GLOSSA psycholinguistics

Cremers, A., Fricke, L., & Onea, E. (2022). The importance of being earnest: How truth and evidence affect participants' judgments. *Glossa Psycholinguistics*, 1(1): 23, pp. 1–17. DOI: https://doi.org/10.5070/G6011172

eScholarship University of California

# The importance of being earnest: How truth and evidence affect participants' judgments

**Alexandre Cremers,** FLF, Vilniaus Universitetas, LT, alexandre.cremers@gmail.com

**Lea Fricke,** Institut für Germanistik, Karl-Franzens-Universität Graz, AT; Germanistisches Institut, Ruhr-Universität Bochum, DE, lea.fricke@ruhr-uni-bochum.de

**Edgar Onea,** Institut für Germanistik, Karl-Franzens-Universität Graz, AT, edgar.onea-gaspar@uni-graz.at

Truth-value judgments are one of the most common measures in experimental semantics and pragmatics, yet there is no standardized way to elicit such judgments. Despite anecdotal remarks on how proper choice of prompts or response options could help disentangle pragmatic from semantic effects, little is known regarding the relation between parameters of the task and what it actually measures. We tested a range of prompts and two response options for their sensitivity to truth of the target sentence, prior evidence, and the interaction between these two factors. We found that participants attribute high value to true statements, even when they are not backed by evidence. Moreover, our results confirm that prompts vary wildly in their sensitivity to pragmatic factors, and should allow researchers to make an informed choice depending on what they want to test. There was no difference between the results generated by the response options, although the Likert scale required fewer participants and may therefore be preferable. In addition, we discuss some theoretical consequences of our results for pragmatics, philosophy of language, and social psychology.

# 1. Introduction

Truth-value and acceptability judgment tasks are commonly used in experiments in cognitive science and linguistics (Crain and McKee, 1985, see e.g., Thornton, 2017 for an overview). In designing an experiment, researchers have to make several methodological decisions, among them, the choice of a prompt and response options, which should be well-informed. To this end, we present an experimental study which compares different experimental prompts and the efficiency of two response tasks.

The effect of different response options has been explored in Marty et al. (2020) and references therein, but only for studies of the syntactic well-formedness of sentences. In this case, continuous sliders and Likert scales seem to offer an advantage in statistical power. It has been claimed that multiple or continuous response options may be more sensitive to contextual factors when studying the meaning of sentences (van Tiel, 2014), or that the use of binary response options in experiments can fail to correctly represent the pragmatic abilities of participants (Jasbi et al., 2019; Veenstra & Katsos, 2018). However, there are no systematic investigations of the effect of response options on semantics tasks, and differences in statistical power may explain some of these isolated observations. We aim at closing this research gap.

The choice of a certain prompt relies on the assumption that it targets the aspect of meaning that is to be investigated. For example, the prompt "Is [the speaker] right?" is assumed to address the truth of a statement (e.g., Noveck, 2001). However, prompts may be sensitive to other (pragmatic) factors as well. In particular, the decisions made by participants in experiments may be influenced by whether a statement was assertable given a situation. While assertability has various components, in this paper we focus on Grice's Maxim of Quality.

In Grice's theory of communication (Grice, 1975), which introduces four maxims governing how a cooperative speaker behaves, the first of these maxims – the Maxim of Quality – requires both truth and evidence for a statement. It states "Try to make your contribution one that is true" and is further broken down into two sub-maxims presented in (1) below. The first bans blatant lying and is uncontroversial. The second requires the speaker to have evidence for her statement.

(1) a. Do not say what you believe to be false.
b. Do not say that for which you lack adequate evidence.

According to Grice's Maxim of Quality, truth is not itself a requirement for the assertability of a statement. What is required is positive evidence (1b) and lack of strong negative evidence (1a). Hence, if assertability in the Gricean sense, in addition to truth value, had an impact on the results in truth-value judgment tasks, this would only show up in cases in which a sentence is true but not sufficiently supported by evidence and vice versa, if a sentence is strongly supported by evidence but false.

One could expect certain prompts to be sensitive to the truth of a statement, others to be exclusively sensitive to assertability, and potentially some to be sensitive to both categories. In the present study, we compare different prompts in regard to their sensitivity to truth and the evidence the speaker had when making her statement.

## 2. Experiment

We investigated how naive participants judged statements made by a speaker who had definite or probabilistic information about a situation. We varied whether the statement was factually true or false, and how much evidence the speaker had before making her statement. We tested six different prompts and two possible response options (binary forced choice or Likert scale).

### 2.1. Material and design

Participants had to complete a one-page survey such as the one presented in **Figure 1**. It consisted of a short background story detailing a simple dice game played between two characters, Anna and Kate, to decide who would drive home from a party, the loser of the game being the designated driver, and the winner being determined by the highest dice value. The game was structured in such a way that in extreme cases the winner could be decided after the first person's, Anna's, dice roll, but in most cases the winner could only be decided after both players rolled. The critical sentence in the survey was a statement made by another character – Sue – about the outcome of the game, without waiting for the end of the game. This was followed by four questions (one target and three controls).

We manipulated three factors: the prompt, i.e., our target Question 2, shown in **Figure 1**, the response options to the prompt question as well as the dice outcome for Anna and Kate.

The factor DICE OUTCOME (both Anna and Kate's results) had eight levels, presented in **Table 1**. Depending on Anna's result (2–12), which is the only information available to Sue at the time she makes her statement, the probability that Anna would be the driver ranges from zero (2) to one (12). For now, we take this probability as a direct measure of the "evidence" available to Sue. In the general discussion, we review more sophisticated notions of probabilistic evidence proposed in the literature.[1] If Anna's dice roll does not immediately decide the outcome of the game, there are two possible scenarios: In one, Anna loses and Sue's statement turns out to be false; in the other, Anna wins and Sue's statement is ultimately true.

