Proceedings of the Society for Computation in Linguistics

Volume 3 Article 3

2020

Modeling Behavior in Truth Value Judgment Task Experiments

Brandon Waldon
Stanford University, bwaldon@stanford.edu

Judith Degen
Stanford University, jdegen@stanford.edu

Recommended Citation
Waldon, Brandon and Degen, Judith (2020) "Modeling Behavior in Truth Value Judgment Task Experiments," *Proceedings of the Society for Computation in Linguistics*: Vol. 3 , Article 3.
DOI: https://doi.org/10.7275/sg32-aq30
Available at: https://scholarworks.umass.edu/scil/vol3/iss1/3

Modeling Behavior in Truth Value Judgment Task Experiments

Brandon Waldon
Stanford University
bwaldon@stanford.edu

Judith Degen
Stanford University
jdegen@stanford.edu

Abstract

Truth Value Judgment Task experiments (TVJTs) are a common means of investigating pragmatic competence, particularly with regards to scalar inference. We present a novel quantitative linking function from pragmatic competence to participant behavior on TVJTs, based upon a Bayesian probabilistic model of linguistic production. Our model captures a range of observed phenomena on TVJTs, including intermediate responses on a non-binary scale, population and individual-level variation, participant endorsement of false utterances, and variation in response due to so-called scalar diversity.

1 Introduction

In Truth Value Judgment Task experiments (TJVTs), participants are asked whether a given sentence is, e.g., 'right' or 'wrong' (or 'true' or 'false', etc.), often in a context of evaluation. In the field of experimental pragmatics, participant judgments in TVJT paradigms have been particularly important for investigating pragmatic competence, especially as it relates to scalar implicature (Noveck, 2001; Noveck and Posada, 2003; Bott and Noveck, 2004; De Neys and Schaeken, 2007; Geurts and Pouscoulous, 2009; Chemla and Spector, 2011; Degen and Goodman, 2014; Degen and Tanenhaus, 2015). On the traditional view of pragmatic competence and its link to TVJT responses, scalar implicature is assumed - following Grice (1975) - to be a binary and categorical phenomenon, in the sense that a given utterance is assumed to categorically either give rise to an implicature or not, depending on contextual, cognitive, and linguistic factors. To experimentalists operating on this assumption, a participant's judgment on a particular trial in a TVJT reflects whether or not a scalar implicature was computed in context.

For example, a 'wrong' judgment of the sentence *John ate pizza or a sandwich*, in a context in which the stronger utterance alternative *John ate pizza and a sandwich* is true and equally relevant, is typically interpreted as a "pragmatic" judgment: participants must have recognized that in such a context, the *or*-sentence is true yet underinformative. Pragmatically enriching it to *John didn't eat both pizza and a sandwich* via scalar inference makes it contextually false. Conversely, an answer of 'right' on this view reflects a "literal" semantic interpretation whereby the implicature is not computed (i.e. *John ate pizza or a sandwich - and possibly both*).

This linking assumption underpins the vast majority of TVJT literature relating to scalar inference (Noveck, 2001; Papafragou and Musolino, 2003; Geurts and Pouscoulous, 2009; Doran et al., 2012; Potts et al., 2015). In an early example, Papafragou and Musolino (2003) observe that children accept true but underinformative sentences in a TVJT at a relatively high rate relative to adults, and that this rate is modulated by the particular linguistic scale invoked on a given trial of the experiment (i.e. *some/all* vs. *finish/start* vs. cardinal numbers). The authors argue from this result that scalar implicature computation is dependent upon linguistic scale as well as on a child's recognition of the communicative goals of her interlocutor.

Though widely employed, this linking assumption for TVJTs is associated with a host of problems discussed by Jasbi et al. (2019). Following those authors as well as Tanenhaus (2004), we take these open problems to be indicative of a larger issue in linguistics, namely that the linking hypotheses which bridge linguistic theory and experimentally-elicited behavior are often underdeveloped, underspecified, or (in some cases) absent in experimental studies. In the service of providing a proof of concept for how this issue may be addressed by future researchers, we propose and evaluate a novel account of participant response in TVJT paradigms based on an explicit and quantitatively specified linking function rooted in a probabilistic theory of pragmatic competence. The general idea is that participants' responses in TVJT experiments are related to the probability with which a cooperative pragmatic speaker would have produced the observed utterance (e.g., *John ate pizza or a sandwich*) in order to communicate the meaning presented to participants as fact (e.g., that John ate both pizza and a sandwich). This probabilistic *production* based view departs substantially from the previous widespread assumption that truth-value judgments are a measure of *interpretation*.

