Prior probability predicts projection

Author(s)

January 5, 2021

Abstract

Beliefs about the world, often termed "world knowledge", affect language processing and utterance interpretation in a variety of empirical domains. In three experiments, we tested whether subjective beliefs about the probability of certain world states obtaining**JT: do we need this word? it garden-pathed me** influence *projection*, the phenomenon whereby listeners infer that speakers are committed to non-entailed utterance content. We find that the higher the prior belief in a world state, the more projective it is, i.e., the more speakers are taken to be committed to it. Prior beliefs predict projection both at the group and at the by-participant level. This work highlights the importance of combining semantic theories of projection with cognitive theories of language understanding. **JT: (4,098 words)**

1 Introduction

Decades of psycholinguistic research have documented**JT: "Psycholinguistic work has documented" is less hyperbolic** the wide variety of ways in which probabilistic beliefs about the world, often termed "world knowledge", affect language processing (e.g., Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Hagoort, Hald, Mastiaansen, & Petersson, 2004; Hald, Steenbeek-Planting, & Hagoort, 2007; Warren & McConnell, 2007), including syntactic ambiguity resolution (e.g., Chambers, Tanenhaus, & Magnuson, 2004; Bicknell & Rohde, 2014), reference resolution (e.g., Winograd, 1972; Hanna & Tanenhaus, 2004), genericity (e.g., Tessler & Goodman, 2019)), scalar implicature (e.g., Degen, Tessier, & Goodman, 2015)), underinformativity implicatures (Kravtchenko & Demberg, 2015), and the production of redundant referring expressions (Mitchell, Reiter, & Van Deemter, 2013; Westerbeek, Koolen, & Maes, 2015; Rubio-Fernández, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020). In contrast, research in formal semantics, the field devoted to specifying how the meaning of a sentence results compositionally from the meanings of its parts, has traditionally sidelined the importance of world knowledge[1] [jd: what are some examples of papers that explicitly say this isn't the domain of semantics?].**JT: are we not setting up a strawman? pragmatics does consider WK part of its domain. and projection hasn't just been looked at in semantics proper** In this paper, we provide empirical evidence from American English that a key focus of formal semantic research, *projection*, is systematically affected by listeners'[2] subjective beliefs about the world, suggesting that semantic theorizing should take into account rather than ignore the ways in which linguistic content's relation to the world enters into the computation of meaning. We provide a sketch of such an account at the end of this paper. [jd: make sure you do – link up to prior in RSA]

Projection is the phenomenon whereby listeners infer that the speaker is committed to non-entailed utterance content. For instance, a speaker who says *Sam knows that it's raining* is typically taken to be committed to the truth of the content of the complement (henceforth, CC) of *know*, that it is raining. This

[1]Because "knowledge" implies justified true belief but subjective beliefs need not be accurate to affect language processing in systematic ways, we henceforth avoid the term "world knowledge" and instead refer to "(subjective prior) beliefs about the world."

[2]We include readers, writers, and signers in the terms "listener" and "speaker."

on its own is not surprising, because the CCs of verbs in regular declarative sentences are entailments of the sentence**JT: not of all verbs** [jd: add a textbook ref here for people who really know nothing about sematnics?] However, even when *know* is realized under well-known entailment-canceling operators like polar questions (*Does Sam know that it's raining?*) or negation (*Sam doesn't know that it's raining.*), the inference that the speaker is committed to the CC that it is raining persists. This suggests that the inferred speaker commitment cannot be due to entailment; instead, such inferences are attributed to *projection* of the CC (Langendoen & Savin, 1971; Beaver & Geurts, 2014). [jd: this still doesn't explain what projection is, really. is it ever positively defined, or just negatively as content the speaker is taken to be committed to even though it occurs under entailment-canceling operators?]**JT: speakers may be taken to be committed to the truth of utterance content. for some of those, semantics has an easy explanation: entailment. for others, they don't, because it goes against compositionality: name for that subset of content is projective content**

Recent experimental work has found that projection appears to be gradient rather than an all-or-none phenomenon: listeners' inferences about speaker commitment to utterance content vary in strength and are affected by contextual factors including the discourse status of the CC and the prosody of the utterance (for an overview see Tonhauser, Beaver, & Degen, 2018). The hypothesis that listeners' prior beliefs influence projection was initially put forth by Stevens, de Marneffe, Speer, and Tonhauser (2017) and Tonhauser et al. (2018), who observed by-item projection variability for different CCs of clause-embedding predicates like *know* and *discover*. They argued that one source of the observed variability may be that more a priori likely content (*Kim flew to New York*) projects more strongly than less a priori likely content (*Kim flew to the moon*) when realized as the CC of a clause-embedding predicate (as in *Did John discover that Kim flew to New York/the moon?*). This idea can straightforwardly be made sense of under recent Bayesian accounts that treat pragmatic utterance interpretation as a matter of combining uncertain prior beliefs about the world with uncertain beliefs about likely speaker production choices via Bayes' rule (Goodman & Frank, 2016; Degen et al., 2015): a CC that is more likely a priori (*before* observing an utterance) is also more likely a posteriori (*after* observing an utterance), even if the particular predicate biases the listener away from taking the speaker to be committed to the CC.

Support for the hypothesis that listeners' prior beliefs influence projection comes from Mahler (2020), who investigated the projection of politically charged CCs of English clause-embedding predicates. For example, the politically charged content in (1) is that Obama improved/damaged the American economy. The prior probability of the content was manipulated by the speaker (Cindy in (1)) speaking at the club meeting of either the College Republicans or Democrats.

(1) Cindy, at the College Republicans/Democrats club meeting:
Ben doesn't know that...
  a. ...Obama improved the American economy.
  b. ...Obama damaged the American economy. (Mahler, 2020, 784f.)

Higher prior probability content (e.g., a liberal content like (1a) uttered by a Democrat) was more projective than a lower prior probability content (e.g., a liberal content uttered by a Republican).

In contrast, Lorson (2018) did not find empirical support for the hypothesis that listeners' prior beliefs influence projection in a study of the effect of prior probability on the projection of the pre-state content of the English change of state verb *stop*. Prior probability was manipulated through gender stereotypes reported in Boyce, von der Malsburg, Poppels, and Levy (2018). For instance, because men are more likely than women to be plumbers, the pre-state content of (2a), that James has worked as a plumber, was hypothesized to be more projective than the pre-state content of (2b), that Linda has worked as a plumber.

(2) a. Did James stop working as a plumber?

b. Did Linda stop working as a plumber? (Lorson, 2018, 38)

Several differences between Mahler's and Lorson's investigations could be implicated in the differential support for the hypothesis: a) the projective content investigated (contents of clausal complements vs. pre-state content of *stop*); b) the entailment-canceling environment (negation vs. polar question); c) the manipulation of the prior probability (political party affiliation vs. gender stereotypes); and d) how explicitly the prior-manipulating information was provided to participants (statement of political party affiliation vs. use of a male or female name to indicate gender). The three experiments reported in this paper investigate the hypothesis on the basis of the projection of the contents of complements of clause-embedding predicates that are embedded in polar questions. In contrast to Mahler, 2020, our experiments include 20 clause-embedding predicates (rather than just 7). We manipulated the prior probability of content in a variety of ways. [jd: what does this mean? we just did it one way, via facts, no?]**JT: mahler only manipulated the speakers political affiliation. we manipulated 20 different things about people** The prior-manipulating information was explicitly provided. [jd: is this a contrast to mahler? if so, need to say explicitly. generally, i find it hard at this point to see how what we're doing is qualitatively different from mahler. can we bring that out more?]**JT: no, that is not in contrast to mahler. we differ from mahler in the following 3 ways: she manipulated something about the speaker, we manipulated the CC / she considered factive vs. non-factive verbs, we look at 20 verbs / she investigated the hypothesis at the group level, we do group- and by-participant level**

Both Mahler, 2020 and Lorson, 2018 measured projection using the 'certain that' diagnostic (see, e.g., Tonhauser, 2016; Tonhauser et al., 2018; de Marneffe, Simons, & Tonhauser, 2019): Participants were asked whether the speaker is certain of the relevant content and they responded on a sliding scale, with one end labeled negatively and the other positively. Both investigations assumed that the more positive the response, the more projective the content. Our experiments also used the 'certain that' diagnostic with such a sliding scale. [jd: my impression is that up to this point, everything in this paragraph can just go in the methods section of exp 1 (nad in fact, is already there)]**JT: we usually explain how the certainty diagnostic works, that is not done in section 2, if we can just say that this is a standard diagnostic for projection and say that mahler and lorson also used it, that's cool and will save some space** Exp. 1 (section 2) investigated the effect of prior beliefs on projection by measuring prior probability of contents and projection in a within-participant design. This design allowed us to investigate whether individual participants' beliefs affected their projection ratings. In Exps. 2 (section 3), prior probability and projection were measured in separate groups, as in Mahler, 2020 and Lorson, 2018. In both experiments, more a priori likely contents were more likely to project. We reflect on the implications of this result for theories of meaning and language processing in section 4.

