Cognition 265 (2025) 106265

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

Word and construction probabilities explain the acceptability of certain long-distance dependency structures

Moshe Poliak [*], Curtis Chen [1], Edward Gibson

Department of Brain and Cognitive Sciences, MIT, United States

ARTICLE INFO

Keywords:
Sentence processing
Frequency effects
Long-distance dependencies
Syntactic islands

ABSTRACT

The factors that affect the acceptability of long-distance extractions have long been debated, with multiple accounts proposed. Liu et al. (2022) proposed a succinct probability-based account of a sub-class of these kinds of materials, wh-questions with long-distance dependencies across sentence-complement verbs (e.g., "What did Mary whine that John bought?"). The explanation that they proposed was that the acceptability of such sentences depends on the probability of the verb-frame of the intermediate verb (e.g., "whine that"). In the current work, we evaluate some potentially simpler probability-based accounts on Liu et al.'s original data set, and show how an alternative (but also probability-based) approach accounts for the data better. We replicate their experiment and conduct the same analysis on the new dataset, finding the same results. Finally, we apply the same analysis to wh-questions with predicate adjectives (e.g., "What was Mary glad that John bought?"), and again find similar results. We conclude that the acceptability of such constructions is higher the more probable the words and constructions that make up the sentence are.

1. Introduction

An enduring puzzle in language science has concerned differences in the acceptability of sentences with a dependency across sentence-complement verbs. Consider the following examples in English:

1

(1a) Mary said that Bill kicked something.
(1b) Mary hated that Bill kicked something.
(1c) Mary murmured that Bill kicked something.
(1d) What did Mary say that Bill kicked?
(1e) What did Mary hate that Bill kicked?
(1f) What did Mary murmur that Bill kicked?

It has been observed that declarative sentences like **1a- 1c** are more acceptable than parallel wh-questions like **1d-1f**, which have a dependency relation from the wh-word "what" to the embedded verb "kicked", which crosses another verb (e.g., "say" in **1d;** "hate" in **1e**; "murmur" in **1f**). Going back to Ross (1967), less acceptable long-distance dependencies as in **1e and 1f** have been called syntactic "islands" for the long-distance dependencies (with the metaphor being that there is an extracted element (the wh-word), and it can't get off the "island").

Many researchers have attempted to explain the low acceptability of some long-distance dependency structures in terms of their syntactic configurations (e.g., Baltin, 1982; Chomsky, 1973, 1977, 1986; de Cuba, 2018; Huang, 1982; Kiparsky & Kiparsky, 1971; Rizzi, 1990; Snyder, 1992; Stoica, 2016; Stowell, 1981). In cases like those above, it has been proposed that the verbs "say", "hate," and "murmur" differ in their syntactic representations, giving rise to the differing acceptability of **1d** vs. **1e** and **1f**. In the case of a verb like "say" in **1a** and **1d**, the complement is an embedded clause (e.g., "that Bill kicked something" in **1a**). But the complement of a factive verb like "hate" in **1b** and **1e** is proposed to have an additional embedded empty noun phrase, headed by an empty noun with a meaning like "fact" (e.g., Kiparsky & Kiparsky, 1971). Then the unacceptability of **1e** is proposed to be explained to be parallel to the unacceptability of a sentence like (**2**):

(**2**) What did Mary hate the fact that Bill kicked?

In **2**, it is difficult to connect the wh-word "what" to the complement of "kicked" across the embedded noun phrase headed by "fact", and people will usually find such examples quite bad. Other researchers proposed a similar analysis of the representations associated with verbs like "murmur" (e.g., Snyder, 1992), leading to the proposed badness of **1f**. But a serious problem with these types of accounts is that the only

* Corresponding author.
E-mail addresses: moshepol@mit.edu (M. Poliak), egibson@mit.edu (E. Gibson).
[1] Deceased.

https://doi.org/10.1016/j.cognition.2025.106265
Received 27 June 2024; Received in revised form 14 July 2025; Accepted 16 July 2025

reason for the additional proposed structure in examples like **1e** and **1f** is to make them fit a syntactic complexity account. There are no independent reasons to propose these covert complex structures (Ambridge & Goldberg, 2008; Liu, Ryskin, et al., 2022): the verbs "say", "hate", and "murmur" have superficially similar argument structures in **1a-1c**, and in **1d-1f**. We would need more evidence for a syntactic structure with empty elements. Such observations prompted a variety of theoretical explanations for the unacceptability of these constructions, which leverage semantics, information structure, discourse, and lexical frequency.

While many syntactic theories have grappled with the unacceptability of structures like in **1e, 1f** and **2**, many non-syntactic factors are known to affect acceptability. One key factor that affects acceptability, for any kind of sentence, is its average surprisal, where the surprisal of a word is the negative log probability of a word given the context (Hale, 2001; Levy, 2008; Shain et al., 2024; Smith & Levy, 2013; Wilcox et al., 2023). Note that while the term *frequency* is used often in psycholinguistics, it is usually a proxy for surprisal. Frequency is derived from a corpus, and the corpus has a finite size. Thus, frequency divided by corpus size is probability, and surprisal is merely the negative log probability, a monotonic, strictly decreasing function. Past research has found that the probabilities of words and constructions affect acceptability: individual verbs are rated as more acceptable the more probable they are (Ambridge, 2013); Verb-adjective bigrams are rated as more acceptable the more probable they are (Bybee & Eddington, 2006); each variant of the English dative alternation ("gave the item to the person" vs. "gave the person the item") is rated as more natural the more probable it is (Bresnan, 2007; Bresnan & Ford, 2010); and the more probable a certain argument structure within a verb is, the more acceptable it is (White & Rawlins, 2020).

Building on these ideas—and specifically the approaches of Dąbrowska (2008) and Kothari (2008)—Liu, Ryskin, et al. (2022) proposed a simple probability-based explanation of why certain sentences with dependencies across sentence-complement verbs are more acceptable than others: the more probable a verb-frame is (e.g., "notice that"), the more acceptable the sentence. That is, they propose that, in the examples above, **1d** is more acceptable than **1e** and **1f** because the verb-frame "say that" is more probable than the verb frame "hate that", or the verb frame "murmur that". In line with probability-based explanations, it is plausible that part of why **1a - 1c** are more acceptable than **1d - 1f** is because declarative sentences are more probable than wh-questions. Across 4 experiments that employed wh-questions and clefting, Liu, Ryskin, et al. (2022) found support for the role of verb-frame probability in the acceptability of sentences with a dependency across sentence-complement verbs. Verb-frame probability and construction (regular declarative vs. wh-question or cleft) were associated with significant main effects that predicted acceptability, with no interaction between probability and construction.

The goal of the current work is to better understand the role of probability in the acceptability of sentences with dependencies across sentence-complement verbs. The explanation in Liu, Ryskin, et al. (2022) is that the acceptability of wh-questions with a dependency across sentence-complement verbs depends on the verb-frame probability (the bigram probability of "{verb} that"). However, this quantity is a complex one, being the product of 2 simpler quantities: the lexical probability of the verb and, within the verb, the probability of the argument structure that takes a sentence complement with the complementizer *that* (See Eq. (1)). There is no reason to start the investigation of the effects of probability with verb-frame probability, if one can start with quantities that are simpler a priori. We therefore evaluate simpler but related proposals of what counts as probability in explaining the acceptability of sentences with extractions across sentence-complement verbs, thus pushing the envelope on the probability-based account provided by Liu, Ryskin, et al. (2022). We do so in the spirit of constructionism (e.g., Bybee, 2006, 2010; Croft, 2010; Fillmore, 1988; Goldberg, 1995, 2006, 2019; Steels, 2011, 2013), investigating the idea that the acceptability of wh-questions with dependencies across sentence-complement verbs depends on the probabilities of the lexeme and the constructions it is involved in—the probabilities of the surface verb-frame (as in Liu, Ryskin, et al., 2022), the verb, and the specific argument structure within a given verb, P(that | verb). Throughout this manuscript, probabilities will be used with the log transform for its favorable mathematical properties (not bound in [0,1]) and better connection to psychological constructs (Shain et al., 2024; Smith & Levy, 2013). Unlike surprisal, we will not multiply the log probability by (−1), to preserve the intuition that higher log probability is tied to higher corpus frequency.

The chain rule relates the bigram lexical frame probability to the unigram lexical probability and the probability that a lexeme takes a sentence complement.

$$P(lexeme, that) = P(lexeme)*P(that \mid lexeme) \quad (1)$$

There are several possible patterns that may emerge when predicting acceptability using verb probability, P(that | verb), and their product, verb-frame probability. If verb probability is a significant predictor, then we learn that more probable verbs result in more acceptable sentences with long-distance dependencies across sentence-complement verbs. If P (that | verb) is a significant predictor, then, within each verb, how probable it is to take a sentence complement also explains acceptability. If their interaction is significant, it means that, above and beyond these individual measures, the surface probability of "{verb} that" (the verb-frame probability) also influences acceptability. In all cases, a main effect of construction is predicted such that interrogatives are less acceptable than declaratives. A priori, we do not know which combination of the 3 factors above will play a role in predicting acceptability, and the goal of the current investigation is to test them. Finally, since the results of Liu, Ryskin, et al. (2022) show that there must be some probability effect involved, there is no reason why this effect should be confined specifically to verbs. Therefore, we also similarly investigate the role of probability in the acceptability of sentences with long-distance dependencies across sentence-complement adjectives, as in **2**.

(2.1a) Mary was glad that Bill kicked something.

(2.1b) What was Mary glad that Bill kicked?

To investigate the role of the probabilities of lexemes and their constructions on the acceptability of wh-questions with a dependency across sentence-complement verbs, we combine corpus analyses with analyses of acceptability-judgment data. We start with reporting a corpus analysis that syntactically parsed sentences from the Corpus of Contemporary American English (Davies, 2008) retrieving P(verb), P (that | verb), and P(verb, that) for the sentences in Experiment 2 of Liu, Ryskin, et al. (2022) (of their 4 experiments, only the first 2 investigated wh-questions, and of these 2, the second experiment used a larger set of verbs (48) than in the first experiment (24)). We then reanalyze the data of Experiment 2 from Liu, Ryskin, et al. (2022) using the quantities derived from the corpus analysis. Next, we report a novel replication of the same experiment, adding filler items and fixing potential confounds in a few sentences, and analyze the experiment using the same quantities. Then we turn to adjectives, reporting another corpus analysis using the same corpus and methodology, and then a novel experiment that used an identical design to investigate the acceptability of wh-questions with adjective-frames, and a subsequent replication of that experiment. We conclude with a general discussion of the findings. All the materials, data, and analyses are available on OSF at: https://osf.io/ukyfn/

2. Corpus analysis—verb

Large corpora contain vast information about how language is used. Liu, Ryskin, et al. (2022) tapped into this well of knowledge by extracting the probability of verb-frames by searching for bigrams like "{verb} that." While the extracted quantity—the verb-frame probability, P(verb, that)—explains much of the variation in acceptability ratings of their experimental stimuli, it leaves open some questions. First, the verb-

frame probability, P(verb, that), is a joint probability, and, by the chain rule, a product of two other probabilities: the lexeme probability of the verb, P(verb), and the probability that the verb will have a dependent clause that starts with the complementizer "that" (P(that | verb); see Eq. (1)). Thus, it is possible that it is not verb-frame probability that influences acceptability, but the simpler quantities of verb probability P (verb) and the probability that a verb takes a sentence complement, P (that | verb). Second, reliance on bigram probability, rather than syntactic analysis, may introduce inaccuracies. For example, a word may intervene between a verb and the sentential complement: "Mary said, unconvincingly, that Steve bought a car." These cases may cause the bigram approach to underestimate the true probability of such verb-frames. In contrast, a syntactic parse would, correctly, identify that the verb in this case is taking a sentence complement, even if separated by other words. Similarly, some words, like *guarantee*, may function as a verb or a noun, and so the unigram probability of the string *guarantee* necessarily inflates the true probability of the verb. To understand which quantities are involved in determining acceptability, there is a need for a new corpus analysis that involves parsing the corpus syntactically.

