A nearly-exhaustive experimental investigation of bridge effects in English

Nick Huang
National University of Singapore

Diogo Almeida
New York University Abu Dhabi

Jon Sprouse
New York University Abu Dhabi

September 2024

ABSTRACT

In many languages, finite clause-embedding verbs vary in whether they allow wh-dependencies to cross from the embedded to the matrix clause—a phenomenon we call 'bridge effects'. Why bridge effects exist has been the subject of much debate; we argue that contributing to the lack of consensus is the relatively small samples of verbs (12-75 for English) previously tested in the literature. To resolve this issue, we report two new data sets of bridge effects covering a nearly-exhaustive sample of 640 English verbs. We use these data sets to address three research questions: Are there bridge effects at all? How well do leading theories of bridge effects explain observed variation across the full range of verbs? And are there new patterns emerging from our data that could lead to a better theory? We ultimately argue in favor of a multivariate approach, drawing upon existing ideas while including a novel morphosyntactic licensing component identified from our data. We also discuss implications for theories of locality and explore how context might affect the acceptability of wh-dependencies.[*]

KEYWORDS: bridge verbs, wh-dependencies, locality constraints, sentence processing, pragmatics, English

[*] Our thanks to two anonymous reviewers and the editors for their constructive feedback, which has helped to improve the article. For comments and suggestions on earlier studies by the first author that have informed the current project, we are deeply grateful to Colin Phillips, Howard Lasnik, Jeffrey Lidz, Valentine Hacquard, Alexander Williams, and Robert DeKeyser. We would also like to thank Sandra Villata, Jessica Hsieh, Spencer Lim, Firdaus Moner, and Wong Zi Shu for their help with various aspects of data collection and preparation. All errors remain our own. This research was supported by faculty grants from New York University Abu Dhabi and the National University of Singapore, as well as funding from the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2021-FRC3-002).

1. INTRODUCTION. Wh-dependencies (also called filler-gap dependencies) appear to be unbounded by distance: a wh-item like *what* can be separated ('extracted') from its point of semantic interpretation, called its *gap* location by analogy to the position it would occupy in a declarative sentence, by any arbitrary distance, calculated either linearly in number of words or structurally in number of clauses. At the same time, there appear to be non-distance locality constraints on wh-dependencies. The one we focus on in this article is that wh-items can only originate within the complement clause of certain clause-embedding verbs, such as *say* and *think*, as in 1, which are commonly referred to as 'bridge verbs' (so named by Erteschik-Shir 1973; first observed by Dean 1967). Extraction from the complement of other verbs, such as *shout* in 2, is less acceptable. We will call these between-verb differences in acceptability 'bridge effects'.

(1) What did Jo think that Sam said that Kim saw _?

(2) ??What did Jo shout that Sam said that Kim saw _?

As one might expect from over 50 years of research, the space of theories of bridge effects is fairly robust. There are at least three prominent theories that can cover the full range of clause-embedding verbs, each positing a distinct source for bridge effects: information structure-based theories (e.g., Erteschik-Shir 1973, Goldberg 2006, Ambridge & Goldberg 2008, Richter & Chaves 2020), template-based processing theories (e.g., Dąbrowska 2008, 2013; see also Verhagen 2005, 2006), and frequency-based processing theories (e.g., Kothari 2008, Liu et al. 2022). We will review these in detail in Section 2. Each has amassed some amount of experimental evidence in its favor. For example, Ambridge and Goldberg (2008) report an exceptionally strong correlation between backgroundedness and bridge effects (Pearson r=-.83, p=.001), which supports information structure approaches; Liu and colleagues (2022) present evidence suggesting there are no bridge effects at all (i.e., no differences in acceptability among verbs) once the frequency of the finite complement clause is taken into account, supporting an extreme version of the frequency-based processing approach; while Richter and Chaves (2020), responding to potential methodological issues in Liu and colleagues' study, find limited evidence for a frequency-based approach, arguing instead in favor of an approach based on semantics and information structure. The picture that emerges is of a highly variable empirical landscape that makes theoretical progress difficult.

Recent empirical studies of bridge effects have recognized that the variability in the results is at least partially related to variability in how many verbs were studied. This has led to an increase in sample sizes over time: 8 in Featherston 2004 in German; 12 in Ambridge & Goldberg 2008; 24 and 48 in Liu et al. 2022; 75 (experimental) and 136 (corpus) in Richter & Chaves 2020. Given that the results appear to change with sample sizes, the logical conclusion is that the field could benefit from testing a comprehensive set of verbs. Though creating the data set would be resource-intensive, it would both license more accurate evaluations of existing theories and open new pathways for theory construction. It would also act as a benchmark data set, thereby relieving each study of the need to invest resources in collecting yet another sample of verbs. To that end, we collected two extremely large-scale data sets of bridge effects for 640 finite clause-embedding verbs in English—a nearly exhaustive list of such verbs in English (see Appendix A). The first measured acceptability with sentences presented in isolation; the second with preceding (supportive) context. So that we have good estimates, each experiment targeted responses from 60 participants per verb, requiring the recruitment of over 9,000 participants on Amazon Mechanical Turk (Section 3). We also compiled measures for each verb that are

relevant to theories of bridge effects (Section 4). For information-structure theories, this entailed a third large-scale experiment to collect backgroundedness judgments using over 5,000 additional participants. For template-based processing theories, we collected twelve measures of semantic similarity from the natural language processing literature. And for frequency-based processing theories, we collected two measures of frequency from COCA. We have made our data sets publicly available on our websites for researchers to use in their own studies of bridge effects.

As a first set of studies with these new, nearly exhaustive data sets, this article addresses three research questions.