Before turning to the specifics of the account, we briefly review some of the open problems in the TVJT literature that motivate the re-thinking of linking functions in TVJT paradigms:

Intermediate judgments: When provided more than two response options in a TVJT, a sizable proportion of participants rates underinformative sentences using the intermediate response options - for example, as only 'kind of [right / wrong]', or 'neither [right nor wrong]'. Katsos and Bishop (2011), for example, provided participants with three response options and observed substantial selection of the intermediate option. They interpreted the choice of this intermediate option as being the result of the computation of an implicature, but a priori, there is no reason to favor this linking assumption over one whereby the intermediate response is associated with a literal semantic interpretation. More generally, it is not clear how the outputs of a binary model of scalar implicature (i.e. implicature or ¬implicature) should relate to non-binary responses on TVJTs.

Population-level variation: In order to explain behavioral variation in contexts where one expects a scalar inference, an adherent to the categorical view of scalar implicature must stipulate that a) not all participants calculated the implicature; or b) some participants who calculated the implicature showed divergent behavior due to some independent mechanism which masked the 'correct' implicature behavior; or some combination of (a) and (b). However, and despite the prevalence of variation at the population level in reported TVJT experiments, even a qualitative analysis of this kind of variation is largely absent from the experimental scalar implicature literature.

Scalar diversity: Doran et al. (2012), following Papafragou and Musolino (2003) inter alia, report that judgments of true but underinformative sentences vary according to the particular linguistic material contained within the sentence, in particular the relevant linguistic scale. They conclude that variation among scalar implicatures is a function of the scale itself (see also van Tiel et al. 2014 for further support for scale-based scalar diversity in a non-TVJT paradigm).

Whether this variation is truly due to inherent features of the linguistic scale (or, e.g., prior world knowledge, or other linguistic material, or other confounding features of the experimental context) is an open question which warrants investigation beyond the scope of this paper. Below, we analyze data from a TVJT where different rates of exhaustive interpretation were observed between a putative lexical scale (<*and, or*>) and a putative ad-hoc, context-dependent pragmatic scale. Our analysis of the data suggests that in this instance, (at least some) variation at the level of linguistic scale may be reduced to more general aspects of pragmatic competence.

Endorsement of false utterances: Invariably, a proportion of participants in TVJTs accepts strictly false sentences. For example, in the study we analyze below, a substantial number of participants rated conjunctions $A \wedge B$ as partially correct in contexts where only A was true. The most common approaches to this type of data are either to use it as the basis of an exclusion criterion or to simply consider it meaningless noise. Doran et al. (2012), for example, exclude participants whose performance deviates by more than two standard deviations from the mean response on 'control' sentences whose semantic contents are consistent with the context of evaluation (and which do not admit of potentially contradictory pragmatic enrichments) or whose semantic contents contradict the context. Katsos and Bishop (2011) report that 2.5% of false sentences in their experiment were endorsed by child participants. On the standard linking assumption, these data are difficult to make sense of, but we will show that they are within the scope of a satisfactory analysis of TVJT behavior.

The remainder of the paper is structured as follows: in Section 2, we summarize the results

Condition	Response Options
Binary	'Right', 'Wrong'
Ternary	'Right', 'Neither', 'Wrong'
Quaternary	'Right', 'Kinda Right', 'Kinda Wrong','Wrong'
Quinary	'Right', 'Kinda Right', 'Neither', 'Kinda Wrong', 'Wrong'

Table 1: Response-option conditions of Jasbi et al. (2019)'s TVJT study.

of a recently reported TVJT study that exemplifies the features discussed above: intermediate judgments, population-level variation, scalar diversity, and participant endorsement of false utterances. Section 3 presents our novel quantitative model of the data from that study. Building on insights from the Bayesian probabilistic literature on pragmatic competence (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013), we model participants as making judgments about a soft-optimal pragmatic speaker whose production choices are a function of utterances' contextual informativeness. On our analysis, participants furthermore expect that the speaker sometimes produce strictly false utterances that are nonetheless somewhat contextually useful. We show that this analysis provides broader empirical coverage over the traditional assumptions discussed above.[1]

2 TVJT Data

2.1 Experiment Materials, Design and Procedure

Jasbi et al. (2019) report the results of a TVJT designed to test whether linking hypothesis and number of response options modulate the researcher's inferences about scalar implicature rates. In their study, number of response options varied between two and five as a between-subjects manipulation. Conditions are summarized in Figure 1. Participants (n = 200) were first shown six cards (Table 2) featuring one or two of the following animals: a cat, a dog, and an elephant. On every trial, participants saw one of the six cards, and a blindfolded cartoon character Bob made guesses as to what animals were on the card. Participants were asked to rate Bob's guesses using the response options available in their particular condition.

Bob made the following guess types: simple declaratives (e.g., *There is a cat*), conjunctions (e.g., *There is a cat and a dog*), and disjunctions (e.g., *There is a cat or a dog*). Card types were crossed with guess types in this study such that a card containing an animal X could be presented with a guess of *There is an* X, *There is an* X *or a* Y (where Y is some animal distinct from X), *There is an* X *and a* Y, or *There is a* Y; cards containing two animals X and Y could be presented with a guess of *There is an* X, *There is an* X *or a* Y, *There is an* X *and a* Y, or *There is a* Z (where Z is some animal distinct from X and Y).