[1] We calculated the evidence values taking into account that if there is a tie, the game is repeated and thus Anna still has a 50% chance to win. In case Anna's result ensured that she had immediately won (12) or lost (2), the last sentence ("Meanwhile… ") was absent from the survey.

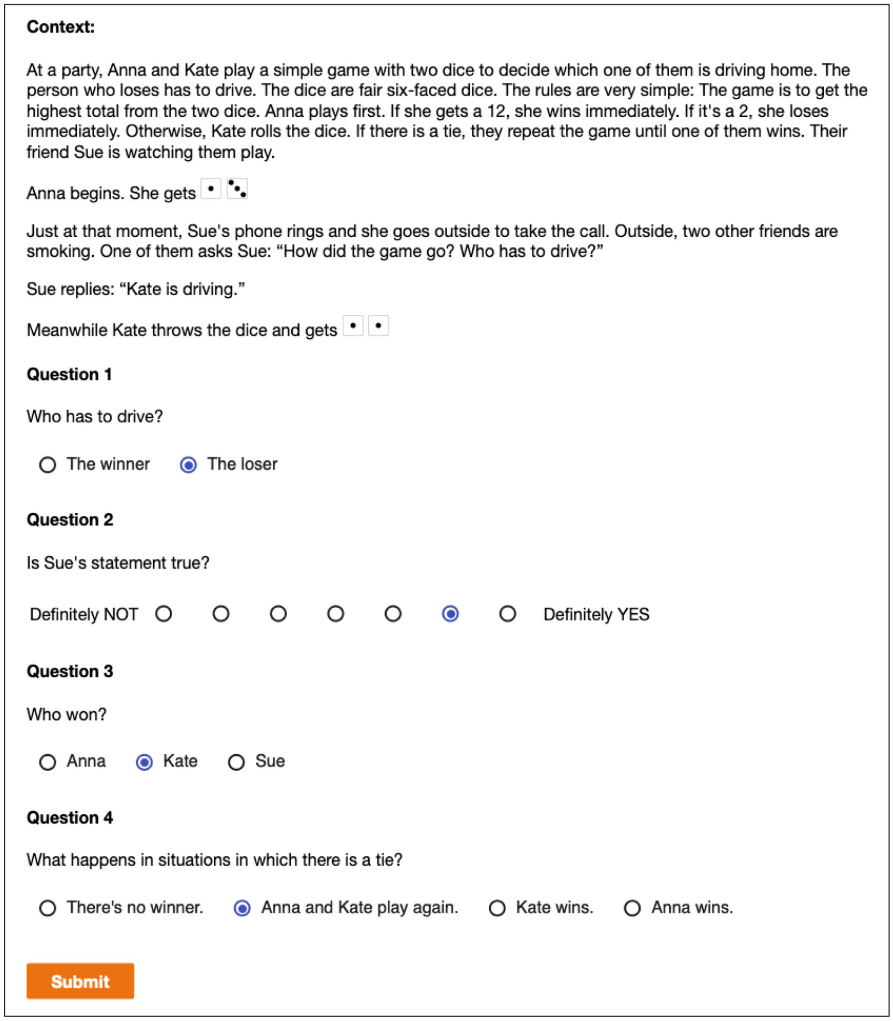



**Figure 1:** Example survey with low evidence but true statement.

**Table 1:** Possible dice outcomes and how they determine the evidence for Sue's statement and its ultimate truth or falsity.

| Anna's total | Available evidence | | Kate's total | Sue's statement |
|---|---|---|---|---|
| 2 | null | 0% | — | false |
| 4 | low | 12.5% | 2 | true |
| | | | 7 | false |
| 7 | medium | 50% | 4 | true |
| | | | 10 | false |
| 10 | high | 87.5% | 7 | true |
| | | | 12 | false |
| 12 | perfect | 100% | — | true |

The six different PROMPTS are listed in **Table 2**. We expected the first three prompts to be more sensitive to factual truth, and the other three to be more susceptible to target evidence, i.e., the probability that the statement is true given the information available to the speaker. The two RESPONSE OPTIONS were a binary forced choice judgment (*Yes* vs. *No*) or a 7-point Likert scale with the endpoints labelled as *Definitely NOT* and *Definitely YES*.

**Table 2:** Prompts used in the target Question 2.

| Level | Prompt |
|---|---|
| true | Is Sue's statement true? |
| correct | Was the answer correct? |
| right | Is Sue right? |
| trustworthy | Based on her behavior in this situation, do you consider Sue trustworthy? |
| natural | Was Sue's statement natural in this context? |
| justified | Was Sue justified in saying that? |

The three control questions tested story comprehension, in particular, whether participants correctly associated losing the game with having to drive. In all, the survey took about 3 minutes to complete.