Our first question is empirical: Do bridge effects exist? Section 5 evaluates this question on our full set of verbs, responding to two claims in the literature. First, Liu and colleagues (2022) argue that once the frequency of the verb co-occurring with finite clauses is accounted for, bridge effects disappear (i.e., there is a main effect of frequency and a main effect of extraction on acceptability, but no interaction). Second, information structure accounts have reported that supportive context can improve long-distance wh-extraction for certain non-bridge verbs, inviting the prediction that context can reduce or even eliminate bridge effects. Using the analysis approach adopted by Liu and colleagues, we find that bridge effects exist even after accounting for each of the main predictors from each of the three prominent theories: frequency (contra Liu and colleagues), backgroundedness, and semantic similarity. With regard to context, we find that context does reduce the penalty for wh-extraction, but the effect is relatively small on average. Crucially, context neither eliminates nor reduces the between-verb differences in penalties that characterize bridge effects, contrary to what one might expect from information structure accounts.

Our second question is the central theoretical debate in the literature: What is the source of bridge effects? Section 6 presents confirmatory analyses to evaluate predictions of the three theories, using the approach adopted in Ambridge & Goldberg 2008: simple linear regressions between the predictor variables for each theory and our acceptability data. Anticipating our results, we find that the information structure-based theory performs best overall, but that (subjectively) none of the fits are particularly strong. This provides a window for understanding the seemingly incompatible results across studies—the relatively poor fits are more susceptible to sampling error based on the number of verbs tested. Our results further suggest that research on bridge effects could benefit by considering a wider range of theories beyond these three.

Therefore, our third question is exploratory: Are there new patterns visible in our data sets that could lead to a better theory? Section 7 reports just such a pattern: the penalty for long-distance extraction appears to correlate to a fair degree with whether a verb also selects for nonfinite complement clauses. As far as we know, this morphosyntactic property has not been noted before (and indeed, syntactic factors are underexplored in research on bridge effects). We first discuss what this property may be. Building on work by Wurmbrand (2019) on exceptional case marking and indexical shift in finite clauses, we suggest that for certain English clause-embedding verbs—but not others—complement clauses contain a dedicated position on the left periphery, which is variously exploited for exceptional case marking or binding (in nonfinite cases) or long-distance wh-extraction (in finite cases; see also Chomsky 1973, among many others). Then, in order to better account for gradient judgments and exceptions to our nonfiniteness generalization, we build on suggestions in Erteschik-Shir 1973, Richter & Chaves

2020, and Chaves & Putnam 2020 that multiple factors may be required to explain bridge effects. Specifically, we construct a multivariate theory that combines this new morphosyntactic property with the best predictors from existing theories. We show that this theory delivers better empirical coverage, even after controlling for complexity. This finding affirms the value of a multivariate approach and provides a novel argument for the view that morphosyntax can play an important role in bridge effects.

Section 8 then considers a few potential issues noted by reviewers: an alternative hypothesis tying bridge effects to the variation in the number of subcategorization frames allowed by clause-embedding verbs, and whether the design of our acceptability judgment task might have produced biased estimates of bridge effects. We present analyses showing that these suggestions, while reasonable, are not borne out in our data.

Finally, in Section 9, we consider two implications of our results for broader questions beyond bridge effects. We discuss results in Section 6 showing that bridge effects have smaller effect sizes (about .3 z-units) compared with island effects as a potentially relevant fact for efforts to unify bridge effects and island effects within a single theory. We also build on an observation in Section 6 that model fits are higher for sentences presented with context as a way to begin to explore the effect of context on the acceptability of long-distance dependencies.

2. REVIEW OF EXISTING ACCOUNTS.

2.1. DEFINING LONG-DISTANCE PENALTIES AND BRIDGE EFFECTS. Before reviewing the literature, it will be helpful to have a precise definition of bridge effects that follows both the theoretical and the experimental literature. The first step is to define the long-distance penalty that occurs for extraction (for all verbs). We define the long-distance penalty for a verb as the acceptability difference between a wh-question with extraction from the matrix clause (short dependency) and a wh-question with extraction from the complement clause (long dependency), as shown in 3.

(3) Long-distance penalty = short dependency rating – long dependency rating

a. ??What did Jo shout that Sam saw _? *long dependency*

b. Who _ shouted that Sam saw the movie? *short dependency*

The direction of this subtraction means that penalties will be positive when long-distance extraction is worse than short. We note that this matches the analysis done in Ambridge & Goldberg 2008, but not the general norm in the experimental literature to perform subtractions in the other direction (experimental condition – control condition). Another departure is our use of the short dependency as a baseline, rather than the declarative clause (e.g. *Jo shouted that Sam saw the movie*) typically used in prior experimental studies of bridge effects, such as Ambridge & Goldberg's. We defer to Section 8.3 a fuller discussion of this departure; for now, we will note that our setup ensures that both conditions are identical in almost all respects, including clause type (interrogative) and speech act (wh-question), differing only in wh-dependency length.

The second step is to define 'bridge effects' as differences in penalties between verbs—intuitively, the penalty for *shout* (a non-bridge) will be a larger positive number, and the penalty for *say* (a bridge) will be a smaller positive number. Again, this is consistent with how bridge effects have been defined in the empirical and theoretical literatures. However, we note that some discussions of bridge effects appear to focus exclusively on the variation in the absolute acceptability of long-distance extraction (i.e., just one condition). Taken at face value, this would

lead to a confound with the general acceptability of clause-embedding for each verb (as Ambridge & Goldberg 2008 note). We suspect that these discussions are simply using a shorthand—they are assuming that the acceptability of clause-embedding varies less than the acceptability of long-distance extraction, so one can focus exclusively on the interesting condition. But we must include both to quantitatively assess the theories.

2.2. INFORMATION STRUCTURE-BASED THEORIES. We begin our review by considering one of the first comprehensive theories of bridge effects: Erteschik-Shir's (1973) information structure theory (see also Ambridge & Goldberg 2008). The core of Erteschik-Shir's proposal, intended to cover both bridge effects and island effects, is stated in 4.