2 Experiment 1

This experiment tested whether higher prior probability content is more likely to project.**JT: more projective? "more likely to project" sounds like projection is binary** Prior probability and projection ratings were collected for the contents of 20 clauses that realized the complements of 20 clause-embedding predicates.[3]

Participants 300 participants with U.S. IP addresses and at least 99% of previous HITs approved were recruited on Amazon's Mechanical Turk platform (ages: 18-82, median: 35.5; 119 female, 179 male, 1

[3]The experiments, data and R code for generating the figures and analyses of the experiments reported on in this paper are available at [redacted for review]. Exp. 1 was pregistered: [link removed for review]. All experiments were conducted with approval from the IRB of [university redacted] and informed consent was obtained.[jd: i think there's a way of anonymizing osf links?]

other, 1 undeclared). They were paid $1.80.

Materials The prior probability and the projection of the contents of 20 clauses were measured in separate blocks. Each clause (e.g., *Julian dances salsa*) was paired with two facts between participants. The content of the clause was expected to have a higher prior probability in the presence of one fact (e.g., *Julian is Cuban*) than of the other (e.g., *Julian is German*). See Supplement A for the full set of clauses and facts.

[jd: turn (3), (4), (5), and (6) into an image that provides good visual overview of the 2 tasks and the predicates – probably just merge into figure of the tasks below. this will also save words.]**JT: not sure i understand: is figure 1 what you're looking for, or something else?**

In the prior block, the 20 clauses were realized as the complements of *How likely is it that. . . ?* questions. As shown in (3)[jd: change to figure ref], each target stimulus consisted of one of the two facts for that clause and the *How likely is it that. . . ?* question. There were a total of 40 target stimuli in the prior block.[jd: wait, is this true? we elicited high and low prior ratings from each participant?]**JT: no. each participant rated all 20 contents, 10 with high and 10 with low. i was trying to describe how many stimuli there are in this block, not what each participant saw. does one not do it like this?**

(3) a. **Fact:** Julian is Cuban.
How likely is it that Julian dances salsa?

b. **Fact:** Julian is German.
How likely is it that Julian dances salsa?

In the projection block, the target stimuli consisted of a fact and a polar question that was uttered by a named speaker, as shown in (4).[jd: change to figure ref] The polar questions were formed by realizing the 20 clauses as the complements of the 20 clause-embedding predicates in (5).[jd: change to figure ref] There were a total of 800 target stimuli in the projection block.[jd: wait what? that can't be right. didn't each participant see 20 prior and 20 projection trials?]

(4) a. **Fact (which Carol knows):** Julian is Cuban.
Carol: Does Sandra know that Julian dances salsa?

b. **Fact (which Carol knows):** Julian is German.
Carol: Does Sandra know that Julian dances salsa?

(5) *be annoyed, discover, know, reveal, see, acknowledge, admit, announce, confess, confirm, establish, hear, inform, prove, pretend, suggest, say, think, be right, demonstrate*

The 20 predicates in (5)[jd: change to fig ref] include a cross-section of English clause-embedding predicates: They include cognitive predicates (e.g., *know*), emotive predicates (e.g., *be annoyed*), communication predicates (e.g., *announce*), and inferential predicates (e.g., *prove*), as well as so-called factive and non-factive predicates (e.g., *know* vs. *think*)[4]

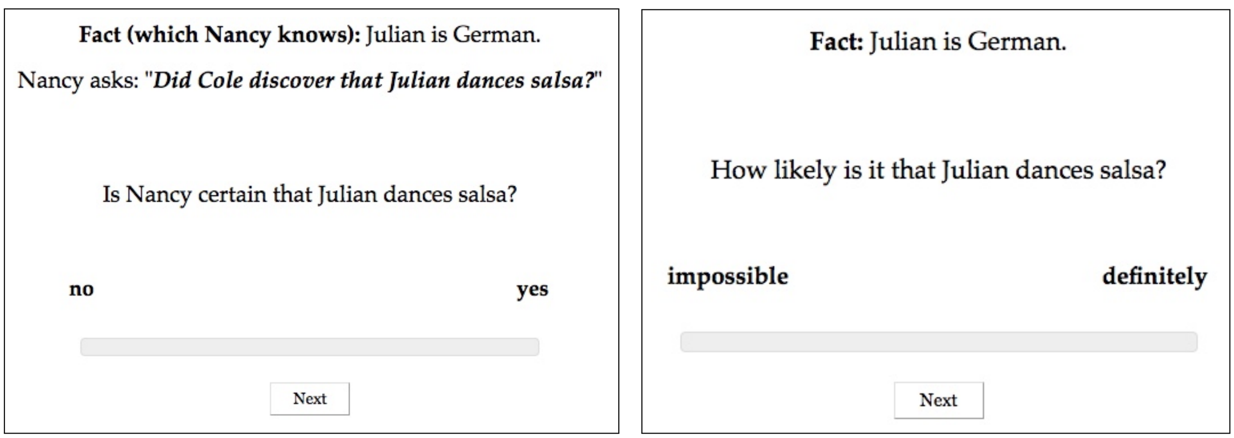
The projection block also included 6 control trials, which functioned as attention checks. The content of these items was expected not to project: For example, in (6)[jd: change to fig ref], the speaker is not committed to the main clause content, that Zack is coming to the meeting tomorrow. The same 6 main clauses were also used to form 6 filler trials in the prior block; these were not used to assess participants' attention. For the full set of items see Supplement A.

[4]Tonhauser and Degen (ms) challenge the distinction between factive and non-factive predicates and we do not analyze the predicates separately, but we have [jd: include some way of marking the different predicate types because i'm not sure it'll get through review otherwise].**JT: factive versus non-factive? will things be different once we have the paper accepted?**

(6) Sample control stimulus in the projection block
Fact (which Margaret knows): Zack is a member of the golf club.
Margaret: Is Zack coming to the meeting tomorrow?

Each participant's stimulus set was semi-randomly generated by first randomly pairing up the 20 predicates and clauses. Half of the items were then randomly assigned the respective clause's higher-probability fact, and half its lower-probability fact. Participants completed a total of 52 trials: 20 target trials in each block, 6 control trials in the projection block and 6 filler trials in the prior block. Block order and within-block trial order were randomized.

Procedure In the projection block, participants were told to imagine that they are at a party and that, on walking into the kitchen, they overhear somebody ask somebody else a question. Participants were asked to rate whether the speaker was certain of the content of the complement, taking into consideration the fact that was presented. They gave their responses on a slider marked 'no' at one end (coded as 0) and 'yes' at the other (coded as 1), as shown in Fig. 1a. In the prior block, participants were told to read facts and to assess the likelihood of events, given those facts. They gave their responses on a slider marked 'impossible' at one end (coded as 0) and 'definitely' at the other (coded as 1), as shown in the sample trial in Fig. 1b



(a) Projection block.