2.1. Method

For this corpus analysis, we used the Corpus of Contemporary American English (CoCA; Davies, 2008), which contains hundreds of millions of words, sampled from the internet, blogs, and the following genres between 1990 and 2019: academic, fiction, magazines, news, spoken, and television and movies. We extracted all the sentences from the corpus that contained any of the forms of the verbs of interest (e.g., think, thinks, thinking, thought). Then, we randomly selected 30,000 sentences and parsed their dependency structure using the Stanza library in Python (Qi et al., 2020). We computed the probabilities that are relevant to the current analysis: P(verb), P(that | verb), and P(verb, that). Additionally, we computed the probability that the verb takes a sentence complement, with or without the complementizer *that*, a quantity that can only be derived via syntactic analysis.

To arrive at the total verb frequency in the corpus, we scaled the verb counts by the total counts of the verb forms in the corpus. So, for example, of the 30,000 sentences that were sampled with the various forms of the verb *think*, the forms were identified as a verb (and not, for example, as a noun) 27,157 times. The total number of sentences that were retrieved from the corpus that included any string that matched any form of the verb "think" (i.e., think, thinks, thinking, thought) was 1,731,371. Thus, we infer that the frequency of "think" as a verb in the corpus is (27,157/30,000)*1,731,371 = 1,567,296. (Some rare verbs, like "blab," had fewer than 30,000 instances in the entire corpus, and thus were analyzed entirely, with a total count of less than 30,000). To transform the frequencies into probabilities, we divided the verb frequency in the corpus by the word count that we retrieved from the corpus. The verb frame counts were divided by the size of the corpus minus one. The probability of a sentential clause given the verb was computed by dividing the counts of the verb frame by the total counts of the verb. Finally, we log-transformed all the probabilities, relating the counts to surprisal (negative log probability given context), which is a core quantity in language cognition (Hale, 2001; Levy, 2008; Smith & Levy, 2013). For the purposes of modeling and visualization, we centered all the log-probabilities.

2.2. Results

The centered log-transformed P(verb) and P(that | verb) for each verb are visualized in Fig. 1. The raw counts for verbs ranged from 393 to 3,157,391, with a median of 36,768. The correlation between the log verb-frame probability used in Liu, Ryskin, et al. (2022) (from Google n-Grams) and the log verb-frame probability in our corpus analysis was

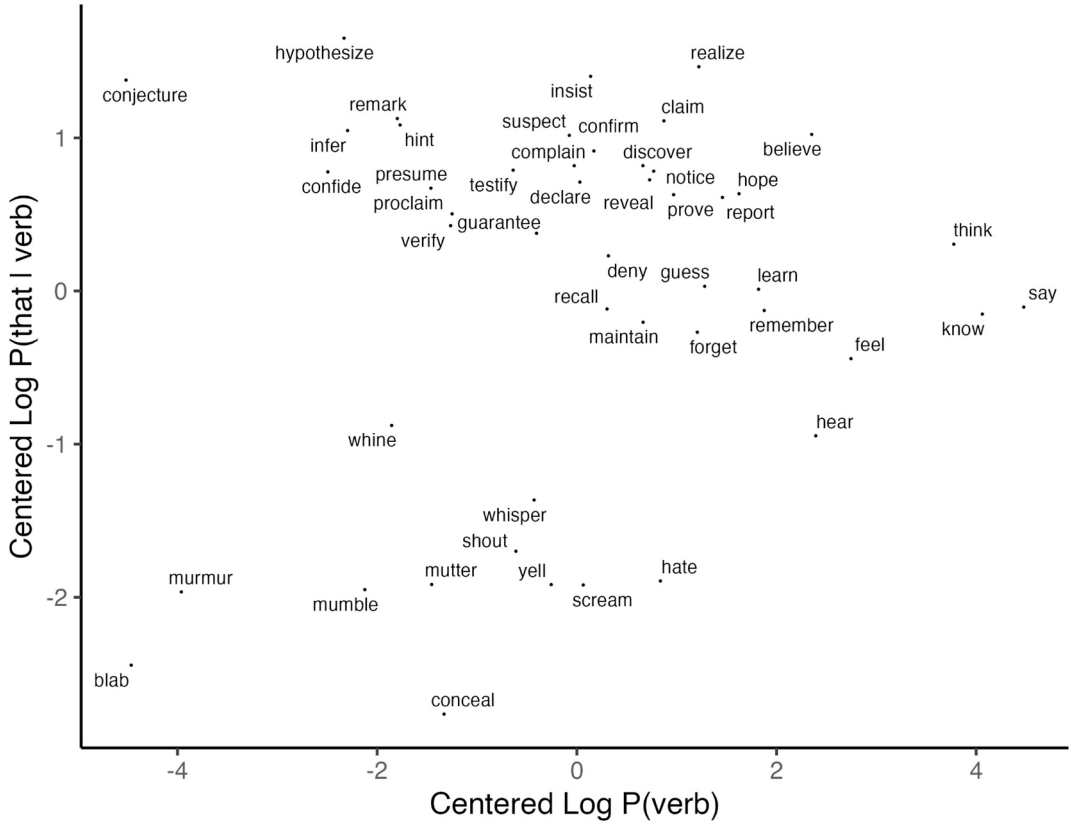

Fig. 1. Verbs varied in their probability, P(verb), and the probability that the verb will take a clause complement with the complementizer *that*, P(that | verb), two largely independent quantities.

0.96. The following probabilities are all computed based on the current corpus analysis: log verb-frame probability had a 0.88 correlation with P (verb) and 0.59 with P(that | verb). Importantly, P(verb) and P(that | verb) were very weakly correlated, with $r = 0.14$. The correlation between P(that | verb) and the probability that the verb had a sentence complement (regardless of the presence of *that*) was 0.98.

2.3. Discussion

The corpus analysis revealed how key probabilities associated with verb-frame probability relate to each other. We found an almost perfect correlation ($r = 0.96$) between the verb-frame probability that Liu, Ryskin, et al. (2022) extracted from Google n-Grams by querying for the surface string "{verb} that" (counting all the forms of the verb) and the verb-frame probability that we extracted by syntactically parsing CoCA. This suggests that the bigram probability of "{VERB} that" was an accurate proxy for the true P(verb, that). Computing the correlation between P(verb) and P(that | verb), we found that they are largely independent, having merely a correlation of 0.14 (and see Fig. 1). This finding is important: had they been highly correlated, there would have been little motivation to study the independent effects of P(verb) and P(that | verb) on acceptability. (And it would prove a challenge as their effects would have been hard to isolate). However, P(verb) and P(that | verb) are largely independent, and therefore we can ask the question at the core of this manuscript: what probability-based quantities affect the acceptability of long-distance extractions and equivalent declaratives.

Note that another related quantity is the probability that the verb takes a sentence complement, independent of an explicit complementizer (e.g., "Mary said John bought something" vs. "Mary said *that* John bought pizza"). However, we will not discuss this quantity in the current paper because (1) in the current corpus analysis it had a 0.98 correlation with verb-frame probability, which suggests that these quantities capture the same construct, and (2) this quantity does not feature directly in the experimental materials, since the experimental constructions always use the explicit complementizer *that*.

3. Reanalysis of Liu, Ryskin, et al. (2022), Experiment 2

The account that Liu, Ryskin, et al. (2022) proposed is that *verb-frame probability* explains the variability in the acceptability of wh-questions with long-distance dependencies across sentence-complement verbs and equivalent declaratives. Liu, Ryskin, et al. (2022) define verb-frame probability for sentence-complement verbs as the sum of the probabilities that the verb in question, in any of its conjugations, is immediately followed by *that* in the Google Books corpus. This quantity, P(verb, that), is the product of two other quantities that are largely independent: the probability of the verb, P(verb), and the probability that the verb will be followed by *that*, P(that | verb). Liu, Ryskin, et al. (2022) did not use these quantities in their analysis. We thus use the results of the corpus analysis above to investigate the effects of these factors on the acceptability of such constructions.

Liu, Ryskin, et al. (2022) conducted 4 experiments: the first 2 with wh-questions and the last 2 with clefts. Of the first 2 experiments, the first used 24 verbs, and the second added 24 more verbs, for a total of 48 verbs. In Experiment 1, with 24 verbs, Liu, Ryskin, et al. (2022) used a Likert scale, and then, in Experiment 2, which included additional materials, they used a binary scale (natural / unnatural). They found similar effects of verb-frame probability regardless of the size of the scale. This insensitivity to scale is consistent with previous literature, which has shown that acceptability judgments are largely invariant given the rating scale (binary, Likert, or magnitude estimation; e.g., Sprouse et al., 2013; Weskott & Fanselow, 2011). Therefore, we reanalyze and replicate the more comprehensive Experiment 2 of Liu, Ryskin, et al. (2022) using their binary scale. We evaluate the sensitivity of participants to the specific labels used in a binary scale in Appendix A (natural vs. unnatural; good vs. bad; acceptable vs. unacceptable; and grammatical vs. ungrammatical), and found that acceptability judgments are not sensitive to the label used. We also evaluate the sensitivity of participants to a binary scale vs. a 5-point Likert scale in Appendix B, and again find that the inference from the task remains the same regardless of the scale.[2]

3.1. Method

We shall now describe the method used by Liu, Ryskin, et al. (2022), with two goals. The first is to introduce the reader to the experiment that Liu, Ryskin, et al. (2022) conducted and that we are re-analyzing. The second is to prepare the ground for the description of the replication of this experiment, which we conducted and are reporting below (Section 4).

3.1.1. Materials

Participants were presented wh-questions with long-distance dependencies across sentence-complement verbs and parallel declaratives, following the patterns below:

{Name 1} {V1} that {Name 2} {V2} something.
e.g.: Mary said that John bought something.
What did {Name 1} {V1} that {Name 2} {V2}?
e.g.: What did Mary say that John bought?

Liu, Ryskin, et al. (2022), Experiment 2, used 48 unique V1 matrix verbs (e.g., *said*) and paired each one with 6 V2 embedded verbs (e.g., *bought*). Each matrix verb was paired once with each of the 6 embedded verbs in each possible construction, resulting in 576 ($=48 \times 6 \times 2$) total experimental items. The materials were split into 2 lists, each with 6 sentences per matrix verb (288 sentences per participant). Most V1 verbs were paired with verbs from the list of V2 action verbs below. However, "feel" and "insist" are incoherent with action verbs (e.g., #"Mary felt that Jon bought something.") and thus were paired with V2 mental verbs. The verbs "hope", "guarantee", "conjecture", and "hypothesize," which are coherent with both action and mental verbs, were paired with a mix of V2 action and mental verbs.