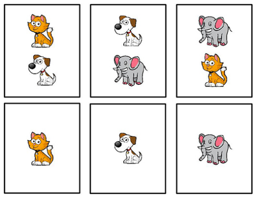

Table 2: Cards used in Jasbi et al. (2019)'s TVJT.

The researchers elicited 3 judgments per participant for each combination of card and guess type.

2.2 Results and Discussion

Proportions of responses for each card-guess type in each response-option condition are shown in Figure 1, with rows presenting behavior aggregated across one and two-card conditions.

The results of the study illustrate the several open empirical issues associated with TVJTs more generally. First, participants routinely reported **intermediate judgments** between 'Right' and 'Wrong' in those conditions where intermediate response options were available. In the Quaternary and Quinary response-option conditions, for example, the intermediate judgment of 'Kinda Right' was the single most-selected response option in two-animal card conditions where Bob's guess was true but underinformative (i.e. either a simple delcarative or a disjunction).

The results also exemplify the issue of **population-level variation**: for example, although behavioral patterns are otherwise fairly categorical in the Binary condition, participant judgments were roughly split between 'Right' and 'Wrong' for underinformative uses of disjunction on two-animal card conditions. A visual inspection of the results suggests even more variation in the population as number of response options increase. The authors furthermore reported **individual-level variation**: qualitatively similar trials (e.g. two trials involving underinformative disjunction) sometimes received different re-

[1] Data and code for all analyses and graphs are available at http://github.com/bwaldon/tvjt_linking.

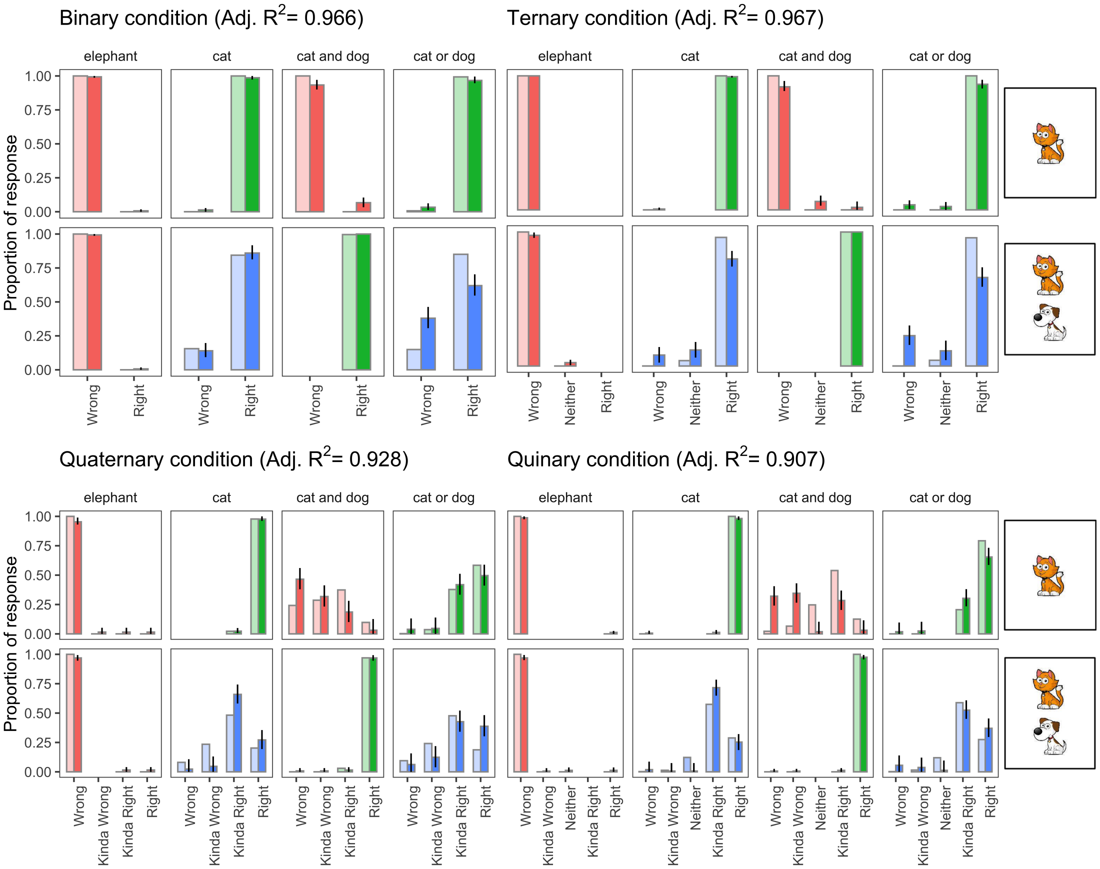

Figure 1: Model predictions (light bars) plotted against empirical results (dark bars) from Jasbi et al.'s (2019) TVJT study. Error bars indicate 95% multinomial confidence intervals. Red and green bars indicate false and true trials, respectively; blue bars indicate implicature trials.

sponses from the same participant.