(b) Prior block.

Figure 1: Sample trials in Exp. 1.

After completing the experiment, participants filled out a short optional demographic survey. To encourage truthful responses, participants were told that they would be paid no matter what answers they gave in the survey.

Data exclusion Data was excluded based on self-declared non-native speaker status and other criteria given in Supplement B, leaving 7436 data points from 286 participants to be analyzed (ages 18-82; median: 35.5; 116 female, 186 male, 1 other, 1 undeclared).

2.1 Results

Prior beliefs. Fig. 2 shows the mean prior probabilities of the 20 contents by fact. Each content's mean prior probability was rated as higher when it was presented with its higher probability fact than when it was presented with its lower probability fact ($\beta = 0.45$, $SE = 0.01$, $t = 31.12$, $p < .0001$), as assessed in a mixed-effects linear regression predicting slider rating from dummy-coded fact type (reference level: 'lower probability') and random by-item and by-participant intercepts and slopes for fact type.[5] This suggests that the manipulation of the prior probability of the 20 contents was successful.

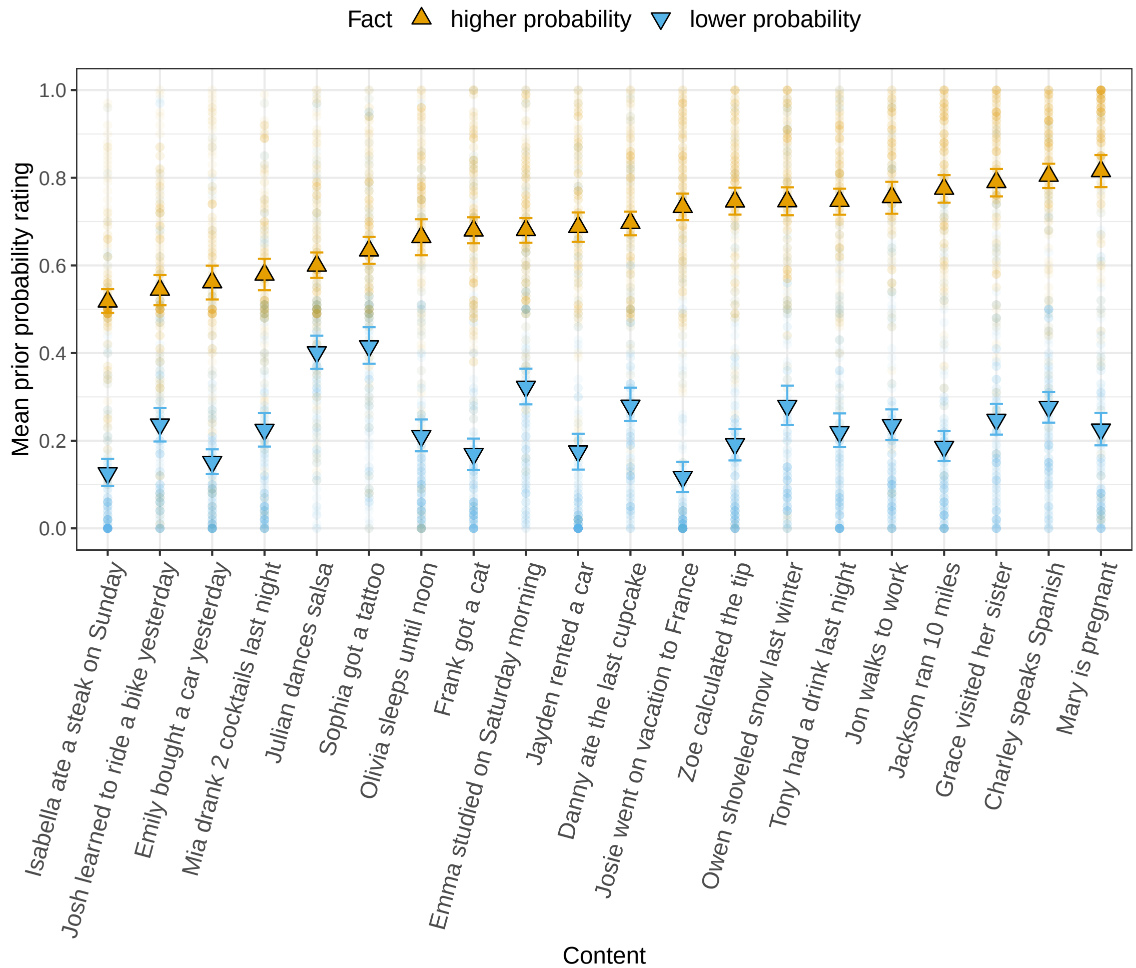


Figure 2: Mean prior probability by content and fact in Exp. 1. Error bars indicate 95% bootstrapped confidence intervals. Transparent dots indicate individual participant ratings.

Do prior beliefs modulate projection? Fig. 3 shows the mean certainty ratings for the CCs by predicate and by fact, as well as the mean certainty rating for the main clause controls (abbreviated 'MC'). The mean certainty ratings were higher for contents presented with higher probability facts than for contents presented with lower probability facts ($\beta = 0.14$, $SE = 0.01$, $t = 12.24$, $p < .0001$), as assessed by a mixed effects linear regression predicting mean certainty rating from dummy-coded fact type (reference level: 'lower probability') and random by-item and by-participant intercepts and slopes for fact type. This result, which holds for all 20 clause-embedding predicates, suggests that participants' prior beliefs about content probability systematically modulated the extent to which they take the speaker to be committed to that content.

We also replicated Tonhauser and Degen (ms)'s result of by-predicate variation in the projection of the CC: for instance, the CC of *be annoyed* was more projective than that of *discover*, which in turn was

[5] All analyses were conducted in R (R Core Team, 2016) using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015).

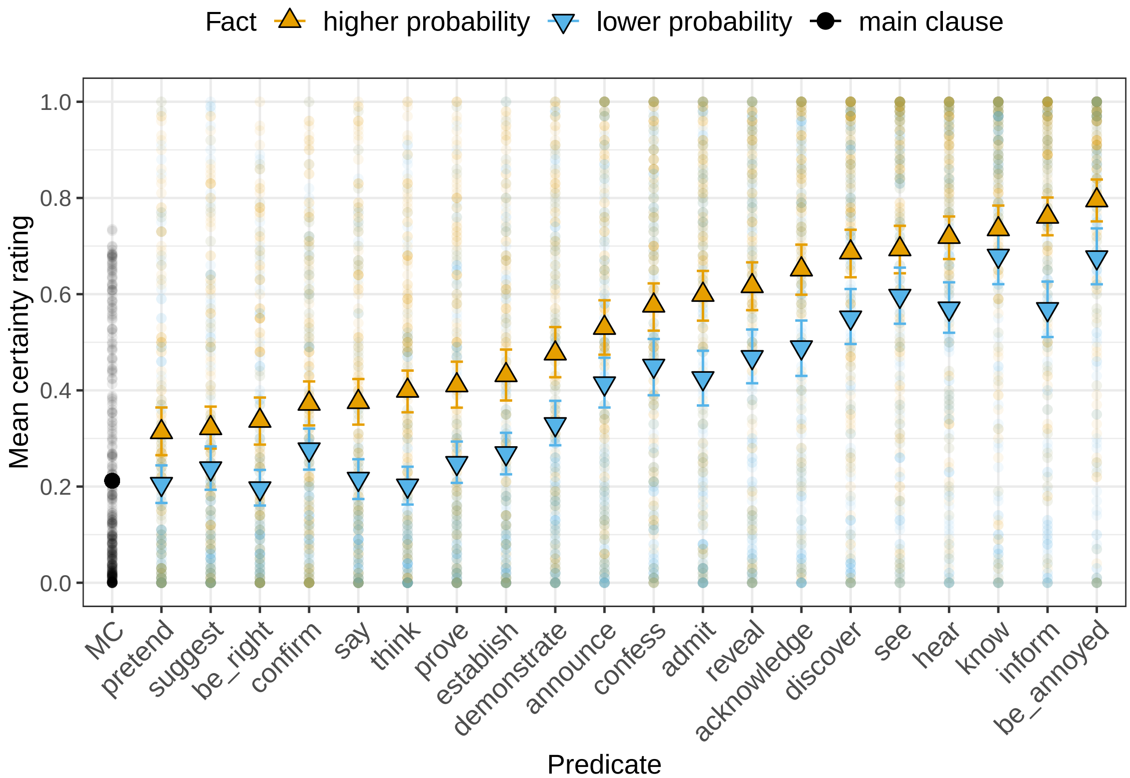


Figure 3: Mean certainty ratings by predicate and prior probability of the content of the complement in Exp. 1. Error bars indicate 95% bootstrapped confidence intervals. Light dots indicate participants' ratings.