Full list of V1:

feel, say, believe, hope, think, report, declare, claim,
know, remember, realize, notice, discover, forget, learn, hate.
whisper, mumble, murmur, mutter, whine, shout, yell, scream
hear, recall, blab, conjecture, conceal, proclaim, hint, remark,
infer, confirm, deny, guess, confide, maintain, testify, reveal,
suspect, verify, prove, insist, guarantee, presume, hypothesize,
complain

Full list of V2:

Action (6): bought, wrote, sold, took, broke, stole
Mental (6): wanted, liked, disliked, preferred, needed, loved

3.1.2. Participants

Participants ($N = 120$) were recruited from Amazon Mechanical Turk (MTurk, and see Liu, Ryskin, et al., 2022, filtering for participants with an IP address in the United States. Liu, Ryskin, et al. (2022)

[2] Note that a binary scale does not imply binary acceptability. Mahowald et al. (2016) observed that acceptability exists on a latent continuum. That is, although the scale on which participants indicate their judgments may be binary, there appears to be an underlying continuum of acceptability from very bad to very good. The reason why a latent acceptability continuum may be inferred from binary judgments is that responses are probabilistic (even the same person might judge the same materials differently on different occasions). This is also reflected in the statistical models that we use to model binary and Likert data: in both cases, acceptability is modeled as a probability of giving a positive response or a rating.

excluded 10 participants due to them indicating that English was not their native language or due to failing more than 15 % of the comprehension questions, leaving data from 110 participants.

3.1.3. Procedure

Participants were asked to indicate whether English was their native language. Then participants were shown the experimental items, each paired with a binary acceptability rating scale (natural / unnatural), followed by a comprehension question (e.g., "Does the sentence mention Andy?").

3.2. Results

The data (see Fig. 2) were modeled with mixed-effects logistic regressions in R (R Core Team, 2024), using the lme4 package (Bates et al., 2015), and the data were processed and visualized using the tidyverse (Wickham et al., 2024) and ggrepel (Slowikowski et al., 2024). First, we fit the same model as Liu, Ryskin, et al. (2022): regressing the binary acceptability judgment on the centered log verb frame probability, P(verb, that), from the corpus analysis above, construction (sum-coded, −0.5 = declarative, 0.5 = interrogative), and their interaction. We used the maximal random effects structure justified by the design (Barr et al., 2013), including random intercepts for participants and matrix verbs, as well as random slopes that mirrored the fixed effects within participants, and a random slope for construction within matrix verbs. Second, we fit a model using the same specifications except that we replaced the term for P(verb, that) with the centered log probabilities P(verb) and P(that | verb) and their interaction, as well as all their interactions with construction. Note that the interaction between P(verb) and P(that | verb) is the superadditive effect of their product, P(verb, that), the verb frame probability. After changing the optimizer and increasing the number of iterations, this model did not converge with the full random effects specification, so we simplified it by removing correlations for participant-level random effects as suggested by Barr et al. (2013). The model with terms for both P(verb) and P(that | verb) outperformed the model that only used P(verb, that): it has lower AIC (9517 vs. 9635), lower BIC (9743 vs. 9777), and it is significantly better based on a likelihood ratio test (Chi-Squared = 137.67, DF = 10, $p < .001$).

We will now report the output of the model with terms for P(verb) and P(that | verb). The model had an intercept of 5.454. Verbs with higher P(verb) were significantly more likely to be rated as acceptable (estimate = 0.320, std. error = 0.063, z-value = 5.104, $p < .001$). Verbs with higher P(that | verb) were significantly more likely to be rated as acceptable (estimate = 0.466, std. error = 0.113, z-value = 4.129, $p < .001$). Wh-questions were significantly less likely to be rated as acceptable (estimate = −2.686, std. error = 0.343, z-value = −7.825, $p < .001$). Finally, an interaction emerged between P(verb) and P(that | verb), representing the superadditive effect of verb frame probability, such that the higher the probability, the more likely participants were to rate the sentence as acceptable (estimate = 0.219, std. error = 0.051, z-value = 4.281, $p < .001$). The other interactions were not significant.

3.3. Discussion

In this section, we reanalyzed the data from Liu, Ryskin, et al. (2022) using a more complex model that uses simpler theoretical quantities. Rather than predicting acceptability using only (verb, that) and construction, we fit models that predict acceptability from P(verb), P(that | verb), construction, and all of their interactions. We found that this model outperforms the original model by Liu, Ryskin, et al. (2022). A priori, this didn't have to be the case: the models could have resulted in a similar likelihood, and the interpretation of the model proposed in the current work may have showed that only a significant interaction is present, reflecting that only P(verb, that) influences the acceptability of the constructions at question. However, this was not the case. While the interaction of P(verb) and P(that | verb) (which represents the superadditive effect of P(verb, that)) has a positive effect on the probability of rating the sentence as natural, so do the main effects of P(verb) and P(that | verb). This suggests a "yes, and"-answer: the acceptability of the constructions in questions is increased in proportion to P(verb), P(that | verb), and P(verb, that). Similar to Liu, Ryskin, et al. (2022) we find a significant effect of construction, such that interrogatives are less acceptable than declaratives. Note that a probability-based account makes a prediction here as well: declarative sentences are more probable than wh-questions and thus are predicted to be more acceptable, though multiple other differences exist between the constructions as well (e.g., dependency length). Like in Liu, Ryskin, et al. (2022), there were no interactions between either probability measure and construction type.

4. Replication of Liu, Ryskin, et al. (2022), Experiment 2

We conducted a replication of Experiment 2 of Liu, Ryskin, et al. (2022) to assess the reliability of their findings, with some changes to the materials in response to possible shortcomings of the original experiment.

4.1. Method

The method was identical to Liu, Ryskin, et al. (2022) except for the changes detailed below.

4.1.1. Materials

The target materials remained the same, except that 4 V2 verbs were exchanged to avoid potential ambiguity[3]: *liked*, *disliked*, *loved*, and *wrote*, were replaced with *required*, *desired*, *understood*, and *made*. Furthermore, we included filler items to address concerns about the lack of distractor (filler) items in the original study (Richter & Chaves, 2020). Filler items differed from critical items by using intransitive V2 verbs that do not take an object in the stimulus sentence, which resulted in yes/no interrogatives and simpler declaratives:

{Name 1} {V1} that {Name 2} {V2}.
ex. John said that Mary cried.
Did {Name 1} {V1} that {Name 2} {V2}?
ex. Did John say that Mary cried?

Name 1, Name 2, and V1 were all reused from the target sentences, and V2 was drawn from a novel set of 12 verbs that could function as intransitives:

Action (7): cried, laughed, smiled, moved, arrived, waited, escaped
Mental (5): agreed, disagreed, concurred, understood, misunderstood

These construction pairs appear similar to the target constructions but importantly do not include interrogatives about the object of the embedded verb. We constructed 6 no-extraction pairs per V1 verb within each construction such that there was a 50–50 balance between targets and no-extraction items. No-extraction items were added to each of the three experimental lists in the same manner as target items (no repeating sentences in any list, with two appearances of each V1 within each construction).

Due to doubling the number of items in the experiment, we were

[3] Consider the following example: "What did Mary say that Bill liked?" There are two possible readings of this wh-question, which come from two possible declarative counterparts, respectively:

(A) Mary said [that [S Bill liked something]].

(B) Mary said [NP [something] [that [S Bill liked __]]].

For our purposes, (B) is a confound; (A) is the intended declarative counterpart from which extraction should take place. This confound occurs for V2 verbs "liked", "disliked", "loved", and "wrote", motivating their removal from the V2 set.

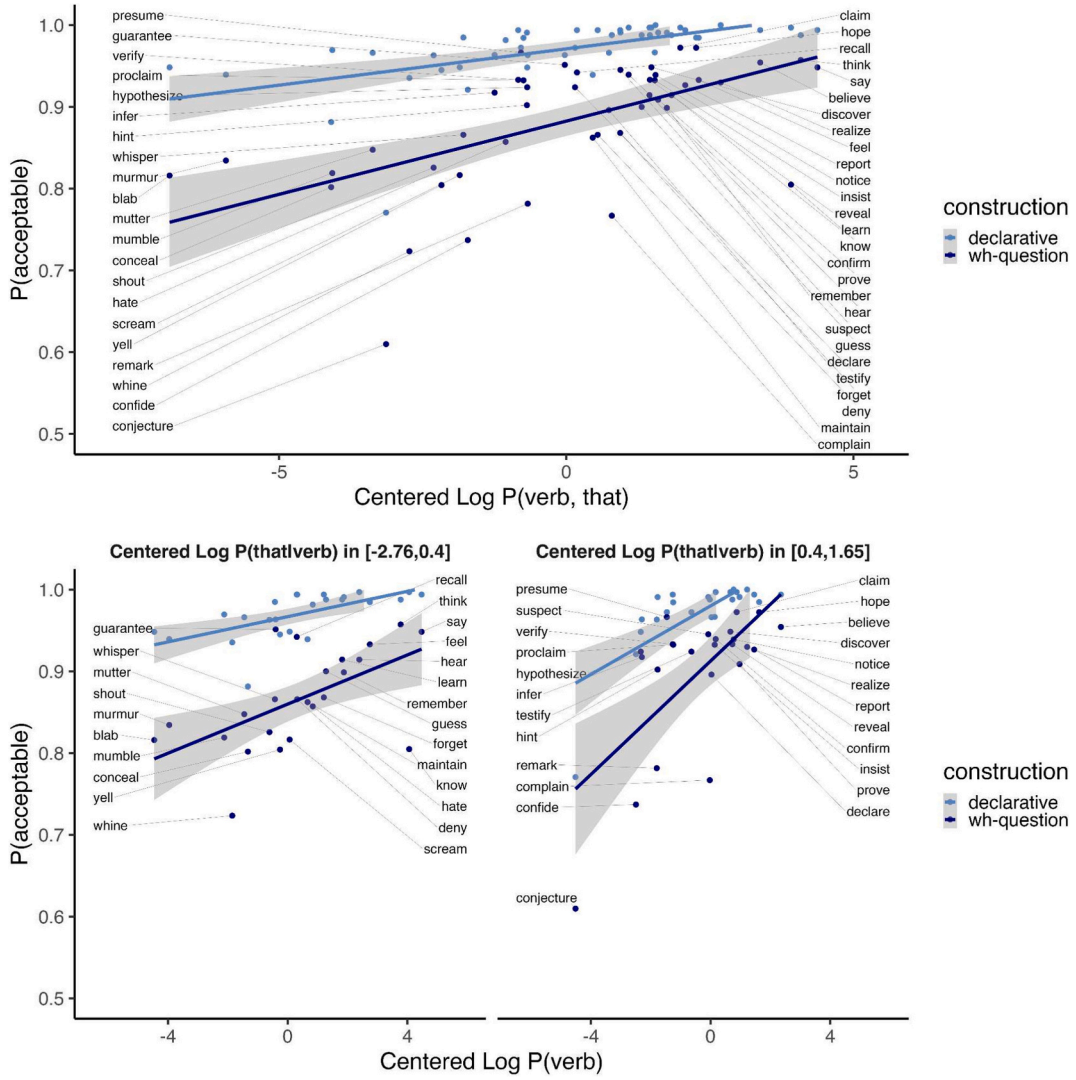

Fig. 2. Top panel: The original analysis from Liu, Ryskin, et al. (2022) involved predicting acceptability from the verb-frame probability, P(verb, that), the construction, and their interaction. (Note that in data were visualized on the log-odds scale, rather than on the probability scale. Their choice reflected more accurately the model representation of the data (a logistic model predicts log-odds), whereas the current choice reflects more accurately the raw data.) P(verb, that) is taken from the corpus analysis reported in the current work, which was syntactically parsed, and not the original quantity that Liu, Ryskin, et al. (2022) used (although they capture nearly identical quantities with a correlation of 0.96). Bottom panel: The same data were reanalyzed with different predictors. Like Liu, Ryskin, et al. (2022), the data are split by construction (color); however, the x-axis now represents the log P(verb). For purposes of visualization, the data are faceted by P(that | verb), such that verbs with lower P(that | verb) appear on the left, and verbs with higher P(that | verb) appear on the right.

concerned that participants may experience experimental fatigue. Thus, we did not include comprehension questions, which were used by Liu, Ryskin, et al. (2022) to exclude inattentive participants. In our experience, Prolific participants rarely fail attention checks, when such checks are included. (Even if participants were inattentive, they would not introduce bias into the results by rating differently verbs that varied in P (verb) and P(that | verb)). Hence, we divided the materials into three lists, such that each participant was exposed to 192 target items (each verb appeared twice in each of the two constructions).