## 2.2. Participants

### 2.2.1. Test

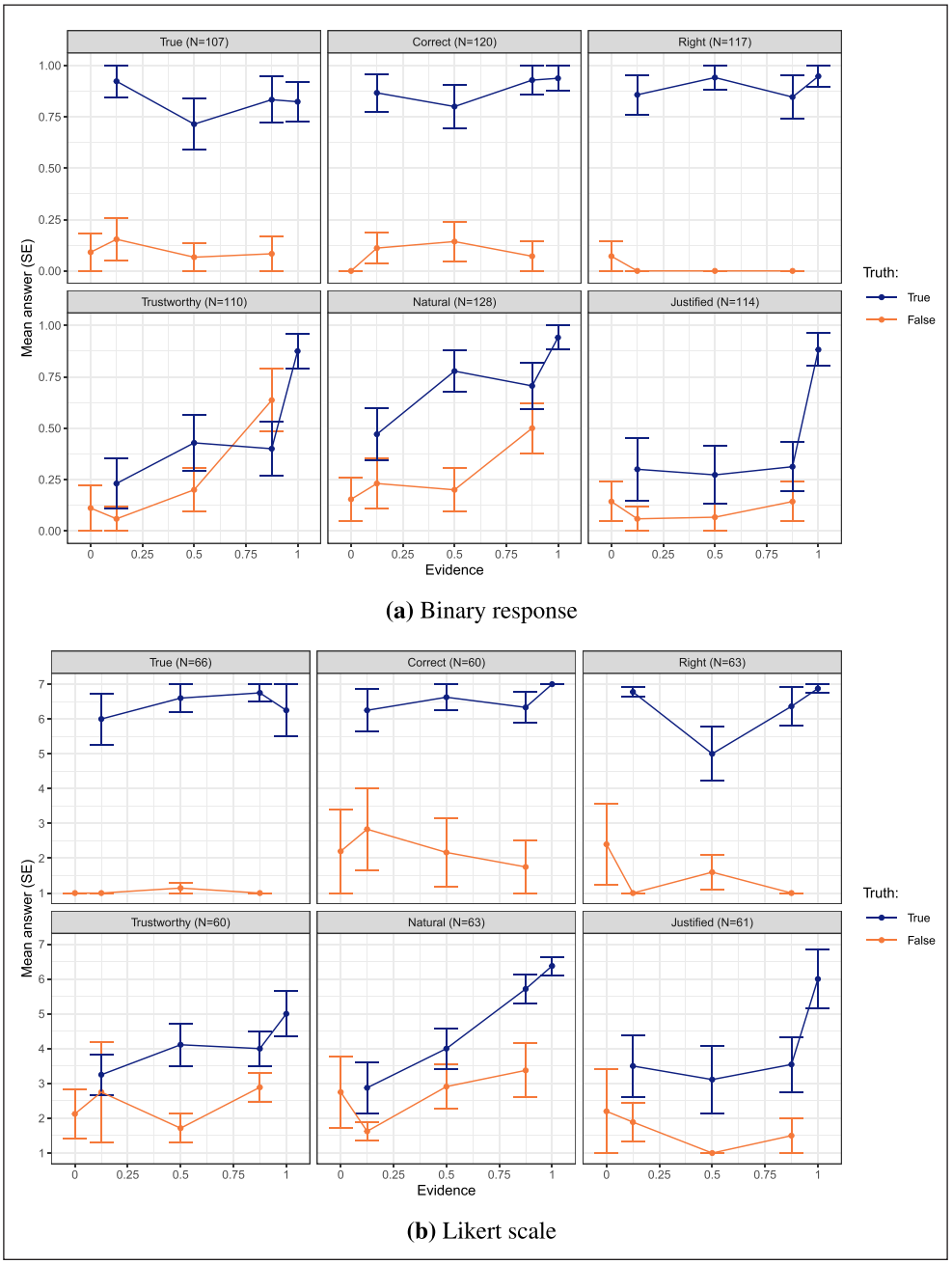
For each combination of degree of evidence, truth, and prompt, we aimed to recruit 20 participants for the binary response option and 10 for the Likert scale option (total: 1440 participants). These sample sizes are conservative estimates based on a pilot study with 530 participants, which showed that the binary option was noisier and therefore required about twice more participants than the Likert scale per condition (in line with previous findings of Marty et al., 2020). We obtained data from 1364 unique participants recruited via Amazon Mechanical Turk, who were paid 47ct as compensation.[2] The average error rate on control questions was 8.6%. For further analysis, we only consider the 1069 participants who passed all 3 control questions.

## 2.3. Results

The results for each prompt are presented in **Figure 2**. There, the mean answer (with SE) for each prompt question is given in a separate graph depending on the evidence level that Sue had for her assertion about the winner of the game. The evidence levels equal the probability that her statement

[2] The discrepancy with our aim of 1440 comes from participants who took the survey more than once, and whose subsequent takes were excluded from the data set.

is true as detailed in **Table 1**. The eventual truth or falsity of Sue's statement is coded by color. Recall that Sue's statement was made without knowledge of Kate's dice outcome, hence truth and evidence are conceptually different factors only amounting to the same in the extremes (the left and right ends of the graphs). The binary and Likert scale responses are presented in separate blocks.




**(a)** Binary response

**(b)** Likert scale

**Figure 2:** Mean answer (with SE), for each prompt, as a function of evidence level and truth.

Likert scale and binary responses gave very similar results. Moreover, some of the prompts (*true*, *correct*, *right*) are very similar and only sensitive to truth and not to evidence. By contrast, the remaining three prompts (*trustworthy*, *natural*, and *justified*) seem to include assertability effects in terms of available evidence; i.e., the results depend on the information available to the speaker when making her statement. For the three other prompts, we observe that assertability depends on the evidence, linearly for *trustworthy* and *natural*; increasing abruptly near 100% for *justified*. In the next section, we use Bayesian modeling to provide a comprehensive statistical picture of these results.

# 3. Data analysis

## 3.1. Descriptive analysis

We first fitted a descriptive model to each combination of prompt and response option using Stan (Carpenter et al., 2017). For binary responses, we used logistic regression; for Likert scale, proportional odds logistic regression. The models included 5 predictors:

- Main effect of truth; coded as 0 or 1.
- Linear effect of probabilistic evidence; z-scored.
- Interaction between truth and probabilistic evidence.
- Categorical positive evidence; 1 if the evidence is 100%, 0 otherwise.
- Categorical negative evidence; 1 if the evidence is 0%, 0 otherwise.

The two categorical evidence effects are meant to model the fact that there is a categorical difference between partial (probabilistic) information and complete certainty regarding the truth or falsity of the statement. Concretely, these two effects allow discontinuity at each end of the evidence spectrum.[3] In **Table 3**, we present the posterior estimates for the 5 parameters, for each of the prompt-response option combinations. The Appendix contains a discussion of effect sizes and a comparison between the statistical power of binary vs. Likert-scale responses.