(4) Extraction can occur out of constituents that can be considered dominant in some context (Erteschik-Shir 1973:27), where 'dominant' is best understood as 'natural to comment on' (Erteschik-Shir ibid.:16) or as 'focusable,' allowing 'the speaker … to draw attention of the hearer' to the constituent (2017:7).

Erteschik-Shir (1973) provides a number of diagnostics for dominance/focusability. One such diagnostic is the 'lie test' (suggested by J. R. Ross). This test shows that, in 5, both the matrix clause and the complement clause of *think* can be dominant, because their propositions can be challenged as lies. In contrast, in 6, only the matrix clause is dominant. Intuitively, this is because a verb like *shout* draws attention to the manner of speaking, while with a factive verb like *know*, the embedded proposition is presupposed to be true.

(5) Fred thinks that Mary won.
  a. That's a lie, he doesn't think Mary won.
  b. That's a lie, she didn't win.

(6) Fred shouted/knew that Mary won.
  a. That's a lie, he didn't shout/know that Mary won.
  b. ??That's a lie, she didn't win.

Ambridge and Goldberg (2008:364-366; also Goldberg 2006; Cuneo & Goldberg 2023) reframe Erteschik-Shir's notion of 'dominance' (or 'focusability') in terms of 'backgroundedness', and propose a constraint that prohibits gaps from appearing inside backgrounded constituents. Ambridge & Goldberg support this proposal with a study testing 12 English verbs. For each verb, they estimated penalty scores with an acceptability judgment survey. They also measured how much the verb 'backgrounds' its complement clause, using a negation test, as shown in 7, which is a variant of the lie test. The negation test works as follows: if sentential negation negates the proposition represented by a complement clause, the clause is focused; if sentential negation fails to negate the complement clause, the clause is backgrounded. Ambridge & Goldberg found a strong correlation between backgroundedness and penalty scores (Pearson r=-.83, p=.001; see also Dąbrowska 2013 for a replication with 16 verbs).

(7) a. *Maria didn't know that Ian liked the cake.*
  Does not imply *Ian didn't like the cake.* (*know* backgrounds its clause)
  b. *Maria didn't think that Ian liked the cake.*
  More likely to imply *Ian didn't like the cake*. (*think* does not)

However, an analysis of only 12 verbs suffers from power issues, which Ambridge & Goldberg themselves (2008:375) note. Underpowered studies with statistically significant results

tend to overestimate the effect size (e.g. Vasishth et al. 2018). Our study tackles this power issue directly by testing a nearly exhaustive set of 640 verbs.

Information structure-based theories have also claimed that supportive context can make non-bridge verbs salient and their complement clause dominant/focusable, thereby improving the acceptability of wh-extraction from the clause (Erteschik-Shir 1973, 2017, Ambridge & Goldberg 2008, Müller 2015, Chaves & Putnam 2020; cf. Kothari 2008 for reading time experiments). The weak version of this claim is that context will reduce penalties by the same extent for all verbs, leaving bridge effects—the between-verb differences in penalties—unchanged. A stronger version of this claim is that context will improve long-distance wh-extraction for non-bridge verbs more than bridge verbs, thereby reducing or eliminating bridge effects. It is not always clear which version of the claim is endorsed in a given paper, but we note that the general discussion in the field tends to implicitly assume the stronger claim. Our two nearly-exhaustive data sets, without and with context, allow us to test both versions of the claim.

2.3. TEMPLATE-BASED PROCESSING THEORY. A second theory derives bridge effects from difficulty experienced during template-based processing. Dąbrowska (2008, 2013, etc.; see also Verhagen 2005, 2006) argue that the processing of questions with long-distance wh-dependencies (henceforth, 'long-distance wh-questions') involves the use of lexical templates like 'WH *do you think* S-GAP' or 'WH *do you say* S-GAP', where WH and S-GAP are respectively variables for a wh-phrase and a finite clause with a gap. Lexical templates are pre-assembled 'lexical formulas' based on the most frequent long-distance wh-questions, which feature second person subjects and a verb like *say* or *think*. Templates free speakers from having to build the representation of a wh-question from scratch. Instead, speakers can insert appropriate phrases into the WH and S-GAP variables in a template.

To process a long-distance wh-question with a verb other than *say* or *think*, one must further alter the template by replacing the verb. As Ambridge & Goldberg suggest, bridge effects might reflect how easy replacing the verb is, which in turn depends on how semantically similar the replacement verb is to *say* or *think*. In this view, long-distance wh-extraction incurs a small penalty for *claim* because *claim* and *say* are semantically similar, and so replacing *say* with *claim* is easy. *Shout*, on the other hand, is associated with a higher penalty because *shout* is less similar to *say*.[1]

This account differs from the information structure theory in two ways. First, bridge effects here are language processing artifacts. Second, this account is not as complete a theory, as it does not explain why *say* and *think* can appear with long-distance wh-dependencies in the first place.

We note that Ambridge and Goldberg (2008:380–382) found little evidence for this theory: bridge effects showed no correlation with similarity measures derived from a judgment survey and a Latent Semantic Analysis calculator (LSA; Deerwester et al. 1990) for their set of 12 verbs. However, one could wonder if their null result might reflect issues with their sample of verbs and/or LSA data set. We address these potential concerns by re-evaluating this theory

[1] This formulation departs from Dąbrowska 2008, 2013. For Dąbrowska, verb similarity might affect the absolute acceptability of long-distance wh-questions, but not necessarily bridge effects (but see Section 2.1).

through our much larger set of verbs and also with four measures of semantic similarity (see Section 4.2).

2.4. FREQUENCY-BASED PROCESSING THEORY. The original frequency-based processing theory posits that verbs that appear more frequently with a finite complement clause will have smaller penalties for extraction from the clause. This builds on the idea that frequency is straightforwardly correlated with processing difficulty and/or acceptability (e.g. Hale 2001, Levy 2008; but see Sprouse et al. 2018, White & Rawlins 2020). Some data in favor of such a view can be found in Kothari 2008, which showed that frame frequency (the frequency that a verb appears with a finite clause) and a verb's bias for a finite clause (i.e. the conditional probability of such a clause given a verb) are both correlated with the absolute acceptability of long-distance wh-questions.