Comparison of judgments of true but underinformative simple declaratives (i.e. *There is an X*) to judgments of true but underinformative disjunctions (i.e. *There is an X or a Y*) on two-animal card conditions revealed some amount of **scalar diversity**. Following Horn (1972), exposure to the disjunctive connective *or* canonically activates an informationally-stronger scalemate *and* as a pragmatic alternative to give rise to an exclusive interpretation. In contrast, the pragmatic scale in the case of the simple declarative is constructed in a more context-dependent manner. To illustrate, in a two-animal card context where the card features both a cat and a dog, the listener considers a partially-ordered pragmatic scale of *cat and dog*, *cat*, and *dog*, where the conjunction outranks its scalemates in terms of informational strength. Thus, an utterance of *cat* activates *cat and dog* as an alternative to give rise to an the exhaustive interpretation (*There is only a cat on the card*).

In the Binary and Ternary conditions, underinformative uses of *or* resulted in substantially higher rates of 'Wrong' responses than did underinformative simple declaratives, suggesting that at the population level, *or* was interpreted more exhaustively than the simple declarative. However, this pattern was reversed in the Quaternary and Quinary conditions, in which underinformative simple declaratives were more likely to be considered only 'Kinda Right' and less likely to be considered simply 'Right' compared to underinformative disjunctions. This pattern suggests that in the Quaternary and Quinary conditions, simple declaratives were interpreted more exhaustively than disjunctions.

Finally, the data in the Quaternary and Quinary conditions also reveal substantial **participant endorsement of false utterances**. Note specifically that in one-animal card trials, the conjunctive guess (e.g. *cat and dog*) is strictly false; thus, we might naïvely expect a priori that participants categorically judge these utterances to be 'Wrong' in all conditions. Yet when given the option to rate this sentence ‘Kinda Right’ or ‘Kinda Wrong’, participants often did so. In all other conditions where the utterance was strictly false (e.g. a guess of *elephant* for a card containing a cat or a cat and dog), behavior was effectively categorical. That is, rates of endorsement of false utterances varied according to the particular way in which the sentence was false in context.

In sum, the data collected by Jasbi et al. (2019) reflect a range of behavioral patterns unaccounted for by the traditional categorical view of scalar inference and corresponding standard linking assumptions. Below, we report an analysis of their data that aims to predict these phenomena.

3 Analysis

3.1 Cognitive model

Our analysis implements a proposal outlined by Jasbi et al. (2019), couched in the Rational Speech Act (RSA) framework (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013). RSA provides a Bayesian, probabilistic account of pragmatic competence. In RSA, the pragmatic inferences drawn by listeners are represented as probability distributions over meanings which the speaker plausibly intended to convey with a given observed utterance. The probability of this listener (L_1) attributing an intended meaning m to a speaker who produces an utterance u is calculated from a prior probability distribution over potential world states P_w as well as from L_1's expectations about the linguistic behavior of the speaker S_1.

$$P_{L_1}(m|u) \propto P_{S_1}(u|m) \cdot P_w(m)$$

P_{S_1} is modeled as a probability distribution over possible utterances given the speaker's communicative intentions m. This speaker produces utterances that soft-maximize utility, where utility is defined via a tradeoff between an utterance's cost C and its contextual informativeness, calculated from the representation of a literal listener L_0 whose interpretation of an utterance u is in turn a function of the truth conditional meaning $[[u]](m)$ and of her prior expectations $P_w(m)$ regarding the likelihood of possible world states. The extent to which the speaker maximizes utility is modulated by a parameter α – the greater α, the more the speaker produces utterances that maximize utility.

$$P_{S_1}(u|m) \propto e^{\alpha(\ln L_0(m|u) - C(u))}$$
$$P_{L_0}(m|u) \propto [[u]](m) \cdot P_w(m)$$

In RSA (and contra the traditional view), pragmatic inferences are not categorical computations of enriched meanings over the semantic denotations of utterances. For example, exclusive interpretations of *or* are represented in RSA as a positive shift in the posterior probability of an exclusive meaning, relative to its prior probability.

In other words, ‘implicature’ is not a theoretical construct in the RSA framework, absent additional stipulations regarding how to go from probability distributions to binary, categorical inferences. This is an advantage: providing a probabilistic representation of both the speaker's utterance choices and the listener's resulting posterior beliefs after observing an utterance puts us one step closer to accounting for the quantitative behavioral patterns observed in tasks such as TVJTs.