more projective than that of *announce*. The Spearman rank correlation between the mean certainty ratings in Exp. 1 (collapsing over facts) and Tonhauser and Degen's (ms) Exp. 1a is .991; see Supplement D for a visualization. Exp. 1 thereby also provides further evidence for the systematic influence of the predicate on projection, but crucially the effect of the prior was observable independently of predicate.

Closer inspection of Fig. 2 reveals by-participant variability in prior probability ratings. This means that individual participants' prior beliefs may not align with the prior probability classification assumed in Fig. 3. For example, given a particular content (e.g., that Julian dances salsa), it is possible that one participant's prior probability rating was lower than that of another participant, even though the first participant was presented with the higher probability fact (Julian is Cuban) and the second one with the lower probability fact (Julian is German). Fig. 4 shows participants' certainty ratings by their individual prior probability ratings. To investigate whether prior beliefs modulate projection at the by-participant level, we conducted the same mixed-effects analysis reported above,**JT: but above we're predicting mean certainty ratings and here were predicting individual certainty ratings, no?** but used participants' individual prior probability ratings as the fixed effect prior predictor. Again, higher-prior-probability CCs were more likely to project (β = 0.28, SE = 0.02, t = 13.85, p < .0001). This suggests that prior probability influences projection even at the by-participant level. In fact, a Bayesian Information Criterion (BIC) model comparison revealed that the individual-level model better captured the variance in the data (group-level model BIC: 2654; participant-level model BIC: 2291), suggesting that individual listeners' prior beliefs systematically modulate the extent to which they take the speaker to be committed to a content: the more they believe it, the more they take the speaker to believe it.

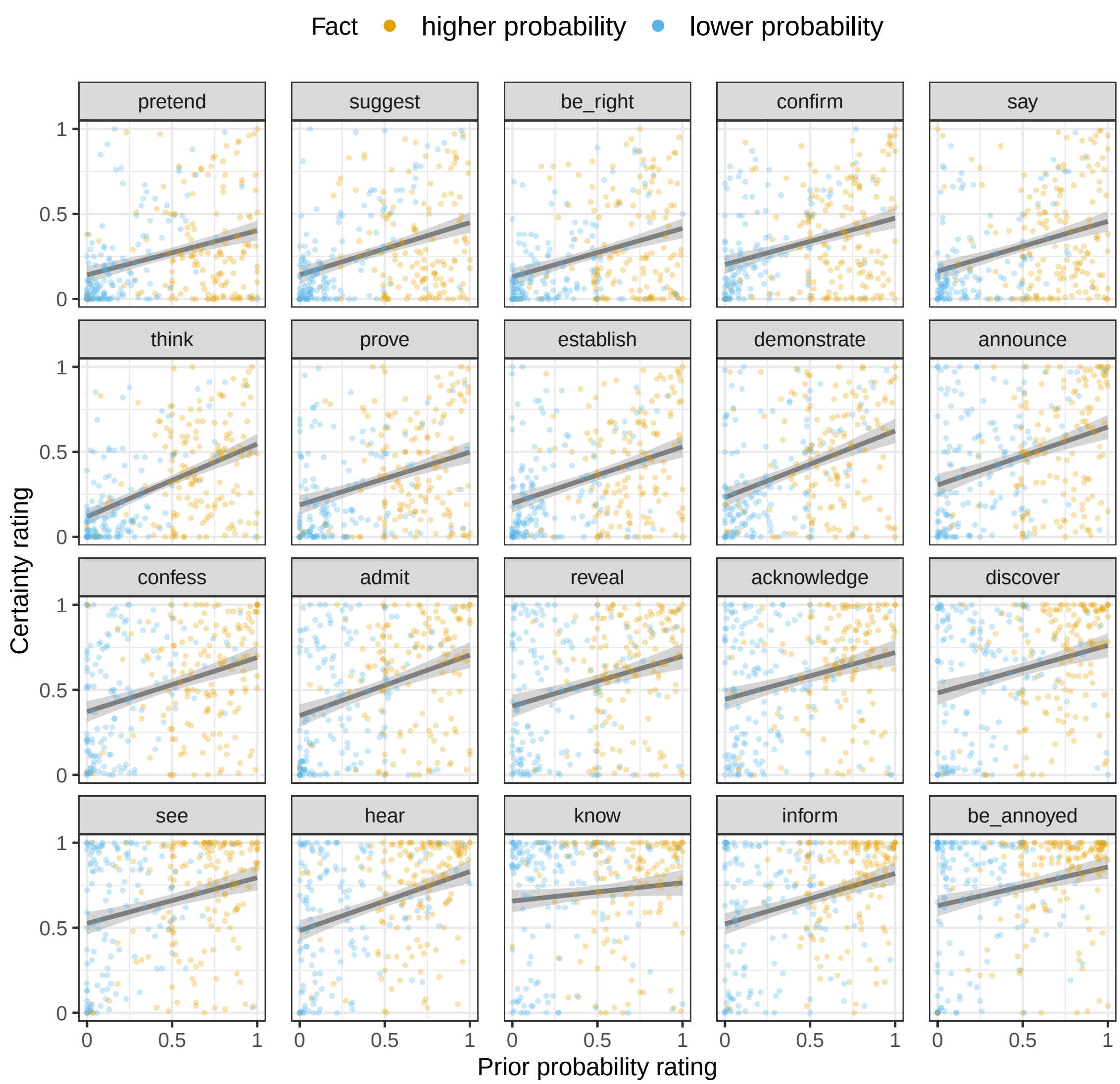


Figure 4: Certainty ratings by prior probability ratings by predicate in Exp. 1. Linear smoothers with 95% confidence intervals are overlaid.

2.2 Discussion

[jd: I wonder whether most of this wouldn't be better served being discussed in the general discussion at the end, and here only briefly bring up the methodological issue that motivates exps 2? (especially given the space constraints?)]**JT: happy to follow your lead.**

The results of Exp. 1 provide empirical support for the hypothesis that higher prior probability content is more likely to project. These results confirm Mahler's (2020) results and expand on them in several ways. First, while Mahler, 2020 manipulated only the political party affiliation of the speaker, the manipulation in Exp. 1 relied on 20 distinct properties of the individuals denoted by the subjects of the 20 clauses (e.g., whether Julian is more likely to dance salsa if he is German or Cuban, or whether Zoe is more likely to have calculated the tip, if she is 5 years old or a math major). Thus, the result of Exp. 1 suggests a general effect of prior probability on projection.[6]

Second, while Mahler, 2020 observed an influence of prior probability of the projection of the CCs of two classes of predicates (so-called factive and non-factive predicates), Exp. 1 observed such an influence for the CCs of 20 predicates. Thus, the result of Exp. 1 supports the assumption that the effect of prior probability on projection is more general than was suggested in Mahler, 2020. In fact, the results of Exp. 1 motivate the hypothesis that prior probability influences projection across the set of English clause-embedding predicates. Finally, prior probability and projection were measured in a within-participant design in Exp. 1, in contrast to Mahler's experiment, which only measured projection. Exp. 1 thus supports the claim that projection is influenced not only by the average prior probability of content but also by the prior probability that individual listeners assign to content. This result suggests that by-participant variability in projection experiments (see, e.g., Tonhauser et al., 2018; Tonhauser & Degen, ms) may be due to participants assigning different prior probabilities to the contents under investigation. It is possible, however, that the within-participant design resulted in participants' responses on either block influencing their responses on the other block. To mitigate against this possibility, we conducted Exps. 2, where prior probability and projection ratings were collected from different populations.