## 3.2. Theory-driven categorization

We model the data based on a range of theoretical models that reflect various types of truth-oriented or evidence-oriented behavior. In the first and simplest model, which we dubbed the *truth*-model, participants are expected to be sensitive only to the truth or falsity of the statement. In this model, we set the priors of all other factors very close to 0. The next model is the *assertability*-model, which requires only an effect of the categorical positive evidence, all other factors being set close to zero in the priors. This model directly follows from the Maxim of Quality. We do not necessarily assume that the categorical positive evidence needs to be equated with a strict sense

[3] We thank an anonymous reviewer and the editor for this suggestion.

of knowledge: while the 87.5% probability we tested (when a 10 is rolled) would certainly not qualify as categorical evidence for winning, a much higher probability just under 1 (say, 99%) could maybe count as definitive evidence. The final model is the *(probabilistic) evidence*-model, which includes only an effect of the probabilistic evidence, all other factors being set close to 0. This model could have various explanations. One is that making a statement which has a high probability of being true can be a rational human behaviour in certain conditions, since a statement with a high probability of being true will often end up being true. On the other hand, making a lower probability statement can be interpreted as a stronger violation of the Maxim of Quality, which could be reflected in participants' judgments.

**Table 3:** Posterior parameter estimates from the descriptive models (mean and 95% credible interval) for each combination of prompt and response option. Parameters for which the credible interval does not include 0 are highlighted.

| Prompt | Response | Parameter | mean | 2.5% | 97.5% |
|---|---|---|---|---|---|
| True | binary | $\beta_{truth}$ | 4.15 | 2.77 | 5.75 |
| | | $\beta_{evidence}$ | −0.36 | −1.24 | 0.47 |
| | | $\beta_{truth \times ev}$ | 0.07 | −1.57 | 1.80 |
| | | $offset_0$ | 0.56 | −1.82 | 3.02 |
| | | $offset_1$ | 0.47 | −1.61 | 2.54 |
| | Likert | $\beta_{truth}$ | 8.64 | 4.77 | 13.38 |
| | | $\beta_{evidence}$ | 0.88 | −4.06 | 5.79 |
| | | $\beta_{truth \times ev}$ | 1.19 | −4.19 | 6.70 |
| | | $offset_0$ | 0.67 | −2.68 | 4.23 |
| | | $offset_1$ | −0.17 | −3.44 | 3.27 |
| Correct | binary | $\beta_{truth}$ | 4.42 | 3.10 | 5.93 |
| | | $\beta_{evidence}$ | 0.13 | −0.66 | 0.93 |
| | | $\beta_{truth \times ev}$ | 0.31 | −1.24 | 1.91 |
| | | $offset_0$ | 1.78 | −0.82 | 4.79 |
| | | $offset_1$ | 0.75 | −2.04 | 3.98 |
| | Likert | $\beta_{truth}$ | 4.64 | 2.96 | 6.52 |
| | | $\beta_{evidence}$ | −0.24 | −1.14 | 0.63 |
| | | $\beta_{truth \times ev}$ | 0.77 | −0.95 | 2.56 |
| | | $offset_0$ | 0.88 | −1.56 | 3.36 |
| | | $offset_1$ | 4.49 | −0.32 | 11.47 |
| Right | binary | $\beta_{truth}$ | 9.68 | 5.44 | 16.22 |
| | | $\beta_{evidence}$ | −1.12 | −3.71 | 0.80 |
| | | $\beta_{truth \times ev}$ | 2.13 | −1.67 | 7.15 |
| | | $offset_0$ | −0.90 | −4.22 | 2.26 |
| | | $offset_1$ | 1.17 | −1.66 | 4.44 |
| | Likert | $\beta_{truth}$ | 6.11 | 4.10 | 8.52 |
| | | $\beta_{evidence}$ | −0.19 | −1.24 | 0.75 |
| | | $\beta_{truth \times ev}$ | 0.51 | −1.37 | 2.58 |
| | | $offset_0$ | −1.16 | −3.75 | 1.37 |
| | | $offset_1$ | 1.53 | −0.96 | 4.57 |

| Prompt | Response | Parameter | mean | 2.5% | 97.5% |
|---|---|---|---|---|---|
| Trustworthy | binary | $\beta_{truth}$ | 0.54 | −0.49 | 1.62 |
| | | $\beta_{evidence}$ | 1.03 | 0.40 | 1.71 |
| | | $\beta_{truth \times ev}$ | −1.27 | −2.61 | −0.02 |
| | | $offset_0$ | −0.77 | −3.23 | 1.91 |
| | | $offset_1$ | 2.21 | 0.38 | 4.31 |
| | Likert | $\beta_{truth}$ | 2.24 | 1.01 | 3.51 |
| | | $\beta_{evidence}$ | 0.78 | 0.10 | 1.47 |
| | | $\beta_{truth \times ev}$ | −0.38 | −1.73 | 0.94 |
| | | $offset_0$ | −0.61 | −2.82 | 1.55 |
| | | $offset_1$ | 0.74 | −1.09 | 2.60 |
| Natural | binary | $\beta_{truth}$ | 1.59 | 0.72 | 2.51 |
| | | $\beta_{evidence}$ | 0.63 | 0.11 | 1.18 |
| | | $\beta_{truth \times ev}$ | −0.15 | −1.23 | 0.92 |
| | | $offset_0$ | 0.02 | −1.89 | 2.04 |
| | | $offset_1$ | 1.71 | −0.59 | 4.64 |
| | Likert | $\beta_{truth}$ | 1.68 | 0.59 | 2.80 |
| | | $\beta_{evidence}$ | 1.15 | 0.51 | 1.82 |
| | | $\beta_{truth \times ev}$ | 0.75 | −0.52 | 2.02 |
| | | $offset_0$ | −0.83 | −2.95 | 1.36 |
| | | $offset_1$ | 0.62 | −1.31 | 2.60 |
| Justified | binary | $\beta_{truth}$ | 1.57 | 0.32 | 2.92 |
| | | $\beta_{evidence}$ | 0.23 | −0.49 | 0.97 |
| | | $\beta_{truth \times ev}$ | −0.30 | −1.73 | 1.15 |
| | | $offset_0$ | −0.87 | −3.25 | 1.38 |
| | | $offset_1$ | 2.96 | 1.12 | 5.09 |
| | Likert | $\beta_{truth}$ | 2.40 | 0.87 | 4.29 |
| | | $\beta_{evidence}$ | −0.31 | −1.28 | 0.53 |
| | | $\beta_{truth \times ev}$ | 0.78 | −0.88 | 2.72 |
| | | $offset_0$ | 0.33 | −2.03 | 2.84 |
| | | $offset_1$ | 2.44 | 0.38 | 4.71 |