3.2 Behavioral model

Jasbi et al. (2019) proposed but did not systematically test a simple linking hypothesis: rather than providing one response if an implicature is computed and another if it isn't, a participant in a TVJT experiment provides a particular response to an utterance u if the probability of u given a meaning represented by m lies within a particular probability interval on the distribution $P_{S_1}(u|m)$.[2] The participant is modeled as a responder R, who in a binary forced-choice task between ‘Right’ and ‘Wrong’ responds ‘Right’ to an utterance u in world m just in case $P_{S_1}(u|m)$ meets or exceeds some probability threshold θ:

$$R(u, m, \theta) = \begin{cases} \text{‘Right’} & \text{iff } P_{S_1}(u|m) \geq \theta \\ \text{‘Wrong’} & \text{otherwise} \end{cases}$$

The model is extended straightforwardly to an experiment in which participants have a third response option (e.g. ‘Neither’), as in the Ternary condition. In this case, the model specifies two probability thresholds: θ_1, the minimum standard for an utterance in a given world state to count as ‘Right’, and θ_2, the minimum standard for ‘Neither’. Thus, in the Ternary condition:

$$R(u, m, \theta) = \begin{cases} \text{‘Right’} & \text{iff } P_{S_1}(u|m) \geq \theta_1 \\ \text{‘Neither’} & \text{iff } \theta_1 > P_{S_1}(u|m) \geq \theta_2 \\ \text{‘Wrong’} & \text{otherwise} \end{cases}$$

Applying a similar logic allows for the specification of linking hypotheses for TVJTs with an

[2] Following Degen and Goodman (2014), the authors argue that conceptually, behavior on TVJTs is better modeled as a function of an agent's representation of a pragmatic speaker rather than of a pragmatic listener.

arbitrary number of response options.

The intuition behind the threshold model is as follows: participants should disprefer utterances that are relatively unexpected. Thus, high S_1 production probability for a given utterance in context makes it more likely that the utterance receives a positive evaluation in the TVJT – expressed by ordered response options above 'Wrong'. Conversely, the more unexpected an utterance is, the more likely it is to be judged as 'Wrong'. Underinformative utterances of the sort that have traditionally been used to assess 'implicature rates' are precisely the kinds of utterances that are unexpected from informative speakers and are therefore likely to be rated as 'Wrong'.

Here, we assess the quality of this linking hypothesis on the dataset from Jasbi et al. (2019). To that end, we first specify the space of possible meanings and utterances that inform a participant's pragmatic competence in this task. We assume that participants have uniform prior expectations of seeing any of the six possible cards in the experiment. We further assume that participants have uniform prior expectations of a speaker producing any of the four utterance types with which a card may have been crossed. For example, if the card featured either just a cat or both a cat and a dog, we represent the participant as having uniform prior expectations of a speaker producing the guesses *elephant*, *cat*, *dog*, *cat and dog*, or *cat or dog* (that is, we do not posit a cost asymmetry between possible utterances).[3]

For illustrative purposes, the 'Simple Bayesian' bars in Figure 2 display marginal distributions over possible utterances produced by S_1 given these assumptions for the utterance and meanings priors, as well as an arbitary value of 1 for the optimality parameter α, and given that the speaker intends either to communicate the meaning that (just) a cat is on the card or that both a cat and a dog are. The speaker distributions reveal two conceptual issues for the threshold response model proposed by Jasbi et al (2019).

First, the probability of S_1 producing the strictly false guess of *cat and dog* should be zero if the card contains just a cat. This is because the literal listener probability P_{L_0} of inferring the 'only cat' meaning given *cat and dog* is zero by virtue of the fact that the utterance is strictly false in this world state. Thus, any model of response that is a function of P_{S_1} as specified predicts that participants categorically rate the *cat and dog* guess as 'Wrong' in this context, contrary to what is observed in the Quaternary and Quinary conditions.

Second, the probability of producing disjunctions is lower than the probability of producing simple declarative guesses in two-animal card contexts. This asymmetry is advantageous in the case of the Binary and Ternary response data: assuming a threshold for 'Right' positioned between $P_{S_1}(\textit{cat or dog}|\text{cat and dog})$ and $P_{S_1}(\textit{cat}|\text{cat and dog})$, we predict correctly that underinformative simple declaratives should be judged 'Right' more often than underinformative disjunctions. But the asymmetry in S_1 probabilities therefore predicts the wrong pattern of responses on corresponding trials in the Quaternary and Quinary conditions.

We argue that these two seemingly disparate issues can be mediated by a common solution. In particular, we propose a revision to the simple Bayesian inference story above, whereby pragmatically-competent listeners either expect speaker productions as directly sampled from the P_{S_1} distribution, <u>or</u> that those utterance production probabilities inform a second conditional probability distribution of utterances given utterances, the 'Partial Truth' utterance distribution $P_{S_{PT}}$:

$$P_{S_{PT}}(u'|u) \propto \sum_{m \in [\![u]\!]} P_{S_1}(u'|m)^{4}$$

The 'Partial Truth' distribution is a generalized way of modeling a speaker who makes assertions that are sometimes strictly false in light of her intended meaning. Recall that the semantic content of any possible utterance choice made by S_1 is a set of possible worlds and is therefore consistent with meanings unintended by the speaker. S_{PT} models the speaker's soft-optimal production probabilities given these unintended meanings, renormalizing the pragmatic speaker's production probabilities over all possible worlds consistent with utterance choices sampled from P_{S_1}.