3 Experiments 2

Exps. 2a and 2b measured the prior probability and the projection of the 20 contents of Exp. 1, respectively.

3.1 Methods

Participants Participants with U.S. IP addresses and at least 99% of previous HITs approved were recruited on Amazon's Mechanical Turk platform. The 95 participants in Exp. 2a (ages: 21-75, median: 33; 45 female, 50 male) were paid 55 cents. The 300 participants in Exp. 2b (ages: 21-72, median: 36; 145 female, 154 male, 1 undeclared) were paid 85 cents.[7]

Materials and procedures The 40 target stimuli of Exp. 2a were identical to those of the prior block of Exp. 1. Each participants saw the two control stimuli in (6), which were included to assess attention to the task. We expected high prior probability ratings for (6a) and low ones for (6b).

(7) a. **Fact:** Barry lives in Germany.
How likely is it that Barry lives in Europe?

b. **Fact:** Tammy is a rabbit.
How likely is it that Tammy speaks Italian and Greek?

The materials of Exp. 2b were identical to those of the projection block of Exp. 1. Trial order in both experiments was random. The procedures of Exps. 2a and 2b were identical to those of the prior and projection blocks of Exp. 1, respectively.

Data exclusion We excluded data based on the criteria given in Supplement B, leaving data from 75 participants to be analyzed in Exp. 2a (ages 21-75; median: 35; 34 female, 41 male) and from 266 participants in Exp. 2b (ages 21-72; median: 36; 129 female, 136 male, 1 undeclared).

[6] None of our prior probability manipulations relied on gender stereotypes because we only became aware of Lorson, 2018 after running Exp. 2.

[7] 28 participants took both experiments. Given that the experiments were run two weeks apart (Exp. 2a on November 13, 2017 and Exp. 2b on November 28, 2017, it is unlikely that these 28 participants' prior ratings influenced their projection ratings.

3.2 Results and discussion

Exp. 2a replicated the prior probability manipulation of Exp. 1. Fig. 5 plots the mean prior probability ratings in Exp. 2a against those of Exp. 1. The Spearman rank correlation was high, at .977. For a visualization of the by-content prior ratings see Supplement E.

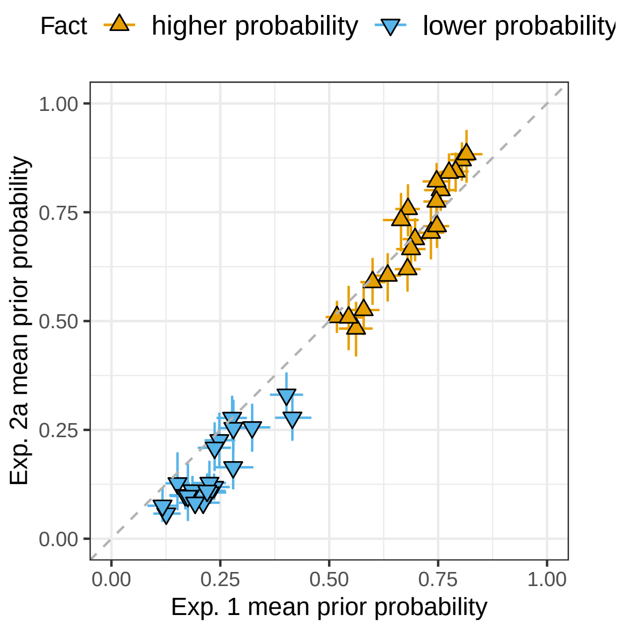


Figure 5: Mean prior probability ratings in Exp. 2a against those of Exp. 1. Error bars indicate 95% bootstrapped confidence intervals. The dotted line indicates **JT: what?**.

Exp. 2b replicated the critical result of Exp. 1, that prior probability influences projection.[8] Fig. 6 shows that mean certainty ratings in Exp. 2b were higher for contents presented with higher probability facts than for contents presented with lower probability facts.

JT: report models predicting a) projection mean Exp 2b from prior mean Exp 2b, b) projection mean Exp 1 from prior mean Exp 2b, and c) projection mean Exp 2b from prior mean Exp 1). USE COHENS D? These results suggest that the result of Exp. 1 is not an artifact of the within-participant design of Exp. 1.

4 Concluding remarks

This paper provided empirical support for Stevens et al.'s (2017) and Tonhauser et al.'s (2018) hypothesis that listeners' beliefs about utterance content influence their inferences about speaker commitment to that content. A pressing question for future research is how prior probabilities interact with other factors that have been shown to influence projection, such as the content's discourse status and utterance prosody. On the theoretical side, our result motivates the development of projection analyses that predict the influence of prior probability on the projection of the content of English clause-embedding predicates. Analyses currently on the market do not predict the results of our experiments because they are limited to subsets

[8]Exp. 2b also replicated Tonhauser and Degen's (ms) result that there is by-predicate variation in the projection of the content of the complement; see Supplement D.

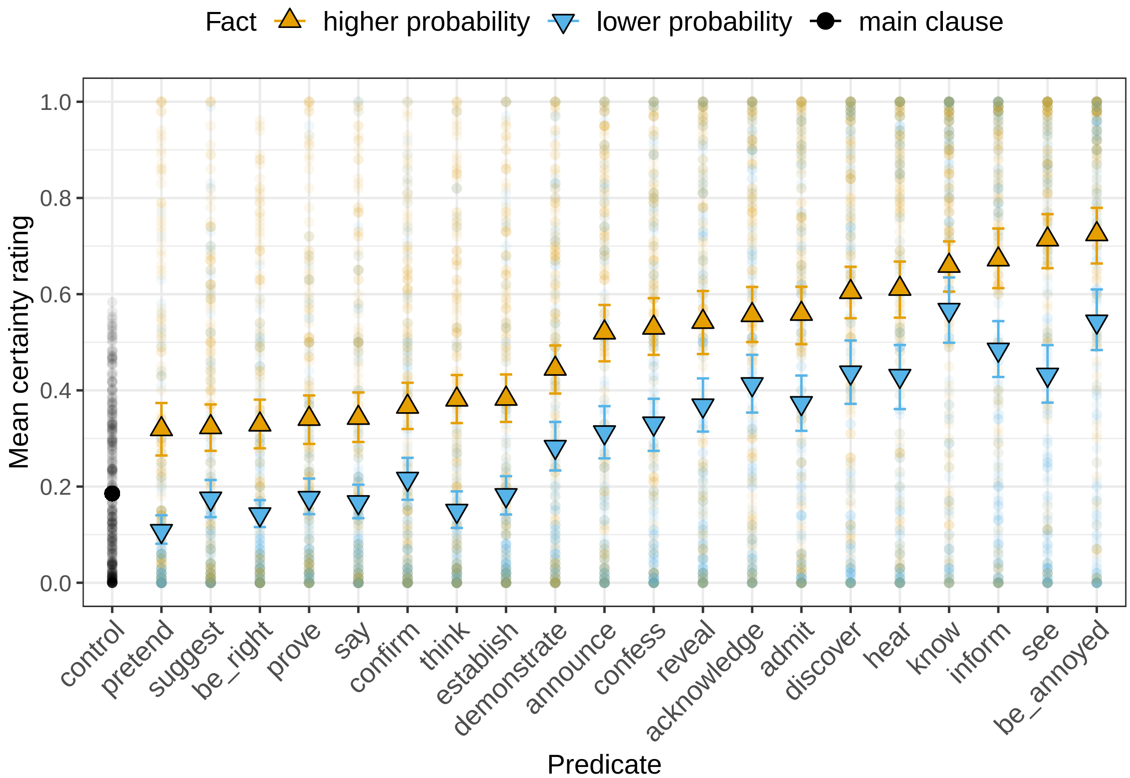


Figure 6: Mean certainty ratings by predicate and prior probability of the content of the complement in Exp. 2b. Error bars indicate 95% bootstrapped confidence intervals. Light dots indicate participants' ratings.

of clause-embedding predicates, like factive ones (e.g., Heim, 1983; van der Sandt, 1992; Abrusán, 2011, 2016; Romoli, 2015; Simons, Beaver, Roberts, & Tonhauser, 2017).