In order to find out which model fits the data best for each of the factor combinations, all models were compared with each other by Bayes factor. **Table 4** lists the detailed results of the model comparisons. There, we take 3 and 30 as Bayes factor thresholds for strong and decisive evidence for a model, respectively. In most cases, there is a clear best model for a given combination, and the choice between binary and Likert does not affect the pattern of responses. The only exceptions are the combinations Trustworthy-Likert and Justified-Likert. For the Trustworthy-Likert combination,

the best model is the evidence model, however, with a Bayes factor of 2.9 this model is only marginally superior to the truth model. The fact that the binary data support the evidence model for trustworthy should, however, provide some additional evidence for this model. In the Justified-Likert case, the truth and assertability models scored nearly equally (BF = 1.1), indicating that there is indeed both an effect of truth and an effect of categorical evidence.

**Table 4:** Model comparison for each pair of prompt and response options. The models are ordered by marginal log-likelihood and we indicate the Bayes factor in favor of the best model against the second best ($BF_{12}$), and in favor of the second against the third ($BF_{23}$).

| **Prompt** | **Response** | **Models** |
|---|---|---|
| True | Binary | truth $\gg$ evidence $\approx$ assertability<br>$BF_{12} = 2.8e6$ $BF_{23} = 1.2$ |
| | Likert | truth $\gg$ evidence $>$ assertability<br>$BF_{12} = 4.7e10$ $BF_{23} = 4.8$ |
| Correct | Binary | truth $\gg$ evidence $\approx$ assertability<br>$BF_{12} = 3.7e7$ $BF_{23} = 1.2$ |
| | Likert | truth $\gg$ assertability $\gg$ evidence<br>$BF_{12} = 1.2e5$ $BF_{23} = 24$ |
| Right | Binary | truth $\gg$ assertability $\gg$ evidence<br>$BF_{12} = 1.1e10$ $BF_{23} = 1.5$ |
| | Likert | truth $\gg$ assertability $\approx$ evidence<br>$BF_{12} = 1.6e8$ $BF_{23} = 1.6$ |
| Trustworthy | Binary | evidence $\gg$ assertability $\gg$ truth<br>$BF_{12} = 130$ $BF_{23} = 98$ |
| | Likert | evidence $\approx$ truth $>$ assertability<br>$BF_{12} = 2.9$ $BF_{23} = 12$ |
| Natural | Binary | evidence $>$ assertability $\approx$ truth<br>$BF_{12} = 13$ $BF_{23} = 1.6$ |
| | Likert | evidence $\gg$ assertability $\approx$ truth<br>$BF_{12} = 512$ $BF_{23} = 2.8$ |
| Justified | Binary | assertability $\gg$ evidence $\approx$ truth<br>$BF_{12} = 43$ $BF_{23} = 1.7$ |
| | Likert | truth $\approx$ assertability $>$ evidence<br>$BF_{12} = 1.1$ $BF_{23} = 4.2$ |

Accordingly, **Figure 3** shows the fit of the best model for each combination.