In more recent work, however, Liu and colleagues (2022) have taken the frequency-based processing theory a step further, and argued that there are in fact no bridge effects once frame frequency is taken into account. Their experiments compare acceptability for a condition with extraction to a condition without extraction for 24 and 48 verbs with varying frame frequencies: they find a main effect of extraction (i.e., a penalty for all long-distance extraction) and a main effect of frame frequency (frequency being correlated with acceptability), but, crucially, no interaction between these two factors. This means that they observe no bridge effects: no meaningful differences in the size of the penalties between verbs.

There are several methodological reasons to follow up on Liu and colleagues' result. For one, as Richter and Chaves (2020) note, the lack of statistical interaction might reflect participant fatigue since Liu and colleagues' experiments were designed to be very long (e.g. 288 items in their Experiment 2) and uniform (all items were critical items, with no filler items). Liu and colleagues' experiments also only tested samples of 24 and 48 verbs—larger than Ambridge and Goldberg's but still relatively small compared to the full range of clause-embedding verbs in English. Richter and Chaves (2020) begin to address this with an experimental study of 75 verbs, ultimately concluding against Liu and colleagues. Our study takes this to the logical conclusion by testing a nearly exhaustive set of verbs (640) using much shorter experiments (31 items) with a 2:1 ratio of (pre-tested) fillers to target items.

3. QUANTIFYING BRIDGE EFFECTS USING ACCEPTABILITY JUDGMENT EXPERIMENTS. In this section, we describe how we obtained quantitative measures of bridge effects, by compiling a set of 640 clause-embedding verbs and collecting the acceptability of wh-questions containing these verbs, in isolation and after a context sentence. In the next section (Section 4), we describe how we compiled predictors for each of the three theories reviewed above. We discuss results starting in Section 5.

3.1. COMPILING A NEARLY EXHAUSTIVE SET OF FINITE CLAUSE-EMBEDDING VERBS. We assembled 640 verbs (Appendix A) from Anand, Grimshaw & Hacquard 2019; Levin 1993; and the MegaVeridicality data set (White & Rawlins 2018). We focused on verbs whose active voice forms must assign thematic roles to a subject. This criterion includes verbs like *say*, *shout*, and other verbs canonically used to illustrate bridge effects, while excluding raising verbs like *seem* and psych-verbs like *surprise*, which can appear in the active voice with an expletive *it* subject. From the resulting set of 641 verbs, we excluded the verb *animadvert* because it is so rare as to be absent from the one-billion-word Corpus of Contemporary American English (COCA; Davies 2020), leaving 640 verbs.

3.2. ACCEPTABILITY JUDGMENT EXPERIMENT 1: JUDGMENTS IN ISOLATION (NO CONTEXT).

CONSTRUCTING WH-QUESTION ITEMS. We first created a set of 10 different frames, each with a distinct combination of nouns and verbs. An example is shown in 8 with the verb *tell*. These frames were intended to be semantically compatible with a majority of the 640 clause-embedding verbs, so that any acceptability variation can be reasonably attributed to the verb being incompatible with long-distance wh-dependencies and not to plausibility.

(8) The party leader told the vice-president that the senators would endorse the governor.

For a subset of 286 verbs, however, we judged that certain frames were unsuitable for specific lexical semantic reasons. For instance, *broadcast* and *editorialize* prefer subjects denoting media organizations, like *the TV station.* We sorted these verbs into 52 smaller classes, based on our intuitions, and adapted the frames accordingly.

The tense and modality of the complement clause also varied for similar reasons. To maximize comparability, we made *would* the default tense/modal marker, as in 8, because some verbs have predictive semantics (e.g. *predict*, *expect*) and are most felicitous with future modals. But we changed the tense/modal marker in the complement clause where required by the verb.

The 10 frames (and modifications) lead to 10 lexically matching pairs of short and long wh-questions for each clause-embedding verb, i.e. two conditions for each verb, consistent with our operationalization of bridge effects (Section 2.1). The two conditions are illustrated in 9 and 10 respectively. Clause-embedding verbs appeared in the past tense, like *told* in 8, with four exceptions: we judged that *stand*, *bear*, and *forgive* should co-occur idiomatically with the modal *couldn't*, while *care* and *mind*, being negative polarity items, should co-occur with the auxiliary *didn't*. As for the wh-word, this was chosen to match the corresponding NP in each frame: if the NP denoted human entities or groups, the wh-word was *who*, otherwise *what*.

(9) Who _ told the vice-president that the senators would endorse the governor?

(10) Who did the party leader tell the vice-president that the senators would endorse _?

The full set of experimental items can be found along with the acceptability ratings data set.

LIST CREATION. To avoid fatigue effects, we made the acceptability judgment surveys short: a total of 31 sentences, consisting of eight target sentences that varied by list and 23 filler sentences that were identical across lists. The eight target sentences consisted of four clause-embedding verbs, each appearing in both short and long conditions, so that we could calculate a within-participant penalty score per verb. We split our 640 verbs into four bins of 160 verbs based on how frequently they occur with complementizer *that* in COCA. We then formed 160 quadruplets to be tested together in a single survey by randomly sampling one verb from each bin, so each quadruplet has a mix of verb frequencies. We also made sure that each quadruplet had a mix of semantic types (e.g., never all manner-of-speaking).

The 10 frames for each verb were then distributed in a Latin Square design, so that participants would never see the same frame across verbs or conditions in their list. Doing so yields 10 different lists for each of the 160 quadruplets, for a total of 1,600 unique lists.