[3] We include *dog* as a possible guess because we posit that participants have no reason a priori to expect the other true and underinformative simple declarative - *cat* - over this equally informative guess in two-animal card conditions.

[4] For our implementation of S_{PT}, we restrict the distribution such that u' must entail (or be entailed by) u in order to have probability above 0. Without this restriction, S_{PT} could in principle assign high probability to utterances which have no relevance to the question under discussion (i.e. "What animals are on the card?"), by virtue of those utterances' assertability in worlds consistent with u. A systematic exploration of the linguistic alternatives available to S_1 (as well as S_{PT}) is a question we must leave to future work.

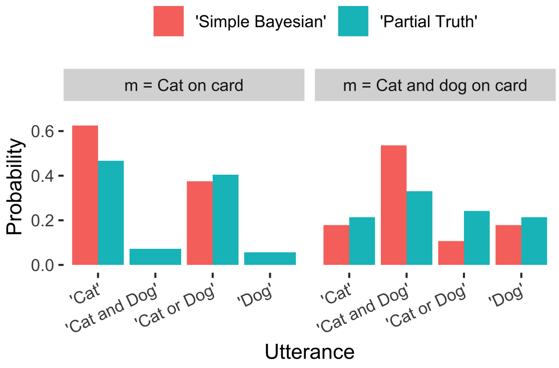

Figure 2: Simulated S_1 production probabilities.

To illustrate: suppose a speaker intends to communicate that many (but not all) of the X are Y, and has quantifier choices *many* and *all*. The only possible utterance choice for the simple Bayesian S_1 speaker is *many*, which is semantically consistent with the intended meaning. But the lower-bounded quantifier *many* is also semantically consistent with an 'all of the X are Y' meaning, which in turn is consistent with the utterance choice *all*. By S_{PT}, we have some nonzero expectation that the speaker will use *all* to communicate the 'many (but not all) of the X are Y' meaning.[5] Thus, a pragmatic listener who hears *all* from the 'Partial Truth' speaker will have a nonzero expectation that *all* should receive an imprecise, non-maximal interpretation. In other words, S_{PT} provides a generalized way of formalizing 'loose-talk' production behavior (Lasersohn, 1999).[6]

The 'Partial Truth' bars in Figure 2 visualize marginal distributions over utterances given an arbitrary 0.6 probability that the speaker samples from the $P_{S_{PT}}$ distribution after sampling from P_{S_1}. The 'Partial Truth' speaker assigns nonzero probability to a guess of *cat and dog* even when the speaker's intended meaning is the single-animal cat card, largely due to the fact that the optimal guess in this context (*cat*) is truth-conditionally consistent with a two-animal card that makes *cat and dog* both true and pragmatically optimal.[7] Moreover, this speaker assigns greater probability to a guess of *cat or dog* in two-animal contexts and down-weights the probability of producing simply *cat*: the optimal utterance in this context (*cat and dog*) is consistent with several world states in which the disjunction *cat or dog* is assertable and with relativley fewer worlds in which *cat* is assertable.

3.3 Quantitative model evaluation

We now turn to a quantitative assessment of the threshold model of response, having addressed two ways in which the unenriched S_1 representation would fail to qualitatively capture behavioral patterns in Jasbi et al (2019)'s TVJT study. Additionally, following Jasbi et al., we recognize that if threshold values were made to be completely invariant across trials of the experiment, then the model would make the undesirable prediction that every participant should have exactly the same response in a given trial type. To allow for population-level variation, the model responder makes a response by comparing the speaker probability against thresholds that are generated from sampling from Gaussian distributions. We thus allow for both population-level and individual level-variation, on the assumption that this sampling procedure takes place whenever a participant is asked to evaluate an utterance in the TVJT.[8]

In order to evaluate the RSA-based threshold model, we conducted a Bayesian data analysis. This allowed us to simultaneously generate model predictions and infer likely parameter values, by conditioning on the TVJT data from Jasbi et al. (separately for each of the four response-option conditions of the experiment) and integrating over the free parameters. Each model assumes uniform priors over utterances and world states as above. We infer the Gaussian threshold distribution parameters and alpha optimality parameters from uniform priors over parameter values using MCMC sampling (observing - for every sample of possible parameter values the expected proportion of responses in that trial type and comparing that distribution to the empirically-observed pattern of response).[9] Additionally, for the Quater-

[5] The effect of this is similar to the use of QUD projection functions for hyperbolic interpretations (Kao et al., 2014).

[6] Formalizing this production behavior is different from analyzing *why* imprecision exists (indeed, is pervasive) in linguistic communication. For the time being, we present this 'loose-talk' speaker model without a thorough assessment of its explanatory power.