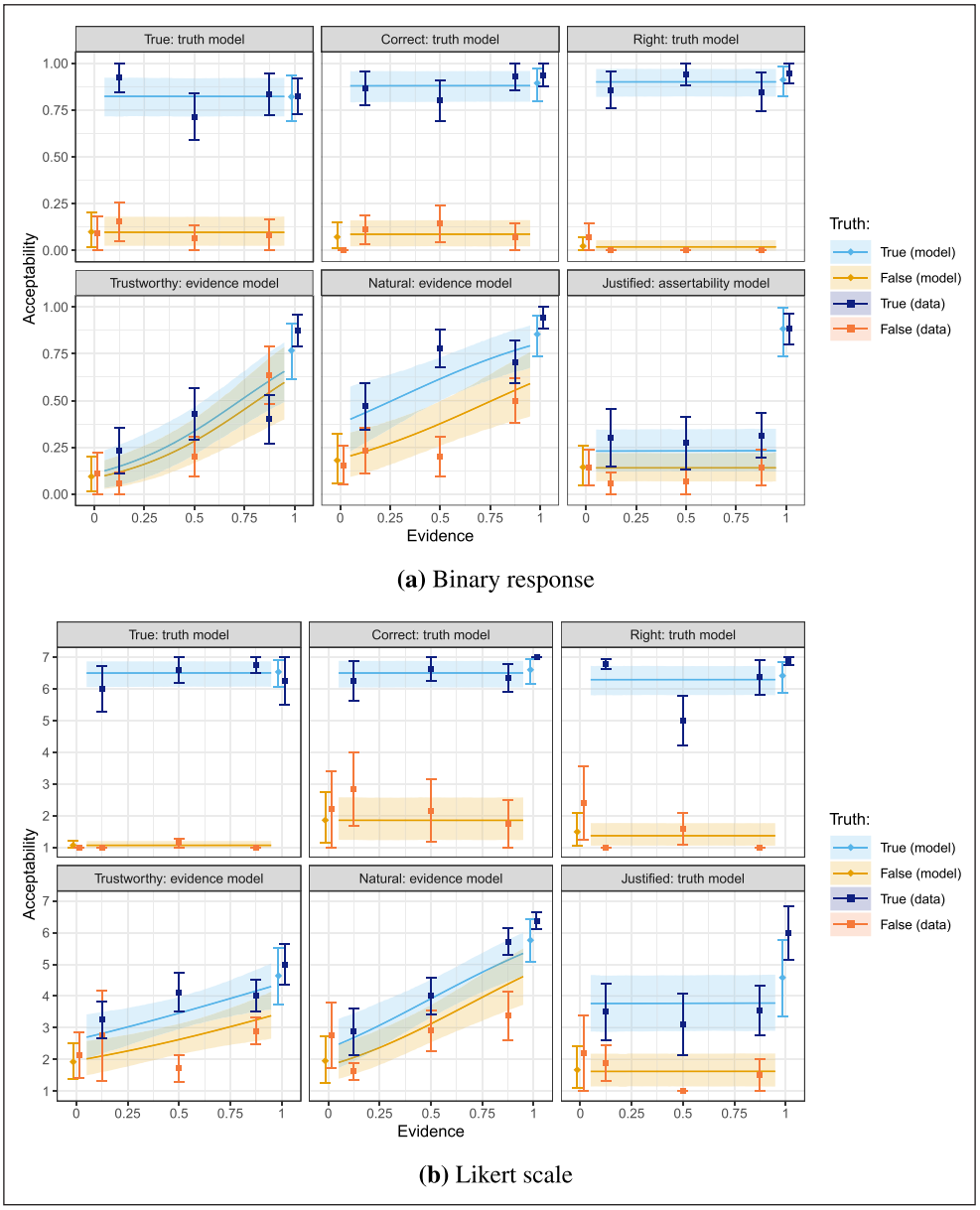



**Figure 3:** Predictions of the best model for each combination of prompt and response option (ribbons and light diamonds, indicating predicted expected value and 95% credible interval), compared to the data (darker squares with error bars, indicating mean and SE).

## 4. Discussion

The main consequences of these results are methodological. The choice of a prompt when designing an experiment should be well-informed in order to fit the goal of the experimenter.

Different prompts will test very distinct aspects of communication, and so this choice cannot be arbitrary.

We found that *true*, *correct*, and *right* all are well-suited for truth-value judgment tasks as they clearly ignore considerations of assertability in terms of Grice's Maxim of Quality. By contrast, *natural* and *justified* appear to tap into issues typically bearing on pragmatic considerations of assertability, though they target different aspects of assertability (probabilistic vs. categorical evidence) and both are to some extent sensitive to factual truth as well. The *trustworthy* prompt cannot be recommended for use in linguistic experiments as it is sensitive to a complex mix of truth, evidence and their interaction, and its behavior may even depend on the response options.

We further found some evidence for the observation in the literature that a Likert scale judgment may be a more economical method as compared to a binary forced choice judgment. Indeed, we found little qualitative or quantitative difference between the two types of response options (the estimates and credible intervals are very similar in **Table 3**), but as shown in the Appendix, the Likert scale yields larger effect sizes than the binary response, and could often achieve the same statistical power with a much smaller sample size. The only qualitative difference happened with the *trustworthy* prompt, but we think that this reflects a problem with the prompt itself rather than a meaningful difference between the two response options, as this prompt gave noisier results than *natural* and *justified* overall.

However, we do acknowledge that it is not obvious to what extent our findings generalize to other types of semantic/pragmatic phenomena, or to task paradigms in which the speaker and the experimental participants both have the same knowledge about past or future situations. It also remains to be shown how the different prompts interact with survey features we did not manipulate in our experiment, but we have at least shown that they can behave very differently in our particular setup. Regarding the truth-sensitive prompts, we do not expect much variation, but the *natural* and *justified* prompts could be more sensitive to manipulations of the context, which is the point if they are meant to test pragmatic effects.

One may also question our operationalization of "$E$ is evidence for $A$" in terms of conditional probability $P(A|E)$. Indeed, more sophisticated notions which have been shown to better capture the use of conditionals (e.g., evidential support $P(A|E) - P(A)$, Douven, 2008, or contingency $P(A|E) - P(A|\neg E)$, van Rooij and Schulz, 2021) could also be better suited to capture evidence in principle, but would not affect the model selection/data interpretation eventually. In our setup, these notions would shift the evidence scale in the following way: what counts as 50% evidence according to our coding would be re-coded as zero evidence; values below 50% on our scale would, accordingly, count as negative evidence; however, the crucial end-point, the 100% value on the evidence scale, would remain the same. Hence, such alternative models would not predict a categorical shift at the 100% mark. But precisely such a categorical shift would be needed to explain the *justified* prompt in terms of evidence alone. Hence, a more complicated coding of

evidence would not have made any substantive difference, and in particular, would not offer a better explanation for the behavior of the *justified* prompt.

A further finding is that truth matters for all prompts, except for the *trustworthy*-binary combination. Even the *justified* and *natural* prompts, which we thought would only target evidence, are sensitive to truth. This premium given to statements that turned out to be true despite a lack of evidence raises important questions for social psychology and may explain well-known confirmation biases (Nickerson, 1998; Wason, 1960). For example, if participants assign more weight to statements that they judge true, irrespective of the evidence the speaker had in support for or against them, it would be difficult to revise false beliefs. Any argument against said belief would be deemed false and therefore less acceptable, even when the person is aware of the evidence supporting the argument.