FILLER COMPOSITION. Our 23 fillers were based on sentences from Sprouse, Schütze & Almeida 2013 that have well-established acceptability ratings and are known to span the range of possible ratings in a 7-point acceptability judgment task. Our fillers were intended to introduce variability in the items, to combat fatigue and boredom, as well as to encourage participants to use the full range of the acceptability scale. Nine of the fillers appeared in a fixed order at the start of each survey; the first seven each have an expected mean of 7, 1, 6, 2, 5, 3, 4 respectively, while the eighth and ninth have an expected mean of 1 and 7. The remaining 14 fillers were composed of two items each with expected ratings of 1 through 7. These 14 fillers and eight target sentences were presented using Ibex's built-in random presentation function, so that no two target sentences were presented consecutively. Lastly, we made sure that fillers did not contain any of the 640 verbs or embedded *that* clauses.

A SYNONYM POST-TEST FOR DETERMINING VERB FAMILIARITY. Many of our 640 verbs appear infrequently, like *grok* and *expostulate*. We implemented a post-test at the end of each acceptability rating survey to check for participants' familiarity with the verbs. Each of the four verbs surveyed appeared in descending order based on frequency, together with an example declarative sentence formed by one of the 10 frames for that verb. Participants were to identify the synonym, i.e. pick a verb or phrase 'closest in meaning', from a set of four choices presented in a random order: a close synonym, an antonym, and two other semantically unrelated verbs. We use this post-test to eliminate both uncooperative participants and trials in which the participant does not know the meaning of the verb (see subsection on data analysis).

TASK AND PRESENTATION. Stimuli were presented using the Ibex experiment platform (Drummond 2012). Participants were instructed to rate sentences on a 7-point scale based on whether they think a native speaker of English could say these sentences in a conversation. Participants were instructed to ignore prescriptivist rules. A rating of 1 indicated that the sentence was 'very bad' and a rating of 7 'very good'. To anchor the scale, three example sentences with suggested ratings of 1, 4, and 7 were embedded within the instructions. These were also taken from Sprouse et al. 2013 with means of 1, 4, and 7. Sentences were presented one per screen. Participants could take as much time as they wanted to respond, but could not go back to previous sentences.

PARTICIPANTS. We recruited participants in two stages, first directly on Amazon Mechanical Turk (AMT) and then through CloudResearch, a platform that allows researchers to target a pool of more reliable AMT participants. All participants were self-reported native speakers of American English. We used AMT and CloudResearch filters to ensure that participants were above 18 years of age and based in the United States. Participants recruited directly on AMT also had to have completed more than 500 tasks and received an approval rating of at least 95% on their previous tasks. Each participant received US$1 for completing a survey; this was based on an hourly rate of US$12/hour and our estimate that a survey, with 31 items total, should take at most 5 minutes to complete.

For each survey, we planned to recruit 60 participants (30 directly on AMT and 30 through CloudResearch), in order to collect 60 responses per verb per condition, although we sometimes inadvertently collected more responses due to the mechanics of Ibex and AMT. We tracked participants so that each participant only completed one survey. We also tried as far as possible to minimize overlaps in the AMT and CloudResearch participant pools. Of the 9,219 unique participants recruited, only 805 participants (8.7%), completed two surveys, once as part of the AMT pool and once as part of the CloudResearch pool. Although the filler items were identical across the two surveys for those participants, the two surveys were run almost two years apart, so the participants were unlikely to remember the sentences.

DATA ANALYSIS AND OUTLIER DETECTION. After data collection, we identified four sentences that were incorrectly presented, and removed those four sentences from analysis. All ratings were z-score transformed by participant to eliminate common forms of scale bias. We only included participants if they were native speakers, read the sentences carefully, responded accurately to the fillers, and knew the verbs, operationalized as follows:

1. They answered yes to two language questions: that they lived in the United States from birth to at least age 13 and that their parents spoke English to them at home.
2. Their median response time to each sentence is at least 2.5 seconds.
3. They responded to at least 12 out of 14 fillers with a rating that is within 2 standard deviations of the mean for that filler.
4. They gave at least three correct responses out of a maximum of four to the synonym post-test.

We included individual trials if they were read carefully and the participant knew the verb, operationalized as:

1. The trial response time was at least 2.5 seconds.

2. The synonym of the verb in the trial was correctly identified in the post-test.

We further excluded a subset of verbs from analysis. We were deliberately liberal in compiling the 640 clause-embedding verbs. Consequently, for certain verbs, clausal complements might be ungrammatical for many native speakers. To identify these verbs, we calculated a mean z-score for each verb in the short wh-dependency condition, where there is no gap in the complement clause. We excluded the 150 verbs with a negative mean z-score for this condition, i.e. below the grand mean of all items in the surveys (designed to have a mean rating near 4, the midpoint of the scale). In addition, for six other verbs, mean short wh-dependency ratings were lower than mean long wh-dependency ratings. This pattern is anomalous under all of the theories under consideration here. We also excluded these verbs out of an abundance of caution. Altogether, these criteria yield a total of 41,958 responses (20,979 pairs of ratings for short and long conditions) for 484 verbs for analysis.

3.3. ACCEPTABILITY JUDGMENT EXPERIMENT 2: WITH PRIOR CONTEXT. Bridge effects are typically illustrated by presenting wh-questions in isolation. Similarly, as far as we know, all previous experimental studies, except one (Kothari 2008), have tested them in isolation. However, as mentioned in Section 2.2, information structure accounts have claimed that supportive context can impact extraction and possibly penalties and bridge effects. We therefore believe it is important to also test the verbs with prior context.

To our knowledge, there are no detailed theories of how supportive context affects penalties. However, previous work has reported that a dialogue format can improve wh-extraction from complements of non-bridge verbs or island structures (Chaves & Putnam 2020; see Ambridge et al. 2015:e120 for islands). Here, we adapt the dialogue format used by Ambridge and colleagues. Although the degree of improvement is disputed (see Pérez-Leroux & Kahnemuyipour 2014), this dialogue is one of the few concrete proposals in the locality literature, and therefore seemed like a reasonable candidate for a first systematic exploration of context effects.

MATERIALS. All target items were presented as a dialogue between two individuals, A and B, as in 11 and 12.