4.1.2. Participants

We recruited 180 participants on Prolific, filtering for participants who indicated English to be their first language, and we set no exclusion criteria. Thus, data could include bilinguals, the IP address was not restricted to the US (a feature which was necessary in the context of MTurk, but is not necessary with Prolific, which is more reliable), and people of different ages and social groups. The data of 11 participants failed to record, and, of the remaining participants, 15 viewed the experiment more than once, so we excluded all trials over 192 (the intended number of trials). All participants stated on the experiment questionnaire that their first language is English. Unlike in Liu, Ryskin, et al. (2022), asking a specific question like "Is English your first language?" may be more accurate than relying on the term "native speaker," the meaning of which may vary across individuals (Cheng et al., 2021). The total number of participants in the analysis was 169.

4.1.3. Procedure

Participants read a consent statement, responded to whether English was their first language and where they were born. Then, participants were presented with the experimental materials in randomized order, where each sentence was followed by a prompt to rate the sentence on a binary scale: natural/unnatural. Participants were not given any additional instructions.

4.2. Results

The data are visualized in Fig. 3. We fit the same models as in the reanalysis of Liu, Ryskin, et al. (2022), with one difference: the model with P(verb) and P(that | verb) as independent predictors converged with the full random effects structure, so this time correlations between the random effects of participants were included in the models. As before, the full model proposed herein outperformed the Liu, Ryskin, et al. (2022) model that only used P(verb, that): it had lower AIC (24,145 vs. 24,429), lower BIC (24,539 vs. 24,572) and was significantly better based on a likelihood ratio test (Chi-Squared = 343.92, DF = 30, $p < .001$).

The new model had an intercept of 2.038. As before, higher P(verb) was associated with significantly higher acceptability (estimate = 0.443, std. error = 0.050, z-value = 8.967, p < .001). Higher P(that | verb) was also significantly associated with more acceptable ratings (estimate = 2.283, std. error = 0.097, z-value = 2.900, $p = .004$). Wh-questions were significantly less likely to be rated as acceptable (estimate = −2.088, std. error = 0.160, z-value = −13.083, p < .001). There was also a superadditive effect of P(verb, that) on acceptability, such that P(that | verb) had a greater positive influence on acceptability for verbs with higher P(verb) (estimate = 1.904, std. error = 0.044, z-value = 4.338, p < .001). Unlike in the reanalysis of Liu, Ryskin, et al. (2022), the model detected a significant interaction between P(verb) and construction, such that the effect of P(verb) was more positive for declaratives than for wh-questions (estimate = 0.192, std. error = 0.055, z-value = 3.468, p < .001). The rest of the interactions were not significant.

4.3. Discussion

The results of the replication are similar to those of our reanalysis of Experiment 2 by Liu, Ryskin, et al. (2022). According to a likelihood ratio test, we found that the proposed model, with P(verb) and P(that | verb) as independent factors, was significantly better than the one that used only verb-frame probability.

Like in the reanalysis of Liu, Ryskin, et al. (2022), we found that P(verb), P(that | verb), and their interaction (the superadditive effect of P(verb, that)) significantly predict an increase in acceptability, and that interrogatives are less acceptable than declaratives. However, there are also some minor differences between the original study and the replication. First, the baseline acceptability rating was much lower for the replication than for the original study (see below for potential explanations). Second, the model detected a significant interaction between verb probability and construction, which is not visually apparent in Fig. 3. This interaction is not a priori predicted by any theory that we are aware of, and is not directly relevant to the current investigation. For example, a syntactic account may predict that wh-questions are somehow categorically different from declaratives and therefore show a sharp preference for probable over improbable verbs. However, this is not the case: the model detected an interaction such that the effect of probability is *smaller* for wh-questions relative to declaratives.

One clear difference between the data in Liu, Ryskin, et al. (2022) and the replication here is the intercept: on average, judgments in Liu, Ryskin, et al. (2022) were much more likely to lean toward "acceptable" than in the current dataset. This is not theoretically relevant to the current project, which is focused on the effects of probability-based measures on the acceptability of specific constructions, but we still consider possible explanations. Potentially, this discrepancy was caused by the experimental context: the replication included simpler filler materials, which are predicted to be more acceptable, as they are simpler (V2 is intransitive, resulting in shorter declaratives and interrogatives without long-distance extractions). If the filler items are more acceptable than the target items, then it makes sense that the target items would be more likely to be judged as not acceptable in the presence of filler items. However, this was not the case: if anything, filler items were rated as acceptable slightly less frequently (filler items were rated as acceptable 74.9 % of the time, while target items were rated as acceptable 77.3 % of the time). Another potential explanation is some difference in the population, since participants in Liu, Ryskin, et al. (2022) were recruited via MTurk and in the replication via Prolific; however, we will avoid hypothesizing what these differences might be because that is an empirical question that can be tested by those who are interested in comparing the populations of these two platforms. In any case, the acceptability baseline is important to note for understanding the data at hand, but it is not theoretically informative in this case.

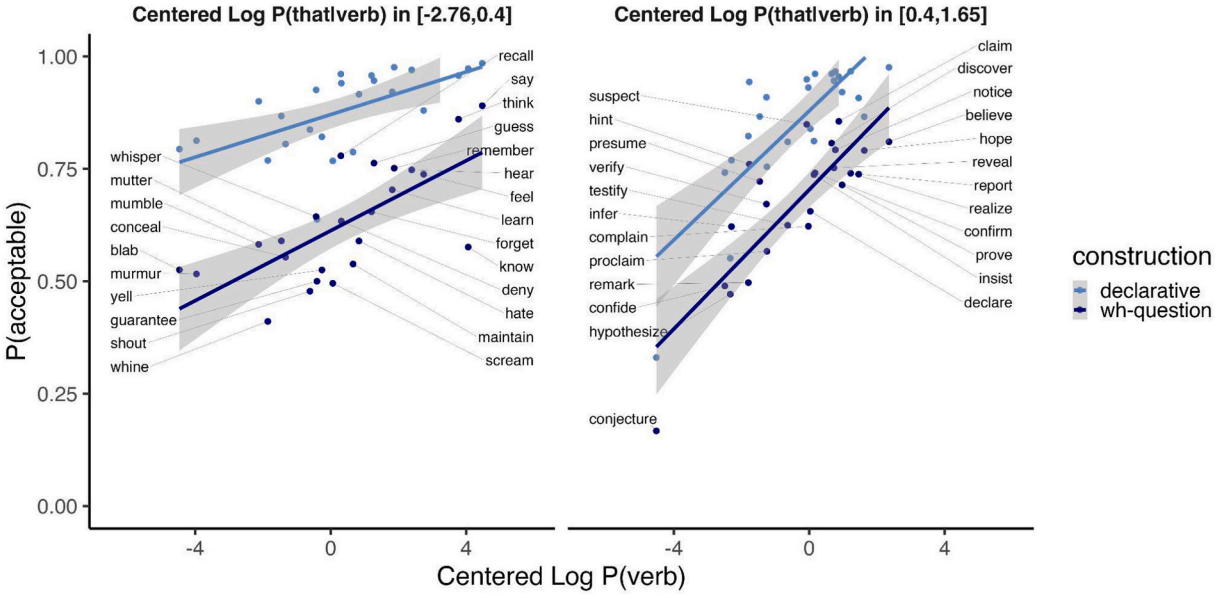

Fig. 3. The mean probability of sentences being rated as acceptable is represented above as explained by construction, P(verb), and P(that | verb). For purposes of visualization, the data are faceted by P(that | verb), such that verbs with lower P(that | verb) appear on the left, and verbs with higher P(that | verb) appear on the right.

5. Extension to adjective-frames

Liu, Ryskin, et al. (2022) proposed that the acceptability of declaratives and wh-questions with long-distance dependencies across sentence-complement verb depends on P(verb, that), or verb-frame probability, and we proposed a reanalysis where it depends on the factors of P(verb, that): verb probability—P(verb)—and the probability that the verb will take a sentential clause as an argument with the complementizer *that*—(P(that | verb)). However, if probabilities strongly predict acceptability, there is no reason to think that this effect should be confined to only sentences with verb predicates. Therefore, we test the same hypothesis—that the acceptability of declaratives and wh-questions with long-distance dependencies across sentence-complement verbs depends on the probability of the lexical item and its probability of having a clause as an argument—with regard to adjectives. For example, we will evaluate the prediction that the acceptability of a wh-question like "What was Mary glad that John bought?" depends on the probability of the adjective *glad* and the probability of *glad* having a *that*-clause as an argument.

5.1. Adjective frames—corpus analysis

We conducted a corpus analysis to retrieve the lexical probability, P (adjective), of 48 adjectives and how likely they are to be completed with a clause with *that* as a complementizer, P(that | adjective). To achieve this, we followed a similar procedure to the corpus analysis of verbs, described above. We selected a list of adjectives that represent a range of frame probabilities and emotional states.

5.1.1. Method

As for verbs, we used the Corpus of Contemporary American English (CoCA; Davies, 2008) and extracted all the sentences from the corpus that contained any of the adjectives listed below. For adjectives that could also function as verbs, we extracted sentences that contained either the past participle (e.g., *surprised*) or the present participle (e.g., *surprising*). From these extracted sentences, 30,000 sentences were selected randomly, and their dependency structure was parsed using the Stanza library in Python (Qi et al., 2020). We counted the appearance of the word in the sentence toward the adjective frequency if the parser produced any of the following results: the word's part of speech was *ADJ* (adjective), the part of speech was a verb in past participle and passive voice (e.g., surprised), or its dependency relation was *amod* (adjective modifier). We counted the sentence as one where the adjective takes a clausal complement with the complementizer *that* if the key word was identified as an adjective according to the criteria above, and it was the head of a word that had a *ccomp* or *csubj* dependency relation to it, and the word *that* was a dependent (or a dependent of a dependent) of the adjective. Again, we computed what frequency we should expect for both variables in the corpus by dividing the obtained counts by the number of sampled sentences and then multiplying this proportion by the total number of sentences in the corpus that contain the key word. As before, we obtained P(that | adjective) by dividing the count of sentences where the adjective was completed with a clause with the complementizer *that* by the adjective frequency, and P(adjective) and P (adjective, that) by dividing the count by the corpus size and corpus size minus 1, respectively. As before, we log-transformed and centered all the probabilities.

List of adjectives:

glad, sad, happy, mad, excited, angry, surprised, shocked, grateful, amused, ashamed, annoyed, irritated, sore, furious, bitter, elated, livid, ecstatic, flabbergasted, flustered, irate, aware, convinced, positive,
confident, certain, sure, doubtful, concerned, worried, satisfied, dubious,
anxious, optimistic, content, hopeful, relieved, jubilant, jealous, impressed,
disappointed, distraught, fearful, indignant, enraged, sorry, proud

5.1.2. Results

The probabilities of P(adjective) and P(that | adjective) for each word are visualized in Fig. 4. The raw counts for adjectives ranged from 515 to 338,913 with a median of 11,304. P(that | adjective) ranged from 0.003 to 0.349, with a median of 0.035. The correlation between P (adjective) and P(that | adjective) was negligible, at 0.112.

5.1.3. Discussion

Like in the case of verbs, the adjective that we examined varied in P (adjective) and in P(that | adjective). Again, we found that the correlation between these two variables was negligible, making it a meaningful question whether the two quantities have a separable effect on predictability.