One possible explanation for the high impact of truth in our experiment is that some participants substituted the question about the naturalness of the statement or the trustworthiness of the speaker, which might not have been so straightforward to answer for them, with an easier question, namely the question whether the statement is true (cf. Strack et al., 1988). The reason behind this valuing of truth over evidence may simply be that what one is immediately interested in is truth. Conceivably, occurrences like, for example, guessing the stock markets foster the association between success and truth. Giving merit to accidentally being right can be classified as an instance of outcome bias, the tendency to evaluate the quality of a past decision based on its outcome (Baron & Hershey, 1988).

However, if participants did in fact judge the naturalness and trustworthiness, this observation may have important implications on society and political discourse, which need to be investigated in future research. It is of relevance to learn more about the factors that affect whether people consider political leaders and experts trustworthy. After all, whatever these factors may be, final action performed by the population, e.g., in terms of election votes, depends on such judgments.

Our findings also inform a debate in philosophy of language around “norms of assertion”. The central issue of this discussion is to define the conditions under which assertions are proper. The main positions in this debate are the knowledge account (Williamson, 1996, i.a.) and the truth account (Weiner, 2005). The former posits that a proposition $p$ can be asserted only if the speaker knows that $p$. On the truth account, $p$ is assertible if it is true. According to Weiner (2005), the speaker must also have reason to believe $p$, but statements based on an epistemic state that is less than knowledge are permissible. A number of experiments provide support for the knowledge norm (Turri, 2013, 2015; Turri & Buckwalter, 2017). In these experiments, participants had to judge whether a statement should be made given different contexts, which varied regarding whether the speaker had knowledge, how strong the evidence was, and whether the statement was true. In contrast, we asked participants to judge a statement that had already

been made. Most importantly, our results show that different prompts can lead to markedly different responses to this question, although truth is a decisive factor. Incidentally, the *justified* prompt reflects Grice's Maxim of Quality: an utterance is considered truly "justified" if and only if it is both true and supported by evidence.

To conclude, we found that humans attribute surprisingly high value to true statements, even when they are not backed by evidence. The extent to which evidence is considered when judging a statement in a linguistic experiment crucially depends on the prompt that is used. Further, the comparison between two response tasks showed that Likert scales constitute a more efficient choice than binary forced choice options for truth-value judgments, as had been previously shown for syntactic acceptability judgments. Future experimental research in the domain of semantics and pragmatics can profit from these methodological insights, as they might contribute to making more informed decisions when designing an experiment.

## Appendix: Statistical power and effect sizes comparisons

**Table 5** gives the estimated ordinal superiority $\gamma$ for each parameter (Ryu & Agresti, 2008). Given two ordinal or binary variables $Y_1$ and $Y_2$ on the same scale, $\gamma$ is defined as:

$$\gamma = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$$

A value of 0.5 indicates no difference between $Y_1$ and $Y_2$. Ordinal superiority provides a measure of effect size for binary and ordinal scales (i.e., it is independent of sample size), and allows us to directly compare the binary and Likert scales. Crucially, Table 5 indicates that for almost all significant effects with the exception of the Trustworthy prompt (where there are qualitative differences between Likert and binary), the Likert scale yields larger effect sizes $\gamma$ than the binary response option. As a consequence, smaller sample sizes are needed to detect the same effects.

**Table 5:** Mean posterior estimates of ordinal superiority $\gamma$ for each factor and combination of prompt and response option and – when $\gamma$ is significantly different from 0.5 (highlighted in blue and orange to facilitate comparison) – estimated minimal sample size to detect an effect at $\alpha = 0.05$ with probability 80%.

| Prompt | Response | Parameter | $\hat{\gamma}$ | $N_{\text{min}}$ |
|---|---|---|---|---|
| True | binary | truth | 0.87 | 12 |
| | | evidence | 0.46 | |
| | | truth×ev | 0.50 | |
| | | $\text{offset}_0$ | 0.57 | |
| | | $\text{offset}_1$ | 0.54 | |
| | Likert | truth | 0.99 | 6 |
| | | evidence | 0.58 | |
| | | truth×ev | 0.52 | |
| | | $\text{offset}_0$ | 0.54 | |
| | | $\text{offset}_1$ | 0.49 | |
| Correct | binary | truth | 0.89 | 10 |
| | | evidence | 0.52 | |
| | | truth×ev | 0.52 | |
| | | $\text{offset}_0$ | 0.65 | |
| | | $\text{offset}_1$ | 0.52 | |
| | Likert | truth | 0.94 | 8 |
| | | evidence | 0.46 | |
| | | truth×ev | 0.56 | |
| | | $\text{offset}_0$ | 0.62 | |
| | | $\text{offset}_1$ | 0.58 | |
| Right | binary | truth | 0.94 | 8 |
| | | evidence | 0.48 | |
| | | truth×ev | 0.52 | |
| | | $\text{offset}_0$ | 0.52 | |
| | | $\text{offset}_1$ | 0.54 | |
| | Likert | truth | 0.97 | 6 |
| | | evidence | 0.47 | |
| | | truth×ev | 0.54 | |
| | | $\text{offset}_0$ | 0.49 | |
| | | $\text{offset}_1$ | 0.60 | |