(11) A: Someone thought that the duchess would invite the arrogant knight.
B: Really? <u>Who thought that the duchess would invite the arrogant knight?</u>

(12) A: The princess thought that the duchess would invite a certain person.
B: Really? <u>Who did the princess think that the duchess would invite?</u>

A's utterance was always an assertion in which a clause-embedding verb appears with a complement clause with no gap. B's utterance was always a wh-question responding to A's assertion, which justifies B's use of the verb and the complement clause. We added *Really?* to signal that B is responding to A's utterance. Participants were instructed to judge only the last sentence in B's utterance, i.e., the question, which was underlined. Before starting, participants saw three similarly-formatted dialogue examples, with suggested ratings.

We constructed these items with the same procedure described in Section 3.2. For short wh-dependency items, A's utterance always featured *someone* or *something* in a subject position, corresponding to *who* or *what* in B's response. For long wh-dependency items, A's utterance always featured *a certain person*/*thing* in the object position, corresponding to *who*/*what* in B's

response. We did not use *someone*/*something* because we judged that they would be likelier to receive a nonspecific reading, making it odd to question them.

For fillers, we reused the items described in Section 3.2, creating suitable declarative sentences for A as context.

To ensure that participants used the context provided to judge target sentences, we added the four catch trials in 13–16. The target sentences contained the presupposition trigger *either*, licensed if the context sentence was negated, or *too*, licensed if the context sentence was in the affirmative. Participants who read the entire dialogue should notice that 15 and 16 are infelicitous, due to presupposition failure, and give lower ratings to the target sentences.

(13) A: The carpenter did not repair the table.
B: The apprentice did not repair the table either.

(14) A: The diver went to the pool.
B: The swimmer went to the pool, too.

(15) A: The boys ate the broccoli.
B: #The girls did not eat the broccoli either.

(16) A: The guide did not board the bus.
B: #The tourists boarded the bus, too.

A synonym post-test was also included at the end of each survey.

PARTICIPANTS. Participant recruitment proceeded as described in Section 3.2; again, we targeted recruiting 60 participants per survey. Participants received US$1.20 for completing a survey. Most participants completed only one survey; only 703 (7.7% of 9,156 unique participants) completed the surveys twice (again, at least two years apart).

DATA ANALYSIS AND OUTLIER DETECTION. For data quality purposes, we applied the same inclusion criteria to the data set, except that:

1. The response time threshold was raised to 3 seconds, since items were longer.
2. We analyzed ratings for the four catch trials to ensure that participants were basing their judgments on the dialogues. Participants must have mean z-scored ratings for the felicitous items that are at least 0.5 units higher than those for the infelicitous items.

So that we can compare ratings for wh-questions with and without prior context, we filtered for the same 484 verbs, yielding a total of 21,732 responses. This number is only about half of the number of responses for the wh-questions presented without context, partly because many participants failed the new catch trial criterion, which requires a difference of 0.5 z-units. Relaxing this criterion to any positive difference only slightly increases the number of included responses (to ~26,000), so we have opted to keep the stricter criterion in place.

4. OBTAINING PREDICTORS OF BRIDGE EFFECTS.

4.1. INFORMATION STRUCTURE THEORIES: THE NEGATION TEST FOR BACKGROUNDEDNESS.

In information structure theories, the key predictor is how much a verb's complement clause is 'dominant/focusable' or the inverse, 'backgrounded'. For our purposes, we assume that this can be quantified using Ambridge and Goldberg's (2008) negation test. For compatibility with their experiment, we will label this measure 'backgroundedness'.

TASK. We collected judgments for the negation test over the internet using Ibex. We pair a sentence in which a clause-embedding verb was negated with another sentence formed by the verb's complement clause, as in 17. Participants were instructed to decide whether the second sentence was true or false, using only information from the first sentence and not any real-world facts. The second sentence was underlined, to make it clear that this was the sentence to judge.

(17) The princess didn't {think/know/…} that the duchess would invite the arrogant knight.
<u>The duchess will invite the arrogant knight.</u>

To the extent that participants judged that the second sentence was true (e.g. for *know* but not *think*), the verb backgrounds the complement clause. We provided a third option, 'Not enough information,' in case participants found it difficult to make True/False judgments; a further advantage was that it lets us calculate two (slightly) different measures of backgroundedness (see 'Data analysis' subsection below). Before starting the experiment, participants saw three example items intended to elicit a True response, a False response, and a 'Not enough information' response, in that order.

We note that our design departs from prior work. Both Ambridge and Goldberg and Liu and colleagues used a Likert scale, where the rating indicates how true participants felt the second sentence was, in light of the first sentence. In pilot testing, we noticed that individuals were unlikely to use the full range of a Likert scale, nor use it in a way that corresponds naturally to an ordinal scale: with a 7-point scale, testers reported often using only three options, typically 1 (False), 7 (True) and 4; testers reported using 4 to indicate uncertainty, and not because they felt that the sentence was midway between being True and False. Therefore we judged that a three-way True/False/Not enough information format would capture participant intuitions more transparently. Crucially, it still yields a gradient measure in the sense that each verb could have a different proportion of responses out of the 60 participants per verb (similar to the way Liu and colleagues use binary acceptability judgments for their experiments).

LIST CREATION. Surveys were 27 items long, with eight clause-embedding verbs per survey and 19 fillers. We divided the 640 verbs into eight bins, based on how often a verb occurs with complementizer *that* in COCA. We formed 80 sets of eight verbs by randomly sampling, without replacement, from each bin.