[7] Because *cat or dog* is a possible S_1 production, and this choice lies in an entailment relation with the simple declarative guess *dog*, we also assign some probability to *dog* as a guess in this context - albeit lower probability than is assigned to the conjunctive guess *cat and dog*.

[8] We also introduce a random noise term in the parameter estimation such that the simulated responder makes random guesses on 1% of trials. This noise term is removed when running the model forward to make predictive estimations.

[9] We used WebPPL (Goodman and Stuhlmüller, 2014) for

Binary condition

α	σ	μ_{θ_1}
1.22	0.125	0.073

Ternary condition

α	σ	μ_{θ_1}	μ_{θ_2}
1.38	0.076	0.061	0.011

Quaternary condition

α	σ	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	PT
2.75	0.159	0.277	0.101	0.048	0.797

Quinary condition

α	σ	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	μ_{θ_4}	PT
4.38	0.099	0.184	0.042	0.005	0.002	0.437

Table 3: MAP estimates obtained from Bayesian data analysis, where α is the optimality parameter, σ and μ are Gaussian threshold distribution parameters, and PT is the probability with which the speaker samples from $P_{S_{PT}}$ rather than directly from P_{S_1}.

nary and Quinary conditions, we infer from a uniform prior the probability with which the speaker samples from $P_{S_{PT}}$ after sampling from P_{S_1}. The intuition for restricting the 'Partial Truth' manipulation to these conditions is that the behavioral patterns which this manipulation is intended to cover are only observed in these conditions.[10]

Posterior distributions over the parameter values are displayed in Figure 3, and model predictions using maximum a posteriori (MAP) estimates of the parameter values (Table 3) are plotted against Jasbi et al. (2019)'s results in Figure 1. Qualitatively, the model addresses each of the desiderata for an empirically adequate linking function discussed above. In all conditions, the model makes predictions for the full range of response options available to participants – thus addressing the issue of **intermediate judgments**. At the same time, the model addresses the issue of **population-level variation**: sampling threshold values from Gaussian distributions allows different judgments in the population for a given utterance (while keeping the speaker production probability of that utterance constant).

Recall that in the Quaternary and Quinary conditions, there was an asymmetry in the judgment of underinformative disjunctions versus underinformative simple declaratives. The model makes use of the 'Partial Truth' speaker function in order to adjust the underlying speaker production probabilities - and hence the distribution of predicted response options - for these utterances. The 'Partial Truth' function also boosts the production probability of strictly false conjunctions, allowing the model to predict responses other than 'Wrong' for this trial type. Thus, the 'Partial Truth' enrichment helps to address both **scalar diversity** and **endorsement of false utterances**.[11]

The correlation between empirical observations and model predictions is high (Adj. $R^2 > 0.9$ in all conditions), suggesting that the threshold responder model is a good model of TVJT behavior overall. Nevertheless, the model makes some undesirable predictions. For example, it over-predicts rates of 'Neither' responses in the Quinary condition. Empirically, this response tended to be disfavored relative to positive and negative response options, for example in the case of strictly false *cat and dog* guesses. The model assumes that the labeling of the response options should have no particular effect on selection, but future work should engage with this assumption.

4 Discussion and Conclusion

Based on a single underlying probabilistic model of pragmatic competence, the presented threshold responder model provides a level of empirical coverage for TVJT data unavailable to existing linking models rooted in the categorical view of scalar implicature. The contribution of this paper is twofold: methodologically, we present this analysis as a proof-of-concept approach to modeling TVJT data for researchers in experimental semantics/pragmatics. We see the presented behavioral model as a starting point for future quantitative analytic work in the TVJT domain – a model against which future models may be assessed.[12]

On the theoretical side, the cognitive model that forms the basis for the behavioral model is non-neutral in its assumptions. In particular, it assumes that TVJT behavior is the result of reasoning about probabilistic utterance choices that

MCMC inference, with 5000 samples (plus a lag of 10 iterations between samples) and a burn-in time of 20,000 iterations. We computed maximum a posteriori values from the marginal posterior distributions over parameter values using the `density` function in R.

[10] We speculate that there may be a link between increasing the number of response options and participants' increased expectation of Partial Truth speaker behavior, which may have been strengthened by the fact that the Quaternary and Quinary conditions explicitly made reference to gradient levels of correctness (i.e. 'Kinda Right' / 'Kinda Wrong'). But this speculation warrants future investigation.

[11] We leave further investigation of the 'Partial Truth' function - in particular its extension to an analysis of linguistic imprecision as sketched above - to future work.

[12] For example, one could in principle link the threshold model to pragmatic listener probabilities of meanings given utterances rather than to speaker production probabilities given intended meanings (as we do in this paper).