References

Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics & Philosophy*, *34*, 491–535. doi: 10.1007/s10988-012-9108-y

Abrusán, M. (2016). Presupposition cancellation: Explaining the 'soft-hard' trigger distinction. *Natural Language Semantics*, *24*, 165-202. doi: 10.1007/s11050-016-9122-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beaver, D., & Geurts, B. (2014). Presupposition. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy.* Stanford University.

Bicknell, K., & Rohde, H. (2014). Dynamic integration of pragmatic expectations and real-world event knowledge in syntactic ambiguity resolution. In *Proceedings of the 31st annual conference of the cognitive science society* (p. 1216-1221). Austin, TX: Cognitive Science Society.

Boyce, V., von der Malsburg, T., Poppels, T., & Levy, R. (2018). Implicit gender in the production and comprehension of pronominal references. In *Proceedings of the 31th annual cuny conference on human sentence processing.*

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Chambers, C., Tanenhaus, M., Eberhard, K., Filip, H., & Carlson, G. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*, 30-49.

Chambers, C., Tanenhaus, M., & Magnuson, J. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology*, *30*, 687-696.

de Marneffe, M., Simons, M., & Tonhauser, J. (2019). The CommitmentBank: Investigating projection in naturally occurring discourse. *Sinn und Bedeutung*, *23*, 107-124. doi: 10.18148/sub/2019.v23i2.601

Degen, J., Hawkins, R. X. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to 'overinformative' referring expressions. *Psychological Review*, *127*, 591621. doi: 10.1037/rev0000186

Degen, J., Tessier, M., & Goodman, N. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. Retrieved from `http://dx.doi.org/10.1016/j.tics.2016.08.005` doi: 10.1016/j.tics.2016.08.005

Hagoort, P., Hald, L., Mastiaansen, M., & Petersson, K. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438-441.

Hald, L. A., Steenbeek-Planting, E., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. evidence from the n400. *Brain research*, *1146*, 210-218.

Hanna, J., & Tanenhaus, M. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*, 105-115.

Heim, I. (1983). On the projection problem for presuppositions. *West Coast Conference on Formal Linguistics*, *2*, 114-125.

Jaynes, E., & Kempthorne, O. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (pp. 175–257). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-010-1853-1

Kravtchenko, E., & Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. In *Proceedings of the 37th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Langendoen, T., & Savin, H. (1971). The projection problem for presuppositions. In C. Fillmore & T. Langendoen (Eds.), *Studies in linguistic semantics* (p. 54-60). New York: Holt, Rinehart and Winston.

Lorson, A. (2018). *The influence of world knowledge on projectivity.* (M.A. thesis, University of Potsdam)

Mahler, T. (2020). The social component of projection behavior of clausal complements. *Linguistic Society of America*, *5*, 777-791. doi: 10.3765/plsa.v5i1.4703

Mitchell, M., Reiter, E., & Van Deemter, K. (2013). Typicality and object reference. *Cognitive Science*.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*, *23*(1), 103–123. doi: 10.3758/s13423-015-0947-8

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas, Part II. *Linguistics and Language Compass*, *10*(11), 591–613. doi: 10.1111/lnc3.12207

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Romoli, J. (2015). The presuppositions of soft triggers are obligatory scalar implicatures. *Journal of Semantics*, *32*, 173-291. doi: 10.1093/jos/fft017

Rubio-Fernández, P. (2016, February). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00153

Simons, M., Beaver, D., Roberts, C., & Tonhauser, J. (2017). The Best Question: Explaining the projection behavior of factive verbs. *Discourse Processes*, *3*, 187-206. doi: 10.1080/0163853X.2016.1150660

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54.

Stevens, J., de Marneffe, M.-C., Speer, S., & Tonhauser, J. (2017). Rational use of prosody predicts projectivity in manner adverb utterances. *Annual Meeting of the Cognitive Science Society*, *39*, 1144-1149.

Tessler, M. H., & Goodman, N. (2019). The language of generalization. *Psychological Review*, *126*, 395-436.

Tonhauser, J. (2016). Prosodic cues to speaker commitment. *Semantics and Linguistic Theory*, *26*, 934-960. doi: 10.3765/salt.v26i0.3788

Tonhauser, J., Beaver, D., & Degen, J. (2018). How projective is projective content? Gradience in projectivity and at-issueness. *Journal of Semantics*, *35*, 495-542. doi: 10.1093/jos/ffy007

Tonhauser, J., & Degen, J. (ms). *Which predicates are factive? An empirical investigation.* (Manuscript, The Ohio State U and Stanford U)

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, *9*, 333-377. doi: 10.1093/jos/9.4.333

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic bulletin & review*, *14*(4), 770–775.

Westerbeek, H., Koolen, R., & Maes, A. (2015, July). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00935

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*(1), 1–191.

Supplemental material for *Prior probability predicts projection*

A Experiment 1: Target and control stimuli

This list gives the 20 clauses of the target stimuli with the lower and higher probability facts, respectively:

1. Mary is pregnant. Facts: Mary is a middle school student / Mary is taking a prenatal yoga class
2. Josie went on vacation to France. Facts: Josie doesn't have a passport / Josie loves France
3. Emma studied on Saturday morning. Facts: Emma is in first grade / Emma is in law school
4. Olivia sleeps until noon. Facts: Olivia has two small children / Olivia works the third shift
5. Sophia got a tattoo. Facts: Sophia is a high end fashion model / Sophia is a hipster
6. Mia drank 2 cocktails last night. Facts: Mia is a nun / Mia is a college student
7. Isabella ate a steak on Sunday. Facts: Isabella is a vegetarian / Isabella is from Argentina
8. Emily bought a car yesterday. Facts: Emily never has any money / Emily has been saving for a year
9. Grace visited her sister. Facts: Grace hates her sister / Grace loves her sister
10. Zoe calculated the tip. Facts: Zoe is 5 years old / Zoe is a math major
11. Danny ate the last cupcake. Facts: Danny is a diabetic / Danny loves cake
12. Frank got a cat. Facts: Frank is allergic to cats / Frank has always wanted a pet
13. Jackson ran 10 miles. Facts: Jackson is obese / Jackson is training for a marathon
14. Jayden rented a car. Facts: Jayden doesn't have a driver's license / Jayden's car is in the shop
15. Tony had a drink last night. Facts: Tony has been sober for 20 years / Tony really likes to party with his friends
16. Josh learned to ride a bike yesterday. Facts: Josh is a 75-year old man / Josh is a 5-year old boy
17. Owen shoveled snow last winter. Facts: Owen lives in New Orleans / Owen lives in Chicago
18. Julian dances salsa. Facts: Julian is German / Julian is Cuban
19. Jon walks to work. Facts: Jon lives 10 miles away from work / Jon lives 2 blocks away from work
20. Charley speaks Spanish. Facts: Charley lives in Korea / Charley lives in Mexico

In the target stimuli of the projection block of Exp. 1, eventive predicates, like *discover* and *hear*, were realized in the past tense and stative predicates, like *know* and *be annoyed*, were realized in the present tense. The direct object of *inform* was realized by the proper name *Sam*. The subject of the clause-embedding predicate and the speaker of the target stimuli were realized by a proper name.