| Prompt | Response | Parameter | $\hat{\gamma}$ | $N_{\text{min}}$ |
|---|---|---|---|---|
| Trustworthy | binary | truth | 0.55 | |
| | | evidence | 0.60 | 118 |
| | | truth×ev | 0.44 | 348 |
| | | $\text{offset}_0$ | 0.50 | |
| | | $\text{offset}_1$ | 0.70 | 42 |
| | Likert | truth | 0.80 | 28 |
| | | evidence | 0.62 | 202 |
| | | truth×ev | 0.47 | |
| | | $\text{offset}_0$ | 0.48 | |
| | | $\text{offset}_1$ | 0.60 | |
| Natural | binary | truth | 0.69 | 46 |
| | | evidence | 0.58 | 210 |
| | | truth×ev | 0.49 | |
| | | $\text{offset}_0$ | 0.53 | |
| | | $\text{offset}_1$ | 0.57 | |
| | Likert | truth | 0.74 | 40 |
| | | evidence | 0.68 | 70 |
| | | truth×ev | 0.56 | |
| | | $\text{offset}_0$ | 0.45 | |
| | | $\text{offset}_1$ | 0.58 | |
| Justified | binary | truth | 0.60 | 118 |
| | | evidence | 0.52 | |
| | | truth×ev | 0.49 | |
| | | $\text{offset}_0$ | 0.50 | |
| | | $\text{offset}_1$ | 0.78 | 22 |
| | Likert | truth | 0.74 | 44 |
| | | evidence | 0.47 | |
| | | truth×ev | 0.54 | |
| | | $\text{offset}_0$ | 0.57 | |
| | | $\text{offset}_1$ | 0.79 | 28 |

## Data accessibility statement

The data and analysis scripts are publicly available on GitHub: https://github.com/Alex-Cremers/TruthEvidenceJudgments.

## Ethics and consent

The experimental research presented in this article was approved by the ethics committee of the University of Graz. The reference number of the approval is GZ. 39/41/63 ex 2019/20. Informed consent to participate in the study was obtained from participants.

## Funding information

This research has been supported by the LingLab program of the University of Graz. A.C. has been supported by a grant from the Research Council of Lithuania and European Social Fund under Measure 09.3.3-LMT-K-712. L.F. and E.O. have been supported by the project "Exhaustiveness in embedded questions across languages", Priority Program SPP 1727, XPRAG.de. The project also benefited from an Arqus Twinning grant between the universities of Graz and Vilnius.

## Acknowledgements

We thank Svitlana Antonyuk, Boban Arsenijevic, Emmanuel Chemla, Petra Hödl and Swantje Tönnis for helpful discussion of this work, and we thank the editor, Florian Schwarz, as well as four anonymous reviewers for their detailed and constructive feedback on earlier versions of this paper. All remaining errors are, of course, our own.

## Competing interests

The authors have no competing interests to declare.

## Author contributions

A.C., L.F. and E.O. designed the experiment. L.F. assumed administrative tasks. A.C. and E.O. programmed the experiment. A.C. and E.O. designed data analysis strategy. A.C. conducted the experiment and performed statistical data analysis. A.C., L.F. and E.O. wrote the manuscript. E.O. supervised the project. All authors are responsible for the content of the report.

## References

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4). DOI: https://doi.org/10.1037//0022-3514.54.4.569

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. DOI: https://doi.org/10.18637/jss.v076.i01

Crain, S., & McKee, C. (1985). Acquisition of structural restrictions on anaphora. *Proceedings of the North Eastern Linguistic Society*, *16*.

Douven, I. (2008). The evidential support theory of conditionals. *Synthese*, *164*(1), 19–44. DOI: https://doi.org/10.1007/s11229-007-9214-5

Grice, P. (1975). Logic and conversation. In D. Davidson & G. Harman (Eds.), *The logic of grammar* (pp. 64–75). Dickenson. DOI: https://doi.org/10.1163/9789004368811_003

Jasbi, M., Waldon, B., & Degen, J. (2019). Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology*, *10*, 37–48. DOI: https://doi.org/10.3389/fpsyg.2019.00189

Marty, P., Chemla, E., & Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics*, *5*(1), 72. DOI: https://doi.org/10.5334/gjgl.980

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. DOI: https://doi.org/10.1037/1089-2680.2.2.175

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. DOI: https://doi.org/10.1016/S0010-0277(00)00114-1

Ryu, E., & Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, *27*(10), 1703–1717. DOI: https://doi.org/10.1002/sim.3079

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*(5), 429–442. DOI: https://doi.org/10.1002/ejsp.2420180505

Thornton, R. (2017). The truth-value judgment task: An update. In M. Nakayama, Y. Su, & A. Huang (Eds.), *Studies in Chinese and Japanese language acquisition: In honor of Stephen Crain* (pp. 13–39). John Benjamins. DOI: https://doi.org/10.1075/lald.60.02tho

Turri, J. (2013). The test of truth: An experimental investigation of the norm of assertion. *Cognition*, *129*, 279–291. DOI: https://doi.org/10.1016/j.cognition.2013.06.012

Turri, J. (2015). Knowledge and the norm of assertion: A simple test. *Synthese*, *192*, 385–392. DOI: https://doi.org/10.1007/s11229-014-0573-4

Turri, J., & Buckwalter, W. (2017). Descartes's schism, Locke's reunion: Completing the pragmatic turn in epistemology. *American Philosophical Quarterly*, *54*(1), 25–45. DOI: https://doi.org/10.2307/44982122

van Rooij, R., & Schulz, K. (2021). Conditionals as representative inferences. *Axiomathes*, *31*(3), 437–452. DOI: https://doi.org/10.1007/s10516-020-09477-9

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177. DOI: https://doi.org/10.1093/jos/fft002

Veenstra, A., & Katsos, N. (2018). Assessing the comprehension of pragmatic language: Sentence judgment tasks. In K. P. S. Andreas H. Jucker & W. Bublitz (Eds.), *Methods in pragmatics* (pp. 257–279). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110424928-010

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140. DOI: https://doi.org/10.1080/17470216008416717

Weiner, M. (2005). Must we know what we say? *The Philosophical Review*, *114*, 227–251. DOI: https://doi.org/10.1215/00318108-114-2-227

Williamson, T. (1996). Knowing and asserting. *The Philosophical Review*, *105*, 489–523. DOI: https://doi.org/10.2307/2998423