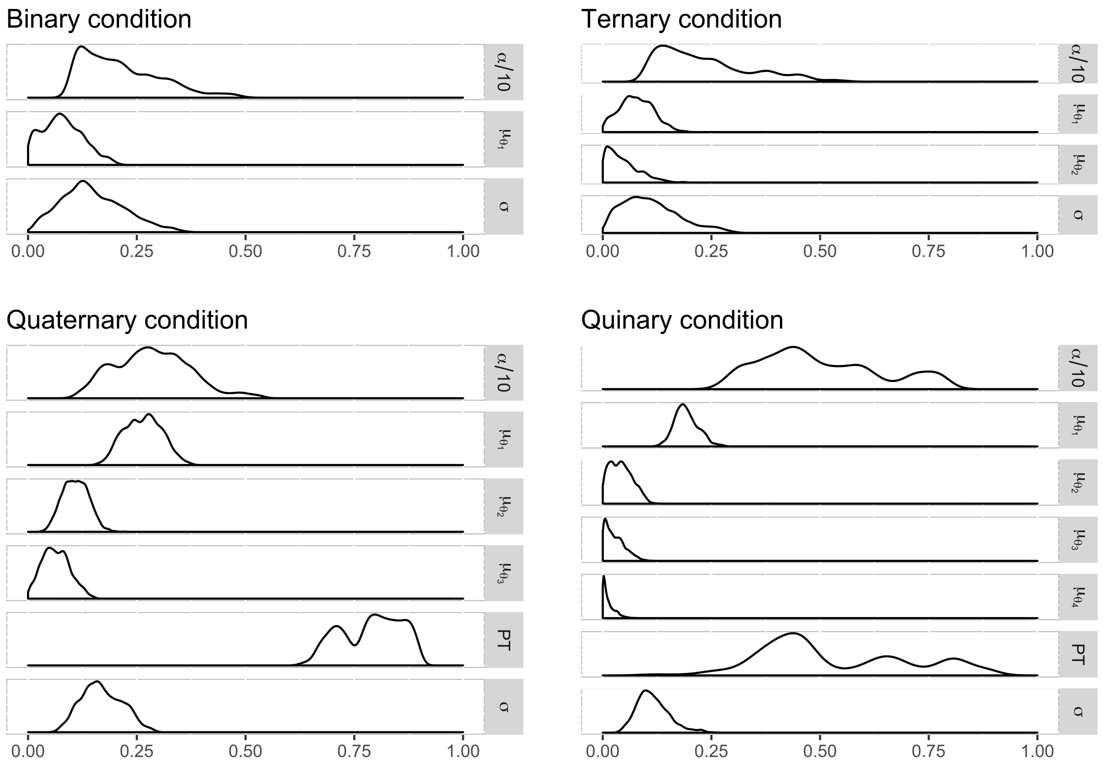

Figure 3: Normalized marginal posterior distributions over parameter values for the threshold responder model in each experimental condition. Note that the posterior distribution for the optimality parameter α has been rescaled for the purposes of this visualization.

are the result of trading off (contextual) utterance informativeness and cost. Under this view, not only does TVJT behavior not quantify implicature rates; the very notion of an implicature evaporates. Rather than finding this undesirable, we believe that this framework allows for more rigorous engagement with the complexities of linking theoretical constructs to behavior (see also Franke 2016), an area of some dearth in experimental semantics/pragmatics.

—

References

Lewis Bott and Ira Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3):437–457.

Emmanuel Chemla and Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics*, 28(3):359 – 400.

Wim De Neys and Walter Schaeken. 2007. When people are more logical under cognitive load - dual task impact on scalar implicature. *Experimental Psychology*, 54(2):128–133.

Judith Degen and Noah D Goodman. 2014. Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 397–402.

Judith Degen and Michael K. Tanenhaus. 2015. Processing scalar implicature A constraint-based approach. *Cognitive Science*, 39(4):667–710.

Ryan Doran, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88:124–154.

Michael C. Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336:998.

Michael Franke. 2016. Task types, link functions & probabilistic modeling in experimental pragmatics. In *Preproceedings of Trends in Experimental Pragmatics*.

Bart Geurts and Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics*, 2:1–34.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–84.

Noah D Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2019-8-8.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Laurence Horn. 1972. *On the Semantic Properties of the Logical Operators in English*. Ph.D. thesis, UCLA.

Masoud Jasbi, Brandon Waldon, and Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology*, 10:189.

Justine Kao, J. Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, 111(33):12002–12007.

Napoleon Katsos and Dorothy V M Bishop. 2011. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, 120(1):67–81.

Peter Lasersohn. 1999. Pragmatic halos. *Language*, pages 522–551.

Ira Noveck. 2001. When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.

Ira Noveck and Andres Posada. 2003. Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language*, 85(2):203–210.

Anna Papafragou and Julien Musolino. 2003. Scalar implicatures: experiments at the semanticspragmatics interface. *Cognition*, 86:253–282.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33(1975):755–802.

Michael K. Tanenhaus. 2004. On-line sentence processing: past, present and future. In Manuel Carreiras and Charles Clifton, editors, *On-line sentence processing: ERPS, eye movements and beyond*, pages 371–392. Psychology Press, London, UK.

Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2014. Scalar diversity. *Journal of Semantics*.

—