5.2. Adjective frames—experiment

To evaluate how P(adjective), P(that | adjective), and their interaction predict acceptability, we conducted an experiment analogous to the replication of Experiment 2 from Liu, Ryskin, et al. (2022).

5.2.1. Method

5.2.1.1. Materials. The materials in the experiment followed the pattern below

What was {Name 1} {ADJ} that {Name 2} {V}?
ex. What was John glad that Mary bought?
{Name 1} was {ADJ} that {Name 2} {V} something.
ex. John was glad that Mary bought something.

Name 1, Name 2, and V were drawn from the same set of common names and V2 verbs as in Liu, Ryskin, et al. (2022). ADJ was drawn from the set of 48 adjectives above. As in the replication experiment, this resulted in 12 appearances for each adjective (each adjective with 6 unique embedded verbs in each of the two constructions), which resulted in 576 critical stimuli. We generated 576 filler stimuli as well, where V was an intransitive verb, resulting in simpler declaratives and yes/no interrogatives without long-distance extractions. As before, the stimuli were split into 3 lists of equal length, and each participant was exposed to only one of these lists.

5.2.1.2. Participants. As in the replication experiment, we recruited 180 participants on the crowd-sourcing platform Prolific, filtering for participants who indicated English to be their first language, and we set no exclusion criteria. Data from 9 participants failed to record. Of the remaining 171 participants, 7 participants viewed the experiment more than once, so we discarded all trials after the last intended trial (#192). Three participants indicated on the experimental questionnaire that their first language was not English (in spite of the Prolific filter), and thus were excluded, leaving a total of 168 participants.

5.2.2. Results

The data are visualized in Fig. 5. We fit the same models as in the replication experiment, except that again we dropped the term for correlations between random effects within participants due to convergence issues. Unlike with verbs, the model with P(adjective) and P(that | adjective) as independent predictors had similar likelihood to the model that used P(adjective, that) only, despite being substantially more complex: it had higher AIC (23,440 vs. 23,429), higher BIC (23,666 vs. 23,571), and was not significantly better based on a likelihood ratio test (Chi-Squared = 8.812, DF = 10, p = .55). As before, we continue interpreting the full model.

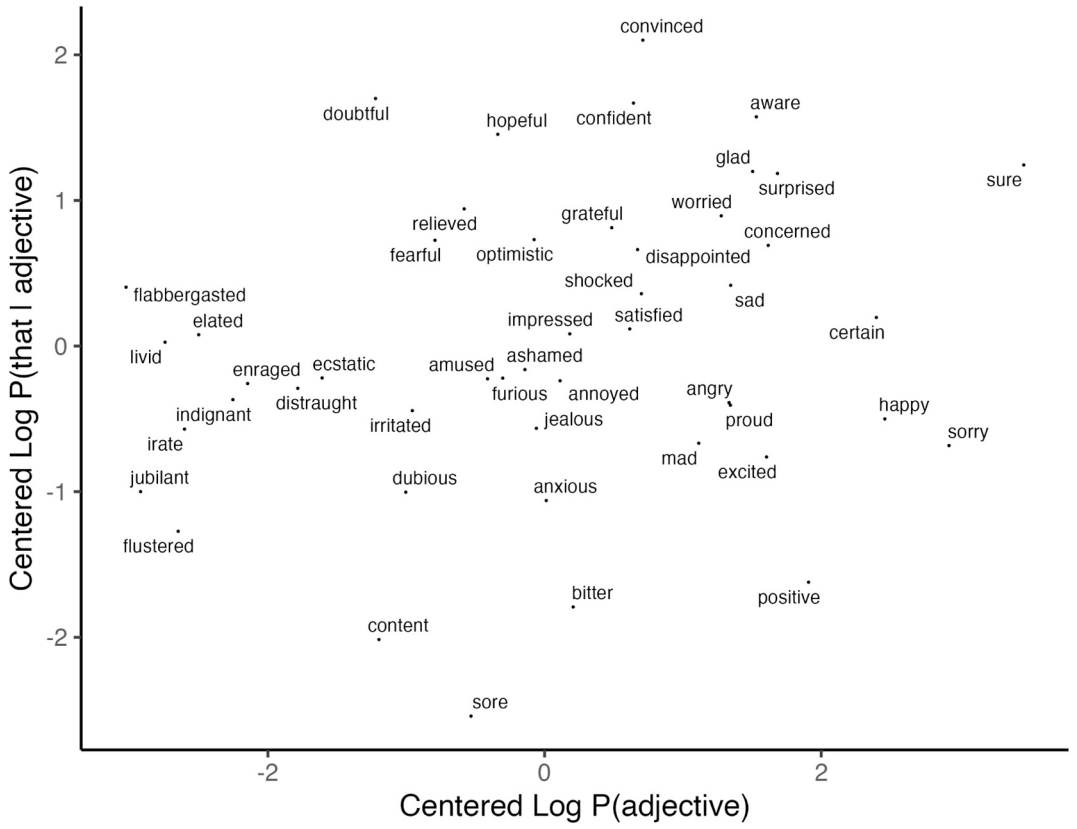

Fig. 4. Adjectives varied in their probability—P(adjective)—and the probability that the adjective will take a clause complement with the complementizer *that*—P (that | adjective).

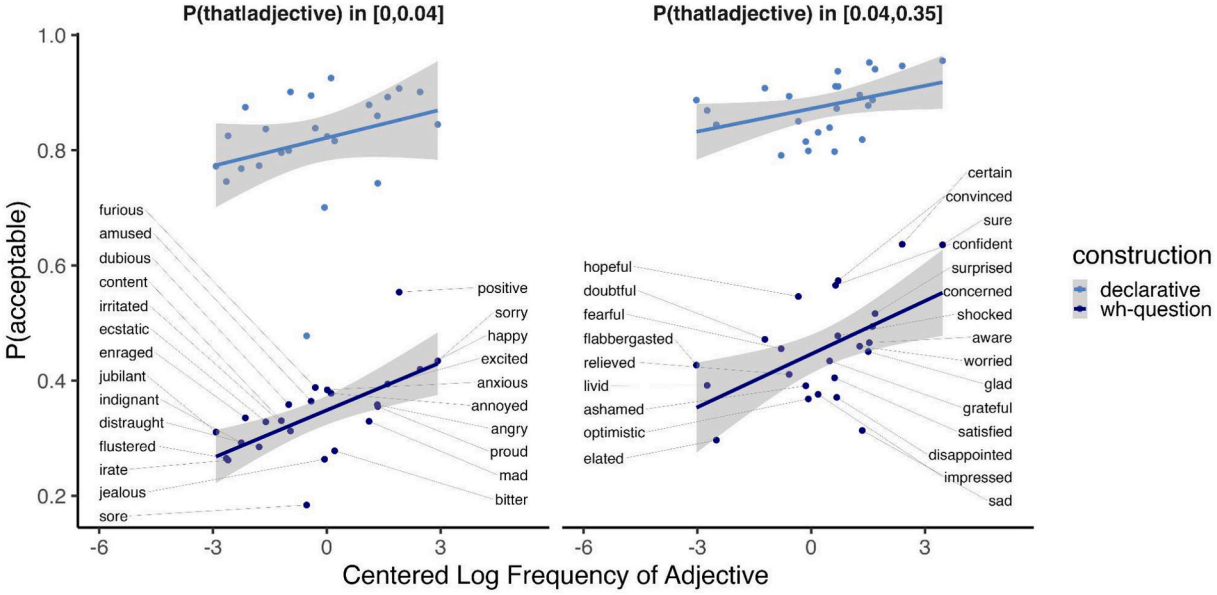

Fig. 5. The mean probability of sentences being rated as acceptable is represented above as explained by construction, P(adjective), and P(that | adjective). For purposes of visualization, the data are faceted by P(that | adjective), such that verbs with lower P(that | adjective) appear on the left, and verbs with higher P(that | adjective) appear on the right.

We proceed to interpret the new model, with P(adjective) and P(that | adjective) as independent predictors. The intercept of the model was 1.032. Higher P(adjective) was associated with a significantly higher chance of the sentence being judged as acceptable (estimate = 0.183, std. error = 0.046, z-value = 3.978, $p < .001$). Higher P(that | adjective) was also associated with a higher chance of judging a sentence as acceptable (estimate = 0.442, std. error = 0.075, z-value = 5.862, $p < .001$). Wh-questions were associated with a lower chance of being judged as acceptable (estimate = −3.793, std. error = 0.250, z-value = −15.198, $p < .001$). There were no significant interactions.

5.2.3. Discussion

Similar to the case of verbs, we find that the acceptability of interrogatives with long-distance extractions and parallel declaratives is strongly predicted by the probability of the underlying constructions: P (adjective), P(that | adjective), and the construction type (declarative vs. wh-question), all in line with the findings regarding verb predicates. Thus, we find evidence that, regardless of the type of the predicate (verbal or adjectival), the unigram probability of the lexeme and the probability that the lexeme takes a sentence complement are both predictive of acceptability.

The findings for adjectives diverged from those for verbs in some minor ways. Unlike with verbs, the model with adjective-frame probability alone—P(adjective, that)—was just as good at explaining the data as the full model that included the independent factors of P(adjective) and P(that | adjective). At the same time, both main effects were significant and not their interaction (whereas the interaction was significant for verbs). Thus, while the full model shows that the underlying variables that predict acceptability are the individual quantities P(adjective) and P(that | adjective), their product—the adjective-frame probability, P(adjective, that)—is a sufficient summary. We also note that the intercept (grand mean log odds of judging a sentence as acceptable) was lower for the adjective data relative to the verb data, potentially reflecting the fact that adjective-frames are less probable than verb-frames.

6. Adjective frames—replication

Due to reviewers' concerns about data quality (missing and repeating observations) in the adjective-frame experiment (Section 5.2), we replicated the adjective experiment (Section 5). The method for this section was identical, except for the details of the participants.

6.1. Method - participants

We recruited 184 participants on Prolific (we aimed for 180, but 4 additional participants completed the study despite the limit that we set on Prolific). We used a filter that is newly available on Prolific and recruited participants who have indicated that they are English-speaking monolinguals. We also only recruited participants from the US. On the experimental questionnaire, all participants were required to complete a CAPTCHA, after which they were asked again if English was their first language (all participants replied positively). Thus, the total number of participants for this analysis is 184.

6.2. Results

The mean acceptability of various adjectives as a function of their unigram probability P(adjective) and probability of taking a sentence complement P(that | adjective) is visualized in Fig. 6. The results patterned identically to Section 5.2.2. The proposed model, which includes P(adjective) and P(that | adjective) as independent measures, relative to the model inspired by Liu, Ryskin, et al. (2022), which used P (adjective, that) as the only probability measure, had higher AIC (28,478, 28452), BIC (28,707, 28596), and was not significantly better according to a likelihood-ratio test (Chi-squared = 0, DF = 10, $p = 1$). In the proposed model, the intercept was 0.685, P(adjective) had a significant positive effect on acceptability (estimate = 0.170, std. error = 0.033, z-value = 5.117, $p < .001$), P(that | adjective) had a significant positive effect on acceptability (estimate = 0.318, std. error = 0.053, z-value = 5.974, $p < .001$), and wh-questions were less acceptable than declaratives (estimate = −3.480, std. error = 0.221, z-value = 15.721, $p < .001$). There were no significant interactions.

6.3. Discussion

The results of the replication of the adjective experiment were qualitatively identical to those of the original. As before, we find that both P(adjective) and P(that | adjective) independently affect the acceptability of sentences, with an additional main effect of construction type (wh-questions vs. declaratives). As before, this model had the same fit as a model parallel to Liu, Ryskin, et al. (2022), using P(adjective, that), which leads us to conclude that P(adjective) and P(that | adjective) are independent predictors of acceptability, and that P(that, adjective) is sufficient summary of these measures.