The following list gives the six clauses that were used in the control and filler stimuli of Exp. 1, with their facts. In the prior block, these six clauses were embedded under *How likely is it that...?*. The projection block featured polar questions variants of the clauses.

1. Zack is coming to the meeting tomorrow. Fact: Zack is a member of the golf club.
2. Mary's aunt is sick. Fact: Mary visited her aunt on Sunday.
3. Todd played football in high school. Fact: Todd goes to the gym 3 times a week.
4. Vanessa is good at math. Fact: Vanessa won a prize at school.
5. Madison had a baby. Fact: Trish sent Madison a card.
6. Hendrick's car was expensive. Fact: Hendrick just bought a car.

B Data exclusion

Table A1 presents how many participants' data were excluded from the analyses based on the exclusion criteria. The first column records the experiment, the second ('recruited') how many participants were recruited, and the final column ('remaining') how many participants' data entered the analysis. The 'Exclusion criteria' columns show how many participants' data were excluded based on the two exclusion criteria:

- 'language': Participants' data were excluded if they did not self-identify as native speakers of American English.
- 'controls': In Exps. 1 and 2b, participants' data were excluded if their response mean on the 6 control items was more than 2 sd above the group mean. In Exp. 2a, participants' data were excluded if their response mean was more than 2 sd below the group mean of the control in (6a) or more than 2 sd above the group mean of the control in (6b).

		Exclusion criteria		
	recruited	language	controls	remaining
Exp. 1	300	3	11	286
Exp. 2a	95	8	12	75
Exp. 2b	300	23	11	266

Table A1: Data exclusion in Exps. 1 and 2

C Model details for Experiments 1 and 2

JT: do we need this for this journal? if not, cool, let's delete. if yes, needs to be adjusted.

This supplement provides details on the data analysis conducted for Exps. 1, 2, and 3. We first motivate the use of Beta regression rather than linear regression in Exps. 1a, 2a, and 3a (section C.1) and then provide a brief primer on how to interpret Bayesian mixed effects Beta regression models (section C.2). We then report the model outputs for Exps. 1, 2, and 3 (section C.3).

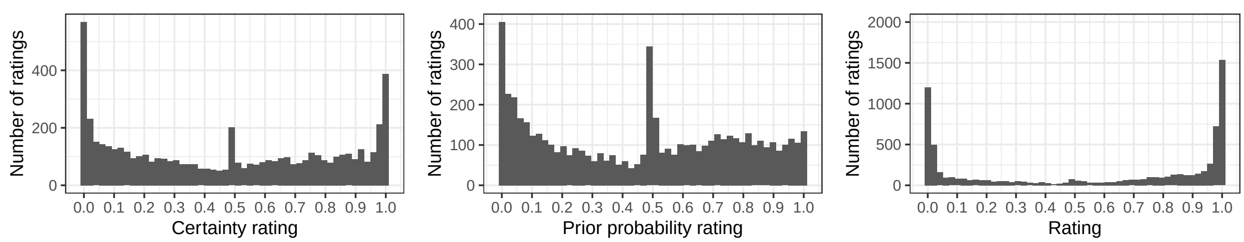
C.1 Motivation for using Bayesian mixed effects Beta regression

There are three separate pieces to motivate: the use of *mixed effects*, the use of a *Bayesian* rather than *frequentist* models, and the use of *Beta regression* rather than *linear regression*.

Using mixed effects refers to the practice of modeling the outcome variable, here slider ratings or proportions of 'yes' ratings, as a function of not just fixed effects of interest (i.e., predicate) but also as the result of possible random variability that is not of theoretical interest (e.g., random by-participant or by-item variability). This is standard practice in psycholinguistic studies and allows the researcher to trust that any observed effects of theoretical interest are true average effects rather than the result of idiosyncratic behavior (e.g., of participants or items). This is also the motivation for using mixed effects in Exps. 1b, 2b, and 3b.

Using Bayesian models rather than frequentist models is increasingly becoming the norm in psycholinguistic studies as computational power has increased and running Bayesian models has become more accessible with the introduction of R packages such as `brms` (Bürkner, 2017). The presence of an effect in frequentist models is evaluated by checking whether the p-value is smaller than .05, where the p-value is defined as the probability of obtaining data that is as skewed or more skewed than the observed data if the null-hypothesis was true, i.e., if the hypothesized effect was absent. Parameter estimates in frequentist models are obtained via maximum-likelihood techniques, i.e., by estimating the parameter values that maximize the probability of observing the data. Bayesian models, by contrast, return a full posterior distribution over parameter values that take into account not just the probability of the data under the parameter values, but also the prior probability of parameter values. In order to evaluate the evidence for an effect of a predictor of interest, one can report 95% credible intervals and the posterior probability $P(\beta < 0)$ or $P(\beta > 0)$ that the predictor coefficient β is either lower or greater than zero, depending on the direction of the expected effect. A 95% credible interval (CI) demarcates the range of values that comprise 95% of probability mass of the posterior beliefs such that no value inside the CI has a lower probability than any point outside it (Jaynes & Kempthorne, 1976; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). There is substantial evidence for an effect if zero is (by a reasonably clear margin) not included in the 95% CI and $P(\beta > 0)$ or $P(\beta < 0)$ is close to zero or one. Posterior probabilities indicate the probability that the parameter has a certain value, given the data and model – these probabilities are thus *not* frequentist p-values. In order to present statistics as close to widely used frequentist practices, and following Nicenboim & Vasishth, 2016, we defined an inferential criterion that seems familiar (95%), but the strength of evidence should not be taken as having clear cut-off points (such as in a null-hypothesis significance testing framework).

Using Beta regression rather than linear regression was motivated by the violation of two of the assumptions of linear regression: first, that residuals be normally distributed (where "residuals" refers to the residual error for each data point after fitting the model), and second, that the error term exhibit homoscedasticity (that it be roughly the same across different conditions). Slider ratings data has the property of being bounded by its endpoints (which we code as 0 and 1, respectively). This often leads to "bunching" behavior at the endpoints (see Fig. A1 for the distribution of raw ratings in Exps. 1a, 2a, and 3a).



(a) Exp. 1 certainty ratings. (b) Exp. 1 prior ratings. (c) Exp. 2 prior ratings.

Figure A1: Histograms of raw slider ratings in Exps. 1 and 2.

This "bunching" behavior, in turn, can lead to the violation of both of the above assumptions of linear regression. Intuitively, these assumptions are violated because conditions that elicit ratings closer to endpoints necessarily have a compressed variance; consequently, a condition's mean and its variance are not independent. Beta regression is useful here because it allows for modeling an arbitrarily distributed outcome variable in the [0,1] interval. The Beta distribution is characterized by two parameters, one capturing the mean μ of the distribution and one capturing its precision ϕ, a measure of dispersion. The greater the precision, the more concentrated the values are around the mean, i.e., the lower the variance of the distribution. We follow Smithson and Verkuilen (2006) in modeling μ and ϕ separately for each predictor. That is, we allow each predictor to affect both the mean and the precision of the outcome variable's distribution.

C.2 Coding choices and interpreting model output

The outcome variable in Exps. 1a, 2a and 3a (slider ratings) contained the values 0 and 1, which Beta regression is undefined for. We therefore applied a common transformation to ratings before the main analysis that rescales values y to fall in the open unit interval (0,1) (Smithson & Verkuilen, 2006). First, we apply $y' = (y - a)/(b - a)$, where b is the highest possible slider rating and a is the smallest possible slider rating. The range is then compressed to not include 0 and 1 by applying $y'' = [y'(N - 1) + 1/2]/N$, where N is the total number of observations.