MATERIALS. Verbs and frames were identical to those used for the acceptability judgment task. For the first sentence, all clause-embedding verbs were negated with *didn't*. We applied this negation to the verbs *stand*, *bear*, and *forgive* for consistency, even though doing so produced ungrammatical sentences for *stand* and *bear*, which co-occur idiomatically with *couldn't*, and potentially changes the meaning of the sentence for *forgive*; these three verbs were ultimately excluded from the analysis. For the second sentence, we reused the complement clause of the first sentence, except for verbs like *prefer* or *ask*, which have subjunctive complement clauses. For these verbs, we added the modal *should* to the second sentences. If the complement clause contained the modal *would*, the second sentence contained the modal *will* instead, as 17 shows. An anonymous reviewer points out that this introduces a mismatch that might affect responses, perhaps resulting in more 'Not enough information' responses. While the exact impact (if any) will have to await further investigation, replacing *would* is necessary because *would* in a main clause has a conditional reading, not the intended future reading. Furthermore, replacing it with *will* is entirely consistent with English sequence of tense: in 17, *would* can be analyzed as *will* that has agreed in tense with the main clause.

As was the case for the acceptability judgment surveys, we included a synonym post-test for the eight verbs tested in each backgroundedness survey.

FILLERS. The 19 filler items each consisted of two sentences. Five fillers were intended to elicit a True response ('True fillers'): The second sentence corresponded to a presupposed clausal subject or adjunct in the first sentence. Eight fillers were intended to elicit a False response ('False fillers'); the first sentence in these fillers featured some kind of negation, while the second sentence was the affirmative variant. The remaining six filler items were likely to elicit a 'Not enough information' response: the first sentence in these fillers typically contained a modal adverb or auxiliary, while the second sentence was the affirmative variant. We note that we introduced more False fillers than True fillers. This was an error relative to our initial intentions (an equal number of False and True fillers), but in retrospect, doing so is not unwelcome. Many clause-embedding verbs are factive or invite a factive reading, meaning test items were more likely to be judged True. The extra False fillers may have serendipitously led to a better balance of responses within the experiment.

PARTICIPANTS. Participant recruitment proceeded as described in Section 3, targeting 60 participants per survey, as before. Each of the 5,069 unique participants completed only one survey. Participants received US$1.20 for completing a survey; this payment rate was based on a US$12/hour rate and our estimate that each survey, consisting of 27 items and a relatively difficult judgment task, might take slightly over 5 minutes to complete.

DATA ANALYSIS. For data quality purposes, we imposed criteria that were identical to the ones in Section 3.2, except that participants must have given (i) 6 or more correct responses to the synonym post-test and (ii) 9 or more correct responses to the 13 'True' and 'False' fillers (i.e. a True response to the True fillers and a False response to the False fillers).

After data collection, we identified one sentence that was incorrectly presented and removed it from analysis. We also excluded all responses for sentences featuring *stand*, *bear*,

and *forgive*, because these verbs were negated with *didn't*, but idiomatically co-occur with *couldn't*.

We calculated two related backgroundedness measures. The first measure, %True, was defined as the percentage of True responses for each verb. The second measure, %Not-False, used a more liberal definition: the percentage of responses that were either True or 'Not enough information'. The higher these measures, the more backgrounded the complement clause is.

4.2. TEMPLATE-BASED PROCESSING THEORIES: SIMILARITY WITH *SAY* AND *THINK*. This theory assumes that the processing of long-distance wh-questions relies on lexical templates featuring *say* and *think*. Bridge effects reflect the difficulty of replacing *say* and *think* in the templates, which depends on how semantically similar the replacement verb is to these two verbs: the greater the similarity, the smaller the penalty.

Ambridge and Goldberg argued against this hypothesis by showing that penalty scores are not significantly correlated with semantic similarity measures derived from Latent Semantic Analysis (LSA). We expand on this work by computing similarity measures using four different data sets from the natural language processing literature:

1. **LSA/Wikipedia** (Ştefănescu et al. 2014). These are word embeddings derived by applying LSA on a 2013 version of English Wikipedia. We chose this data set over the LSA calculator used by Ambridge and Goldberg because this Wikipedia-derived data set is more recent.
2. **Global Vectors (GloVe)/Wikipedia** (Fares et al. 2017). These are word embeddings created by applying GloVe (Pennington et al. 2014), an unsupervised learning algorithm, on a 2017 version of English Wikipedia.
3. **GloVe/Gigaword** (Fares et al. 2017). These are word embeddings created by applying GloVe on the 5th edition of the 4-billion-word Gigaword corpus, compiled from English news outlets (Parker et al. 2011). This provided an alternative to the Wikipedia-based measures above.
4. **WordNet** (Fellbaum 1998). WordNet provides a hierarchy of senses ('synset') for English words. For each verb, we used WordNet's definitions and example sentences, as included in Python's NLTK package (Bird et al. 2009), to identify the synset that most closely corresponds to how the verb is used in the items in the acceptability judgment tasks.

For each data set, we calculated similarity scores for each verb relative to three anchors: similarity with *say*, similarity with *think*, and a hybrid score that chooses whichever score is greater between *say* and *think* (following Ambridge & Goldberg 2008). We define similarity as cosine similarity for the first three data sets and as NLTK's path similarity measure for the WordNet data. Path similarity tracks the distance between a verb and *think* (or *say*) within WordNet's hierarchy.

4.3. FREQUENCY-BASED PROCESSING DIFFICULTY: FRAME FREQUENCY AND VERB BIAS IN COCA. To investigate frequency theories, we calculate two measures from COCA.

1. **Frame frequency**. This is how frequently a verb appears with a finite complement clause. Following Liu et al. 2022, we approximate this with how frequently the verb's lemma immediately precedes the complementizer *that* in COCA.

2. **Verb bias**. This measure is derived by dividing the raw frame frequency by verb frequency, following Richter & Chaves 2020.

We base our estimates on COCA because it is the largest American English corpus, so estimates would align better with our participants' linguistic experience as American English speakers. Furthermore, COCA is lemmatized and tagged for parts of speech, making it straightforward to identify a verb and a complementizer. However, because COCA does not mark the null complementizer, our frequency values are necessarily underestimates. This issue predominantly affects high-frequency verbs appearing in informal registers (e.g. *say* but not *testify*); which are verbs that most often appear with a null complementizer (e.g. Biber 1999).