7. General discussion

The current project investigated the role of probability in the

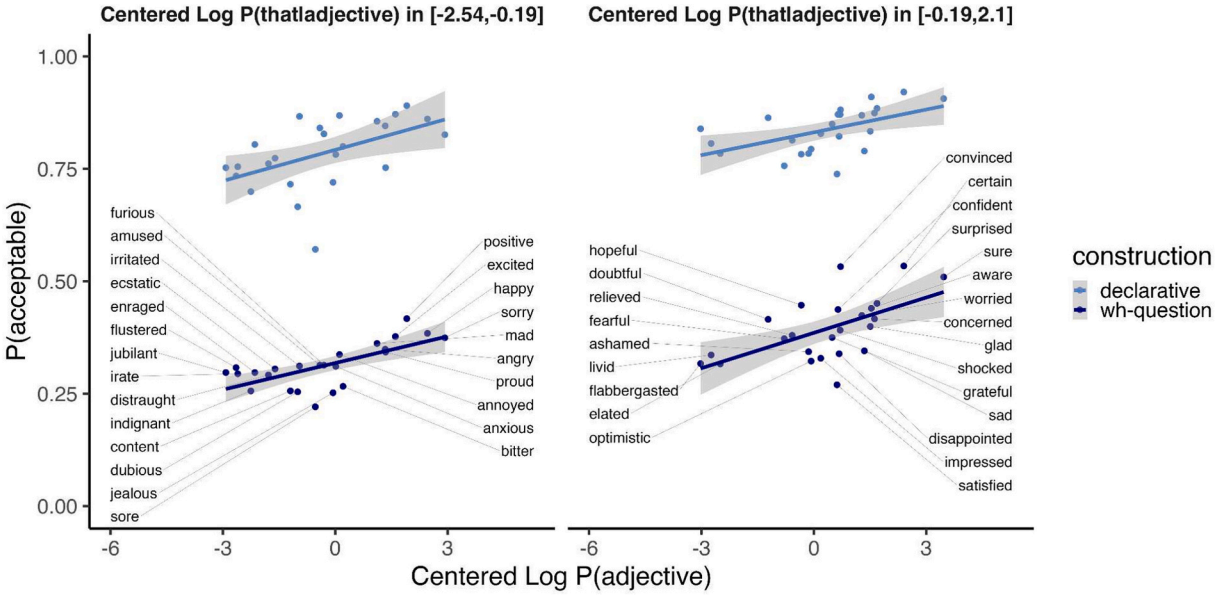

Fig. 6. The mean probability of sentences being rated as acceptable is represented above as explained by construction, P(adjective), and P(that | adjective). For purposes of visualization, the data are faceted by P(that | adjective), such that verbs with lower P(that | adjective) appear on the left, and verbs with higher P(that | adjective) appear on the right.

acceptability of wh-questions with a dependency across sentence-complement verbs (e.g., "What did Mary notice that John bought?"). Liu, Ryskin, et al. (2022) proposed an exposure probability-based account of the variation in the acceptability of such sentences: the more probable the verb-frame (e.g., "noticed that"), the more acceptable the sentence is. While Liu, Ryskin, et al. (2022) start with investigating the verb-frame probability P(verb, that), this should not have been the initial probability-based explanation to explore. The verb-frame probability is the product of two simpler quantities, according to the chain rule: the probability of the—verb P(verb)—and the probability within that verb of an argument structure that allows for a sentence complement—P(that | verb) (see Eq. (1)). It is possible that these probabilities play a role on top of verb-frame probability, as past work has shown that acceptability depends on the probabilities of lexemes and constructions, like a single word, an argument structure, or word sequences (Ambridge, 2013; Bresnan, 2007; Bresnan & Ford, 2010; Bybee & Eddington, 2006; White & Rawlins, 2020).

The current work presents strong evidence that the probabilities of words and their constructions predict the acceptability of wh-questions with a dependency across sentence-complement verbs and comparable declarative sentences. We conducted a corpus analysis of verbs and adjectives on a large parsed corpus. We then reanalyzed the data from Experiment 2 of Liu, Ryskin, et al. (2022), finding that all the quantities of interest—P(verb), P(that | verb), and their interaction, the superadditive effect P(verb, that)—significantly predicted acceptability. As in Liu, Ryskin, et al. (2022), there was a significant and large effect of construction, such that wh-questions were less acceptable than declarative sentences, and no significant interactions between the construction and either of the probability measures (except for one: in the verb replication experiment, where P(verb) had a smaller positive effect on wh-questions than on declaratives). We conducted a replication of this study, accounting for potential concerns in the original design and arriving at very similar results. We then conducted an identically designed study, substituting adjective-frames for verb-frames (e.g., "What was Mary glad than John bought?"). The results were largely similar, although there was no significant effect of adjective-frame probability, P(adjective, that) above and beyond P(adjective) and P(that | adjective). We then replicated the adjective experiment, arriving at the same findings. All of these together suggest that the probabilities of words and constructions together predict the acceptability of wh-questions with a dependency across sentence-complement predicates (verbs or adjectives).

The experimental results of Liu, Ryskin, et al. (2022) replicated despite changes to the methods and using a novel, syntactically-parsed corpus analysis. Whereas Liu, Ryskin, et al. (2022) arrived at P(verb, that) by querying for bigrams on the Google n-Gram viewer, the current project quantified verb-frame probabilities by syntactically parsing the Corpus of Contemporary American English. A priori, these two quantities could have diverged: the corpora are substantially different, and bigram probability may have misestimated P(verb, that) because they do not capture the syntactic relationship between words in the sentence. For example, bigram probabilities would miss all instances where the verb and the complementizer are not immediately neighboring, thus deflating the true verb-frame probability. However, this was not the case: the verb-frame probability derived by Liu, Ryskin, et al. (2022) had a 0.96 correlation with the verb-frame probability that we derived from the syntactic parse of the Corpus of Contemporary American English. Moreover, this shows that simple and quick corpus searches like Google n-Gram viewer can be highly reliable and informative. Experimentally, the replication of Experiment 2 of Liu, Ryskin, et al. (2022) involved (1) changes to several critical items (due to ambiguity in the original materials); (2) the inclusion of filler items without a dependency across the sentence-complement verb; and (3) the removal of comprehension questions (originally intended to catch inattentive participants). Despite all these changes, the results of the replication are remarkably similar—visually and inferentially—to those of the original experiment. The major difference between the two, which is not meaningful theoretically, was that participants were more likely to rate items as unnatural in the replication across the board. This indicates that the original findings are robust to variations like the ones that were introduced in the replication and that they are replicable.

The results from the acceptability study investigating wh-questions with dependencies across adjectival predicates were very similar to the results from the experiments with verbal materials, but the inferential results were not entirely identical. As with verbs, it was the case that both the P(adjective) and P(that | adjective) affected acceptability. However, unlike in the experiments with verbal predicates, the interaction of these quantities—the superadditive effect of adjective-frame probability, P(adjective, that)—was not a significant predictor of acceptability. We are not aware of theoretical explanations for this difference. Possible data-driven explanations could be that adjectives were overall less probable and less likely to take a sentence complement relative to verbs (that is, P(that | adjective) was generally lower than P(that | verb)). Also, P(that | adjective) had less variability, with almost all values being similarly low, relative to P(that | verb), which showed more variability (and see Figs. 1 & 4). This may have limited the extent to which this quantity affected acceptability judgments. The difference in probabilities of P(that | verb) and P(that | adjective) also would explain why the effect of construction was bigger in the experiment with the adjectival predicates than in the experiments with the verbal predicates: verbal predicates are more likely than adjectival predicates to take a sentence complement.

Another difference that emerged between the verbal and adjectival experiments is that the model that included P(adjective) and P(that | adjective) as independent predictors had only a marginally higher log-likelihood, and higher AIC and BIC, than the model that simply used P(adjective, that) alone. This indicates that the model with only P(adjective, that) predicted the data almost as well as the full model, despite being simpler. This may seem confusing given that P(adjective, that) is the product of P(adjective) and P(that | adjective), and that the interaction of P(adjective) and P(that | adjective) was not significant. This is possible because, although P(adjective, that) is not predicting acceptability above and beyond P(adjective) and P(that | adjective), it is correlated with both. Thus, in a model that contains only the adjective-frame probability P(adjective, that), this quantity may come out as significant, and the model may predict the data nearly as well as the model with the independent quantities. This does not influence our conclusions: the full model revealed that it is the two independent quantities of P(adjective) and P(that | adjective) that are significant predictors of acceptability, and not their superadditive effect. Practically, this suggests that P(adjective, that) is a sufficient summary of P(adjective) and P(that | adjective) for the purposes of predicting acceptability.

7.1. Theories of the acceptability of structures with long-distance dependencies

Chomsky (1973) brought the study of the acceptability of structures with long-distance dependencies into mainstream linguistics. He took observations about the acceptability of complex English structures from Ross (1967) together with his own observations, and he proposed a simple syntactic account, such that transformations from a deep structure form could not cross more than two bounding nodes (S and NP for English) when forming a surface structure. This theory accounted for many observations, across a variety of constructions, in what seemed like a simple way at the time. Given his hypothesis, he also proposed that aspects of these structures must be part of an innate Universal Grammar (UG), because it was difficult to see how these constraints on the transformations could be learned. This is a variant of the "poverty of the stimulus argument": there isn't enough evidence in the typical input to a child in order to learn the target rules (Chomsky, 1971). Other researchers then tested and explored similar and different kinds of

hypotheses in order to test Chomsky's claims.

At the current point, it has become clear that Chomsky's ambitious syntax-based explanation for the acceptability of structures with long-distance dependencies is not successful in several ways. First, despite its initial simplicity, these kinds of proposals needed to be much more complex in order to account for simple phenomena, such as the acceptability of materials like **1d** repeated here:

(1d) What did Mary say that Bill kicked?

Furthermore, when the details of the syntactic hypothesis were spelled out, many counterexamples emerged (see Liu, Winckel, et al., 2022 for a summary of some of the general types). As discussed in the introduction, syntax-based theories had been initially extended to account for the differences in acceptability in the kinds of materials that we examine in the current paper, involving long-distance dependencies across sentence-complement verbs like "say", "hate", or "murmur". The proposal was that, for less acceptable examples, there is an additional empty noun phrase that needs to be crossed in the transformation from deep structure to surface structure. But as observed earlier, there is no independent evidence for such a claim. Furthermore, there appears to be a continuum of acceptability from less acceptable to more acceptable, dependent on the probability factors that have been observed and discussed in the current paper. These observations are not easily accounted for within a syntax-based proposal.

At the moment, there doesn't seem to be a syntax-only component to an explanation of the acceptability of structures with long-distance dependencies (Liu, Winckel, et al., 2022). One kind of case that appears to be semantic, is the fronting of full conjuncts, as in **3**:

3

3a. *Who did you invite _ and Lana?

3b. *Who did you invite Lana and _?

3c. *Who did you invite _ and _?

It isn't possible to front one or more full conjuncts, as in **(3)** (Liu, Winckel, et al., 2022). For example, it is not possible to front the first conjunct (**3a**); the second conjunct (**3b**) or both (**3c**). Researchers explain these phenomena in terms of what is termed the "conjunct constraint" (Sag, 2010). In an analysis without transformations, the definition of coordination necessarily implies (at least) two conjuncts. This accounts for the unacceptability of examples in (**3**). The coordination in examples **3a** and **3b** has only one conjunct, and in example **3c** it has no conjunct at all (Sag, 2010; Chaves, 2012).