The mean parameter μ is modeled via a logit link function (default for Beta regression in `brms`), though other links that squeeze μ into the [0,1] interval are possible. The dispersion parameter ϕ is modeled via a log link, which ensures that values of ϕ are strictly positive, which is necessary because a variance cannot be negative.

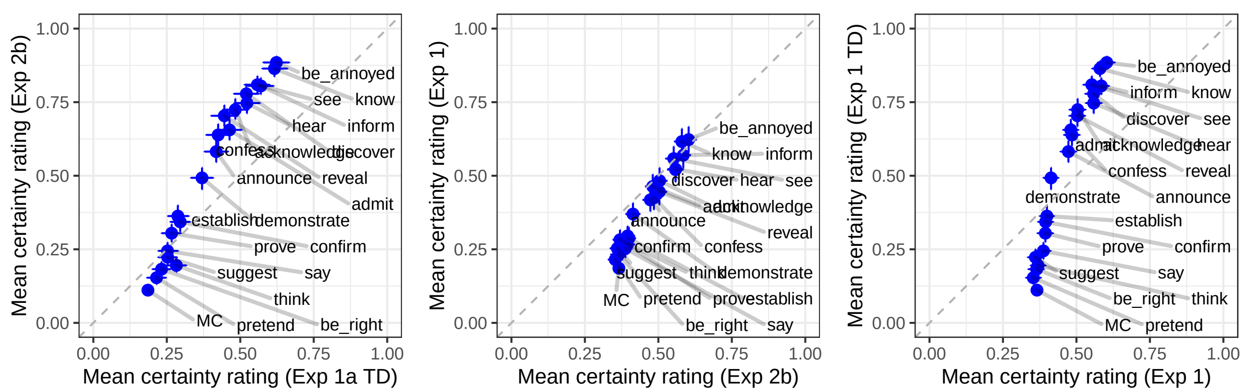
We allowed both μ and ϕ to vary as a function of predicate, with reference level set to main clause control in Exp. 1a, entailing control in Exp. 2a and contradictory control in Exp. 3a. We also allowed random intercept adjustments to each parameter by participant and by item, where item was defined as a unique combination of a predicate and a complement clause. Four chains converged after 2000 iterations each (warmup = 1000, $\hat{R}$ = 1 for all estimated parameters) with a target acceptance rate of .95 and a maximum treedepth of 15.

C.3 Model outputs for Experiments 1, 2 and 3

The three tables in this section show the model outputs for Exps. 1, 2 and 3, respectively: Table **??** for Exps. 1a and 1b, Table **??** for Exps. 2a and 2b, and Table **??** for Exps. 3a and 3b. Each table shows maximum a posteriori (MAP) model estimates for projection ratings from the Beta regression model (left and middle column, mean μ and precision ϕ) and the logistic regression model (right column, β) with 95% credible intervals.

D Projection comparisons

Fig. A2 compares the mean certainty ratings of the predicates and main clause controls in Exp. 1, Exp. 2b, and Tonhauser & Degen, ms Exp. 1a (abbreviated 'Exp. 1a TD'). The Spearman rank correlations were .986 (Exp. 2b vs. Exp. 1a TD), .971 (Exp. 1 vs. Exp. 2b) and .974 (Exp. 1a TD vs. Exp. 1).



(a) Exp. 2b vs. Exp. 1a TD. (b) Exp. 1 vs. Exp. 2b. (c) Exp. 1a TD vs. Exp. 1.

Figure A2: Comparisons of mean by-predicate certainty ratings from Exp. 1, Exp. 2b, and Tonhauser & Degen's Exp. 1a (abbreviated 'Exp. 1a TD'). Error bars indicate 95% bootstrapped confidence intervals.

E Prior probability results in Exp. 2a

Fig. A3 plots the mean prior probabilities of the 20 contents by fact. Participants' ratings are given as light dots. The mean prior probability rating for each content was higher when the content was presented with the higher probability fact than when it was presented with the lower probability fact.

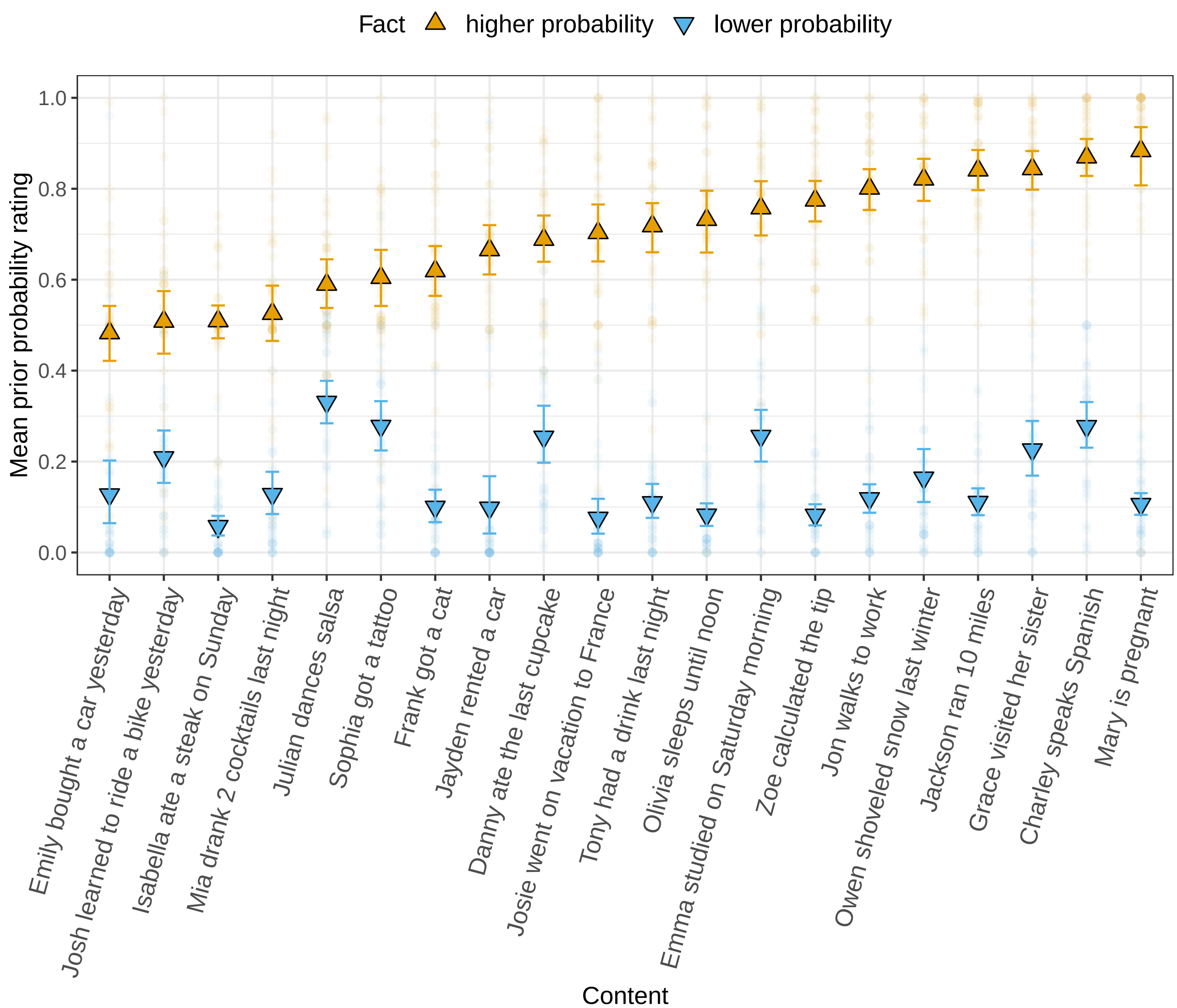


Figure A3: Mean prior probability by content and fact in Exp. 2a. Error bars indicate 95% bootstrapped confidence intervals. Light dots indicate participants' ratings.