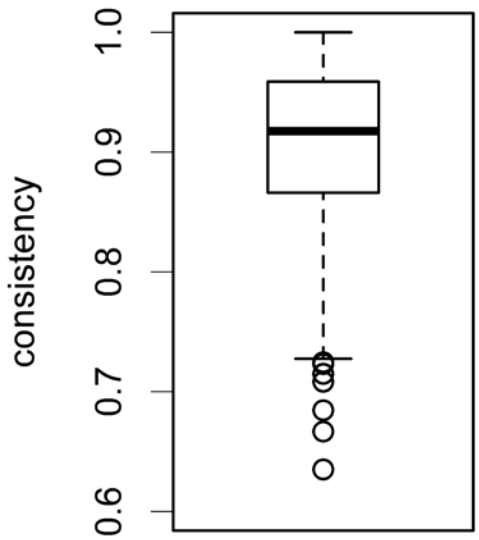

Fig. 1. Probabilistic consistency ratings of the participants based on applying the law of total probability to the probabilities they provided.

2. Experiment 1: Correlation Matrices

Table 1 displays the inter-correlation of the four variables of Experiment 1 and shows that, as expected, all correlations were highly significant. One can also see that, as hypothesized, ΔP seemed to be a better predictor for both relevance and the reason relation than the difference measure.

Table 1. Correlation Matrix and Descriptive Statistics for Measures Obtained in Experiment 1

	Reason relation	ΔP	Difference Measure	Mean (SD)
Relevance	.81	.54	.40	0.73 (2.29)
Reason Relation		.58	.46	0.06 (1.19)
ΔP			.73	0.01 (0.41)
Difference Measure				-0.01 (0.34)

Note. All correlations are highly significant, $p < .0001$. Given the non-independence of data points within participants and within contents, these *p*-values should, however, be read with caution. The ranges for the variables are: directional relevance from -4 to 4, reason relation from -2 to 2, ΔP from -1 to 1.

To appropriately test this hypothesis it is important to consider that the data has replicates both on the level of the participant (since each participant provided one response for each of the 12 within-participant conditions) and on the level of the scenarios (as each scenario could appear in each relevance condition across participants). Due to this dependency structure with conditions repeated within participants and scenarios, standard statistical procedure such as correlation cannot be used. For this reason, a linear mixed model was used in the paper for the analysis.

Out of the six confirmation measures mentioned in Tentori *et al.* (2007), our design only allowed us to test the Keynes and Horwich's ratio measure, $\log(P(C|A)/P(C))$ in addition to the difference measure (which is also listed there). Unfortunately, this measure introduces the problem of extreme ($-\infty$) or undefined values for 24% of our observations. Furthermore, it correlates highly with the difference measure for the reduced sample, $r = .89$:

Table 2. Correlation Matrix and Descriptive Statistics for Measures Obtained in Experiment 1

	Reason relation	ΔP	Difference Measure	Ratio	Mean (SD)
Relevance	.83	.54	.40	.37	0.73 (2.29)
Reason Relation		.57	.43	.40	0.06 (1.19)
ΔP			.69	.61	0.01 (0.41)
Difference Measure				.89	-0.01 (0.34)
Ratio					-0.01 (0.80)

Note. All correlations are highly significant, $p < .0001$. Given the non-independence of data points within participants and within contents, these *p*-values should, however, be read with caution. The ranges for the variables are: directional relevance from -4 to 4, reason relation from -2 to 2, ΔP from -1 to 1.

When adding the ratio measure to our LMM model for Experiment 1, the results indicate that it accounts for no unique variance on its own for either perceived relevance, $F(1, 24.00) = 2.63$, $p = .12$, or perceived reason relation as DV, $F(1, 24.57) = 2.21$, $p = .15$. Indeed, it remains the case that of these three predictors, only ΔP accounts for unique variance for perceived relevance, $F(1, 25.14) = 235.84$, $p < .0001$, and perceived reason relation, $F(1, 22.16) = 216.87$, $p < .0001$.

3. Selection of the Scenarios

For the selection of the scenarios, the full sample of 725 participants was used without applying our exclusion criteria, and as there were no significant differences between the IR_S and IR_D conditions, the difference between them was collapsed for the analysis.

The distinction between these two ways of implementing the irrelevance category was initially introduced in an attempt to implement the notion of ‘topical relevance’ from relatedness logic (Iseminger, 1986; Walton, 2004: ch. 4), which treats two propositions as relevant if they share a subject matter and as irrelevant if they don’t. The way we operationalized this requirement was that two propositions are judged to be relevant, if they concern the same context/content and irrelevant if they didn’t. Accordingly, if Stephen is going on a date, then we

assumed that if two propositions (A, C) both concern preparations for the dating situation then they will share a subject matter, whereas a proposition concerning what Stephen's neighbor likes to eat (B) concerns a different subject matter. Under this assumption, A and C are topically relevant to each other, whereas A and B are topically irrelevant to each other. However, as there were no significant differences between the IR_S and IR_D conditions, the difference between them was collapsed for the analysis, and the IR_D conditions of our stimulus materials were selected for Experiment 2.

To prevent scenario content from becoming a nuisance variable only complete scenarios were selected so that we could ensure that all experimental conditions were represented within each scenario. For each experimental condition, the outputs of the following three equations were z-transformed and the average was taken. This average was used to calculate the 30th percentile with the largest distance from *optimal*: (1) $(\overline{\Delta P} - optimal)^2$, (2) $\left(\overline{P(A)} - optimal\right)^2$, and (3) $\left(\overline{P(C)} - optimal\right)^2$. To illustrate, for the experimental condition IRHH, the optimal value of ΔP would be 0 and the optimal values of P(A) and P(C) would be 100. So to ensure that our selected scenarios were able to implement this experimental condition, the distance of the average ΔP, P(A), and P(C) from these optimal values was used as a selection criterion.

For each scenario, the frequency of its experimental conditions lying within the 30th percentile of the worse experimental conditions was counted. The 30th percentile with the largest number of bad experimental conditions was then used to exclude six complete scenarios. That is to say, scenarios with five or more counts of worse experimental conditions were excluded. The mean frequency of the worse experimental manipulations for excluded scenarios was 5.83, and the mean frequency of worse manipulations for included scenarios was 2.6. In one case, a choice had to be made between two scenarios that both had 5 worse manipulations using boxplots.

In Table 3, summary statistics is shown for the excluded {3, 4, 6, 9, 15, 18} and included scenarios {1, 2, 5, 7, 8, 10, 11, 12, 13, 14, 16, 17}. With ΔP values of almost 0 on average for the IR conditions and the NE and PO conditions differing with ΔP values above |.25| from the IR conditions in the expected directions, the relevance manipulations were successfully implemented. Moreover, with high and low prior manipulations differing on average with |.20| or more from the midpoint of the scale, the priors manipulations was also successfully implemented.

Table 3. Summary statistics of selected scenarios.

	Included	Excluded
PO ΔP mean	.32	.22
NE ΔP mean	-.27	-.21
IR ΔP mean	-.01	.020
Mean high P(A)	.70	.63
Mean low P(A)	.15	.15
Mean high P(C)	.77	.70
Mean low P(C)	.27	.30

However, because complete scenarios were selected some outliers had to be accepted in particular scenarios, which are still in need of further improvement.

References

Iseminger, G. (1986). Relatedness Logic and Entailment. *The Journal of Non-Classical Logic*, 3, 5-23.

Walton, D. (2004). *Relevance in Argumentation*. Mahwah, N.J.: Lawrence Erlbaum Associates.