Going beyond the word-probability cases in the current paper, and the semantics-based examples in (**2**), the most promising current theories of the acceptability of structures with long-distance dependencies are based on discourse properties (Abeillé et al., 2020; Ambridge, 2013; Cuneo & Goldberg, 2023; Deane, 1991; Erteschik-Shir, 1973, 1998; Erteschik-Shir & Lappin, 1979; Goldberg, 2006; Van Valin Jr, 1998; Van Valin & LaPolla, 1997; Winckel et al., 2025). Recently, Cuneo and Goldberg (2023) investigated the idea that the unacceptability of long-distance extractions stems from a conflict in information structure: the extracted element at the beginning of the sentence is usually at-issue; however, when this element is extracted from a position where it is backgrounded (e.g., from a relative clause), the result is infelicitous and results in low acceptability. In an acceptability study of a large set of controlled examples, they found that the more backgrounded an element was—as measured by a negation task and a discourse task—the less acceptable the sentence became when that element was fronted.

Whereas this kind of account is based on meaning, the account proposed by Liu, Ryskin, et al. (2022) and elaborated in the current paper is based on construction probabilities. Because these accounts use potentially independent mechanisms in their explanations, the relationship between these accounts can take any form: they could be true at the same time, or one may explain away the other. We leave the comparison of these theories to future research: it is not the focus of the current study.

7.2. Concluding remarks

Previously, Liu, Ryskin, et al. (2022) showed that verb-frame probability predicted the acceptability of wh-questions with long-distance dependencies across sentence-complement verbs. In the current work, we broadened this account to one that involves simpler probability-based quantities: the lexeme probability of the verb and the probability that it takes a sentence complement. We evaluated this account on the original dataset from Liu, Ryskin, et al. (2022) and found that it explains the data better than the original proposal. We then replicated their experiment and found the same results. Finally, we extended the materials to adjective frames in 2 experiments (an original study and replication), and again found similar results. In sum, this work shows that much of the variability in the acceptability of wh-questions that have been traditionally considered "islands" is explained by probability-based quantities.

The current work merges insights on language processing from construction-based (e.g., Bybee, 2006, 2010; Croft, 2010; Fillmore, 1988; Goldberg, 1995, 2006, 2019; Steels, 2011, 2013) and probability-based approaches (Hale, 2001; Jurafsky, 1996; Levy, 2008; Spivey & Tanenhaus, 1998; Tanenhaus et al., 1995). It provides evidence in favor of these approaches because participants seem to be sensitive to the probabilities of words, sequences of words, and even argument structure, which is a more abstract quantity. Not only do construction probabilities govern language processing, but they are also generalizable across languages and are fairly easy quantities to compute; the frequencies of words and their co-occurrences can be calculated from any corpus in any language. This contrasts with deriving latent, theory-based quantities. For example, to measure the frequency of a hypothetical quantity such as the probability of A-movement (Chomsky, 1993) one must not only syntactically annotate a corpus, which is more demanding than counting words and n-grams, but one must also develop a single framework that can be used to parse any kind of corpus and language. This added complexity inhibits cross-linguistic investigation and, consequently, advancements in cognitive science (Blasi et al., 2022; Henrich et al., 2010). Thus, construction probability is a powerful and potentially highly general tool for explaining acceptability.

CRediT authorship contribution statement

Moshe Poliak: Writing – review & editing, Visualization, Investigation, Formal analysis, Data curation. **Curtis Chen:** Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Edward Gibson:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Appendix A. Participants' behavior is not affected by the choice of labels on binary scales

All the experiments reported above queried for acceptability judgments using a binary scale with the labels *Natural* and *Unnatural*. To pressure-test the robustness of the findings in the main text, we conducted a series of 4 experiments on the same set of materials that varied the labels on binary scales: Natural/Unnatural, Good/Bad, Acceptable/Unacceptable, Grammatical/Ungrammatical.

A.1. Method

Materials. Twenty-four grammatical sentences, ranging in length from 10 to 20 words, were randomly selected from the Universal Dependencies Treebank (Nivre et al., 2020). At random, the sentences were grouped into 4 groups of 6 sentences. Group 1 remained untouched; In group 2, one pair of adjacent words was randomly exchanged; in group 3, three pairs of adjacent words were randomly exchanged; in group 4, all words were shuffled. For each sentence, we computed the Damerau-Levenshtein distance (Levenshtein, 1966), which is the minimal number of deletions, insertions, substitutions, or exchanges of adjacent words that are needed to arrive from the corrupted sentence to the original sentence. We assume that a higher Damerau-Levenshtein distance should result in lower acceptability. All 24 sentences were presented to all participants in the same order.

Participants. Fifty participants were recruited from Prolific per experiment, for a total of 200 participants. On Prolific, we filtered for English monolinguals from the US, and on the survey platform we asked participants again whether English was their first language, and all responded affirmatively.

Procedure. All participants were redirected from Prolific to the same custom study platform as in the experiments in the main text. Participants were not given any instructions besides rating the sentences according to the prompt. The prompt that appeared with each sentence was identical within the experiment. The prompt "Rate how grammatical/good/natural/acceptable the sentence is" appeared below the sentences that participants were asked to rate, followed by two radio buttons that carried the positive label (on top) and negative label (below). Each experiment had only one set of labels, since this was a between-participant manipulation.

A.2. Results

The mean rating (where 1 is the positive label and 0 is the negative label) per item is visualized in Fig. 7. To test inferentially whether labels have an effect on participants' ratings, we fit 3 mixed-effects logistic regressions using the lme4 library (Bates et al., 2015): null, partial, and full. In the null model, response (1 = positive, 2 = negative) was predicted from a fixed intercept, with random intercepts for participants and items. In the partial model, we added a fixed effect of Damerau-Levenshtein distance and a random slope for Damerau-Levenshtein distance within participants. In the full model, we added a sum-coded variable for label (with four levels: good, grammatical, acceptable, natural), its interaction with Damerau-Levenshtein distance, and an additional random slope for label within item. In all 3 models, the random effects structure is the maximal random effects structure justified by the design (Barr et al., 2013). The null model will be used as reference. The partial model is able to account for variability in acceptability due to Damerau-Levenshtein distance, but unable to account for any variability introduced by manipulating the labels; the full model can also capture main effects of label (a baseline change in ratings, e.g., if in one pair of labels the positive option were chosen with higher probability than in the other pairs of labels, regardless of the sentence that was presented), as well as interactions (a differential effect of Damerau-Levenshtein distance on ratings based on label).

We conducted a likelihood-ratio test on the null model against the partial model, and another test on the partial model against the full model. The partial model was significantly better than the null model (Df = 3, Chi-squared = 175.79, $p < .001$), with lower (better) AIC (partial = 3319, null = 3489) and BIC (partial = 3358, null = 3508). In contrast, the full model was not significantly better than the partial model (Df = 15, Chi-squared = 8.23, $p = .914$), and had higher (worse) (partial = 3319, full = 3341) and BIC (partial = 3358, full = 3477). This suggests that accounting for the use of different labels does not improve model fit. Inspecting the full model, the intercept was 1.539, and, as predicted, there was a significant effect of Damerau-Levenshtein distance (estimate = −1.319, std. error = 0.202, z-value = −6.529, p < .001), suggesting that sentences with more corruptions were seen as less acceptable. However, all the main and interaction effects associated with label were not significant.

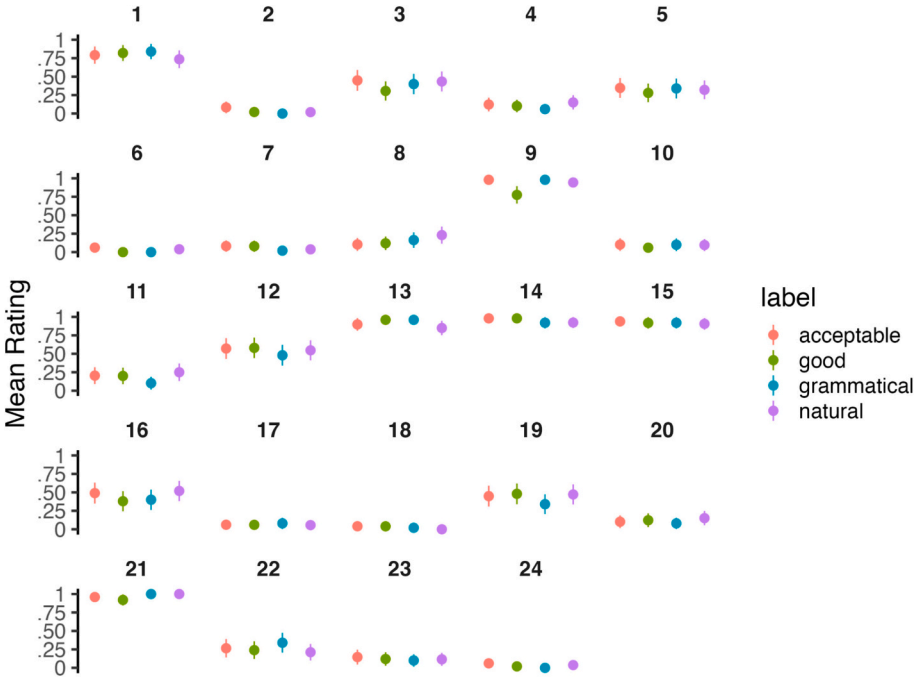

Fig. 7. The mean rating (1 = positive label, 0 = negative label) is plotted per item for each of the four label pairs: Acceptable / Unacceptable, Good / Bad, Grammatical / Ungrammatical, Natural / Unnatural. Error bars are 95 % Confidence Intervals.

A.3. Discussion

This study investigated whether the labels in acceptability rating tasks with binary scales have an effect on participants' behavior, comparing the labels Good/Bad, Acceptable/Unacceptable, Grammatical/Ungrammatical, Natural/Unnatural. The study showed no appreciable effect of label, such that, according to a likelihood-ratio test, the full model that accounted for labels was not significantly better than the partial model, which only accounted for the effect of Damerau-Levensthein distance on acceptability (i.e., number of corruptions). According to AIC and BIC, which penalize models for complexity, the full model was worse than the partial model. Finally, in the full model, not a single effect associated with label was found to be significant. Although absence of evidence is not evidence of absence, this suggests that the effect of labels is not big enough to affect participants' behavior in noticeable ways given a sample size of 50 participants per group.

Appendix B. Replication with a Likert scale

All the experiments reported above queried for acceptability judgments using a binary scale. However, during the review, concerns were raised that results might depend on the use of a binary scale rather than a Likert scale (which has multiple levels, rather than just 2). Sprouse et al. (2013) investigated how behavior and inference may be affected by different scale types (including binary scales and Likert scales) and found no meaningful difference. Moreover, Liu, Ryskin, et al. (2022) conducted Experiment 1, which contained a subset of the materials investigated in the current work in Sections 3–4, using a Likert scale, and arrived at the same results as the current investigation. Nonetheless, materials with adjective frames have not been investigated using a Likert scale. Therefore, we conduct a smaller-scale replication of the adjectival materials using a Likert scale and report them here.

7.3. Method

Completely identical to Section 6, except (1) 60 participants were recruited rather than 180, for economy, and (2) the current experiment used a Likert scale with 5 levels:

5-Natural.
4-Somewhat Natural.
3-Neither Natural nor Unnatural.
2-Somewhat Unnatural.
1-Unnatural.

All participants confirmed that English was their first language.

7.4. Results

The results are visualized in Fig. 8. We fit a cumulative mixed-effects regression with a logit linking function using the ordinal package in R (Christensen, 2023). Except for this, the model specifications were identical to those from Section 6, with one difference. The model with P(adjective) and P(that | adjective) as independent predictors did not converge with the full random effects structure, and therefore, we fit it without slopes for interactions between the predictors within participants. As in the case of binary acceptability judgments, the model detected a significant positive effect for P(adjective) (estimate = 0.121, std. error = 0.036, z-value = 3.353, $p < .001$), a significant positive effect for P(that | adjective) (estimate = 0.306, std. error = 0.060, z-value = 5.077, $p < .001$), and a significant negative effect of construction, such that wh-questions were less acceptable than declaratives (estimate = −2.162, std. error = 0.292, z-value = −7.389, $p < .001$). Unlike in the previous versions of the adjective studies, the model also detected a significant interaction between P(adjective) and construction, such that the effect of P(adjective) was smaller for wh-questions than for declaratives (estimate = −0.071, std. error = 0.031, z-value = −2.318, $p = .020$).

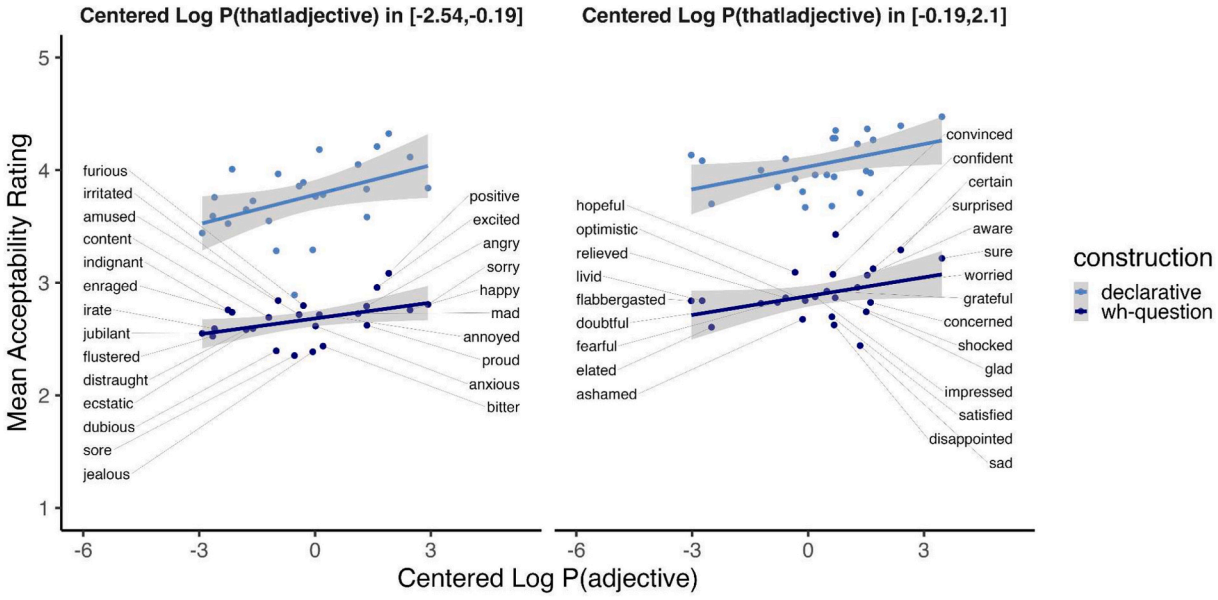

Fig. 8. The mean acceptability of sentences is represented above as explained by construction, P(adjective), and P(that | adjective). For purposes of visualization, the data are faceted by P(that | adjective), such that verbs with lower P(that | adjective) appear on the left, and verbs with higher P(that | adjective) appear on the right.

7.5. Discussion

In this experiment, we sought to investigate whether the effects that were detected for P(adjective) and P(that | adjective) on acceptability ratings depend on whether participants responded to a binary scale or a Likert scale. We conducted a replication of the experiments in Sections 5–6, with the only change being the use of a 5-level scale. Visually and inferentially, the results were nearly identical to those of Sections 5–6, except for a significant interaction effect between P(adjective) and construction, which also appeared in Section 4. As before, the interaction between P(lexeme) and construction is not a priori predicted by any theory that we are aware of, and is not directly relevant to the current investigation. For example, a syntactic account may predict that wh-questions are somehow categorically different from declaratives and therefore show a sharp preference for probable over improbable verbs. However, this is not the case: the model detected an interaction such that the effect of probability is *smaller* for wh-questions relative to declaratives. Moreover, this seems to be a potential side-effect of using a logit linking function, which is used in an ordinal regression: the same difference in probability becomes greater on the logit scale when it is closer to probabilities near 0 or 1. Since declaratives are more acceptable than wh-questions, the same effect of P(adjective) may appear bigger on the logit scale for declaratives than wh-questions, even if the effect on the probability scale is the same. We fit a model using a Gaussian regression (lmer) with the same specifications, and, indeed, this interaction was not significant anymore, suggesting that it was an artifact of using a logit linking function. Critically, the finding that the acceptability of wh-questions with long-distance dependencies across sentence-complement verbs depends on both P(lexeme) and P(that | lexeme) was replicated and remained across binary and Likert scales.

Funding

Funded by Nation Science Foundation award # BCS-2020840.

Data availability

We have shared the materials, data, and analyses on OSF and provided an anonymized link for reviewers.

References

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition, 204*, Article 104293.
Ambridge, B., & Goldberg, A. E. (2008). *The island status of clausal complements: Evidence in favor of an information structure explanation*.
Ambridge, B. (2013). How do children restrict their linguistic generalizations? An (un-) grammaticality judgment study. *Cognitive Science, 37*(3), 508–543.
Baltin, M. R. (1982). A landing site theory of movement rules. *Linguistic Inquiry, 13*(1), 1–38.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01
Blasi, Damián E., Henrich, Joseph, Adamou, Evangelia, Kemmerer, David, & Majid, Asifa (2022). Over-reliance on English hinders cognitive science. *Trends in cognitive sciences, 26*(12), 1153–1170.
Bresnan, J. (2007). Is knowledge of syntax probabilistic. In S. Featherston, & W. Sternefeld (Eds.), *Experiments with the English dative alternation. Roots: Linguistics in search of its Evidential Base (studies in generative grammar)* (pp. 75–96).
Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language, 86*(1), 168–213.
Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711–733.
Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.
Bybee, J., & Eddington, D. (2006). A usage-based approach to Spanish verbs of'becoming. *Language, 82*(2), 323–355.
Chaves, Rui P. (2012). Chaves On the grammar of extraction and coordination. *Natural Language & Linguistic Theory, 30*(2), 465–512.
Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology, 12*, Article 715843.
Chomsky, N. (1971). Problems of knowledge and freedom. *Philosophy, 48*(184).
Chomsky, N. (1973). Conditions on transformations. In S. Anderson, & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). Holt, Rinehart, & Winston.
Chomsky, N. (1977). On wh-movement. In P. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal syntax* (pp. 71–132). Academic Press.
Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.
Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures*. Walter de Gruyter.
Christensen, R. H. B. (2023). Ordinal—Regression models for ordinal data. https://CRAN.R-project.org/package=ordinal.
Croft, W. (2010). *Construction grammar* (pp. 463–508).
de Cuba, C. (2018). Manner-of-speaking that-complements as close apposition structures. *Proceedings of the Linguistic Society of America, 3*, 32–1.
Cuneo, N., & Goldberg, A. E. (2023). The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition, 240*, Article 105563.
Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language, 58*(4), 931–951.
Davies, M. (2008). *The corpus of contemporary American English (COCA): 560 million words, 1990-present*.
Deane, P. (1991). *Limits to attention: A cognitive theory of island phenomena*.
Erteschik-Shir, N. (1973). *On the nature of island constraints* [PhD Thesis]. Massachusetts Institute of Technology.
Erteschik-Shir, N. (1998). The syntax-focus structure interface. In *The limits of syntax* (pp. 211–240). Brill.
Erteschik-Shir, N., & Lappin, S. (1979). *Dominance and the functional explanation of island phenomena*.
Fillmore, C. J. (1988). The mechanisms of "construction grammar". In *Annual meeting of the Berkeley Linguistics Society* (pp. 35–55).
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press. http://catdir.loc.gov/catdir/toc/uchi051/94020705.html.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the North American chapter of the Association for Computational Linguistics*.
Henrich, Joseph, Heine, Steven J, & Norenzayan, Ara (2010). The weirdest people in the world? *Behavioral and brain sciences, 33*(2–3), 61–83.
Huang, C. J. (1982). *Move WH in a language without WH movement*.
Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*(2), 137–194. https://doi.org/10.1207/s15516709cog2002_1
Kiparsky, P., & Kiparsky, C. (1971). Fact. In M. Bierwisch, & K. Heidolph (Eds.), *Progress in linguistics* (pp. 143–173). Mouton.
Kothari, A. (2008). Frequency-based expectations and context influence bridge quality. In *Proceedings of WECOL* (p. 2008).
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006
Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2022). A verb-frame frequency account of constraints on long-distance dependencies in English. *Cognition, 222*, Article 104902. https://doi.org/10.1016/j.cognition.2021.104902
Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2022). Structural, functional, and processing perspectives on Linguistic Island effects. In *, Vol. 8. Annual review of linguistics* (pp. 495–525). Annual Reviews https://doi.org/10.1146/annurev-linguistics-011619-030319.
Nivre, J., De Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., … Zeman, D. (2020). *Universal dependencies v2: An evergrowing multilingual treebank collection*. arXiv preprint arXiv:2004.10643.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python natural language processing toolkit for many human languages*. arXiv preprint arXiv: 2003.07082.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Richter, S., & Chaves, R. (2020). *Investigating the role of verb frequency in Factive and Manner-Of-Speaking Islands*. CogSci.

Rizzi, L. (1990). *Relativized minimality*. The MIT Press.

Ross, J. R. (1967). *Constraints on variables in syntax*. Thesis. Massachusetts Institute of Technology https://dspace.mit.edu/handle/1721.1/15166.

Sag, I. A. (2010). English filler-gap constructions. *Language, 86*(3), 486–545. https://doi.org/10.1353/lan.2010.0002

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10), Article e2307876121.

Slowikowski, K., Schep, A., Hughes, S., Lukauskas, S., Irisson, J.-O., Kamvar, Z. N., … others. (2024). Package ggrepel. In *Automatically position non-overlapping text labels with ggplot2*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319.

Snyder, W. (1992). *Wh-extraction and the lexical representation of verbs*. Unpublished Manuscript. MIT.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1521–1543. https://doi.org/10.1037/0278-7393.24.6.1521

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua, 134*, 219–248.

Steels, L. (2011). *Design patterns in fluid construction grammar* (Vol. 11). John Benjamins Publishing.

Steels, L. (2013). *Fluid construction grammar*.

Stoica, I. (2016). Island effects and complementizer omission: The view from manner of speaking verbs. *Constructions of Identity*, 191–200.

Stowell, T. A. (1981). *Origins of phrase structure* [PhD Thesis]. Massachusetts Institute of Technology.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Van Valin Jr, R. D. (1998). Focus structure or abstract syntax? A role and reference grammar account of some abstract syntactic phenomena. In *2007-03-05. Http://wings.Buffalo.Edu/soc-sci/linguistics/research/rrg/vanvalin_papers/focabstr.Pdf*.

Van Valin, R. D., & LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press.

Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language, 87*(2), 249–273.

White, A., & Rawlins, K. (2020). Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: A Journal of General Linguistics, 5*(1), 105. https://doi.org/10.5334/gjgl.1001

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2024). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of Surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics, 11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2025). Discourse-based constraints on long-distance dependencies generalize across constructions in English and French. *Cognition, 254*, Article 105950.