5. DO BRIDGE EFFECTS EXIST? The first question we address with our data sets is whether bridge effects exist in both our 'no context' and 'with context' data sets. We do so for two reasons. First, recall that Liu and colleagues (2022) concluded, from their survey of 24 and 48 English verbs, that the acceptability of sentences with clause-embedding verbs presented without context can be modeled in terms of only two main effects (namely, presence of a long wh-dependency and frame frequency) and without an interaction effect, thus effectively denying the existence of bridge effects. Second, one possible prediction from information structure accounts is that bridge effects might decrease or even disappear in the presence of supportive context.

5.1. ARE THERE DIFFERENCES IN PENALTIES BETWEEN VERBS? Given that we operationalized bridge effects as differences in penalty scores between verbs, a direct test of whether bridge effects exist is to see whether penalties for each verb actually vary. Figure 1 shows substantial variation in both 'no context' and 'with context' mean penalty scores between the 484 verbs analyzed.

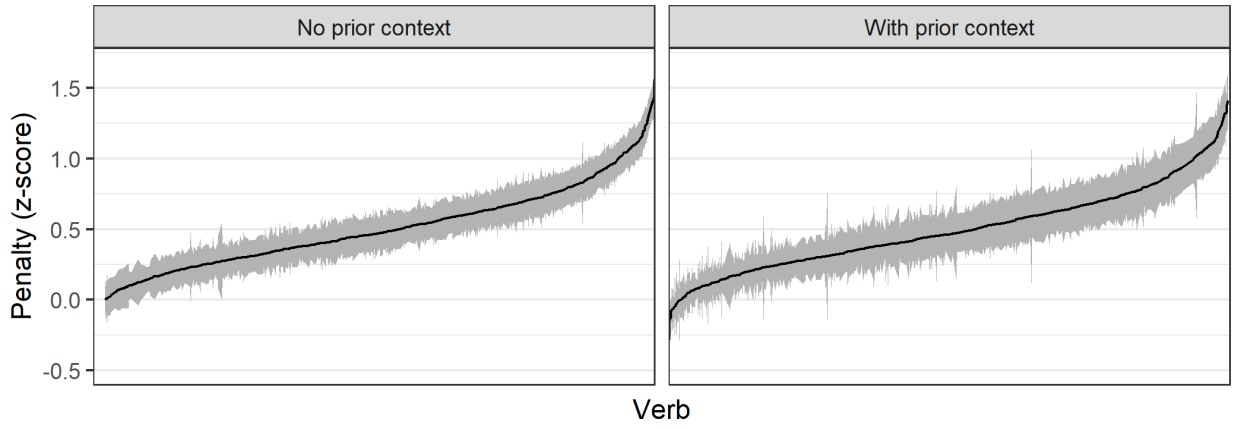


FIGURE 1. Penalty score means and standard errors (grey band) for the 484 verbs of interest, for wh-questions presented without and with prior context

To confirm this variation statistically, we ran a one-way analysis of variance (ANOVA). Using the R lme4 package (Bates et al. 2015), we fitted a linear mixed effects model of penalty scores, with verb as the predictor and random intercepts for participant (but not random slopes, due to model convergence problems, we suspect driven by the sheer size of the data sets). We then ran an ANOVA on this model, calculating *p*-values with the Satterthwaite approximation in the

lmerTest package (Kuznetsova et al. 2017). We found a significant effect of verb on penalty scores, without and with prior context (without prior context, $F(483, 13658)=4.41$; $p<.001$; with prior context, $F(483, 6414)=3.24$; $p<.001$). Put differently, penalty scores show significant between-verb differences, as expected if bridge effects exist both without and with context.

5.2. DO BRIDGE EFFECTS PERSIST AFTER CONTROLLING FOR A SPECIFIC PREDICTOR? We ran a second analysis closely modeled upon Liu and colleagues' analyses. In theory, this second analysis addresses the same question—do penalties differ by verb? But this analysis differs from the analysis in Section 5.1 in two ways: (i) it uses individual judgments of each condition as an outcome variable rather than penalty scores (which likely leaves more variance for the model to partition), and (ii) it uses a theoretical predictor (backgroundedness, frame frequency, or semantic similarity) instead of the atheoretical predictor 'verb' (which allows us to see if the short and long conditions behave differently in the presence of the predictor). We fitted linear mixed effects models predicting sentence acceptability based on crossing two factors: wh-dependency length (short, long) and the continuous independent measure of interest for each of the three prominent theories. For thoroughness, we constructed a different model for each frequency measure as well as each information structure and semantic similarity measure. To the extent that we find interaction effects, that would imply that the acceptability of sentences with clause-embedding verbs cannot be simply modeled using just two main effects (namely, wh-dependency length and each of the various measures), contra Liu and colleagues.

Each model included only by-participant and by-item random intercepts, as not all models converged with random slopes. We take these models to provide evidence for bridge effects when the interaction term's *p*-value is less than .05 (according to lmerTest); and the model's Bayes factor relative to a model without an interaction term (BF_{10}) is above 3 (calculated with the bayestestR package; Makowski et al. 2019), i.e. the data is at least 3 times more likely under a model with the interaction term than without an interaction term (Jeffreys 1961, Kass & Raftery 1995).

As Table 1 shows, significant interaction effects were reliably detected in our data, indicating bridge effects. Figure 2 illustrates the interactions, by plotting one selected measure for each theory. In this figure and subsequent scatterplots, we label non-factive *think* and factive *know*, which are often used in the literature to illustrate bridge effects. For information structure and frequency-based theories, both 'no context' and 'with context' data sets have interactions (in the predicted direction) with *p*-values below .05 and Bayes factors above 3. For template-based theories, the two GloVe models and the Wordnet model meet these criteria; only the LSA-based models do not. Overall, we take these results as converging evidence that there are bridge effects as classically defined in the literature (contra Liu et al. 2022).[2]

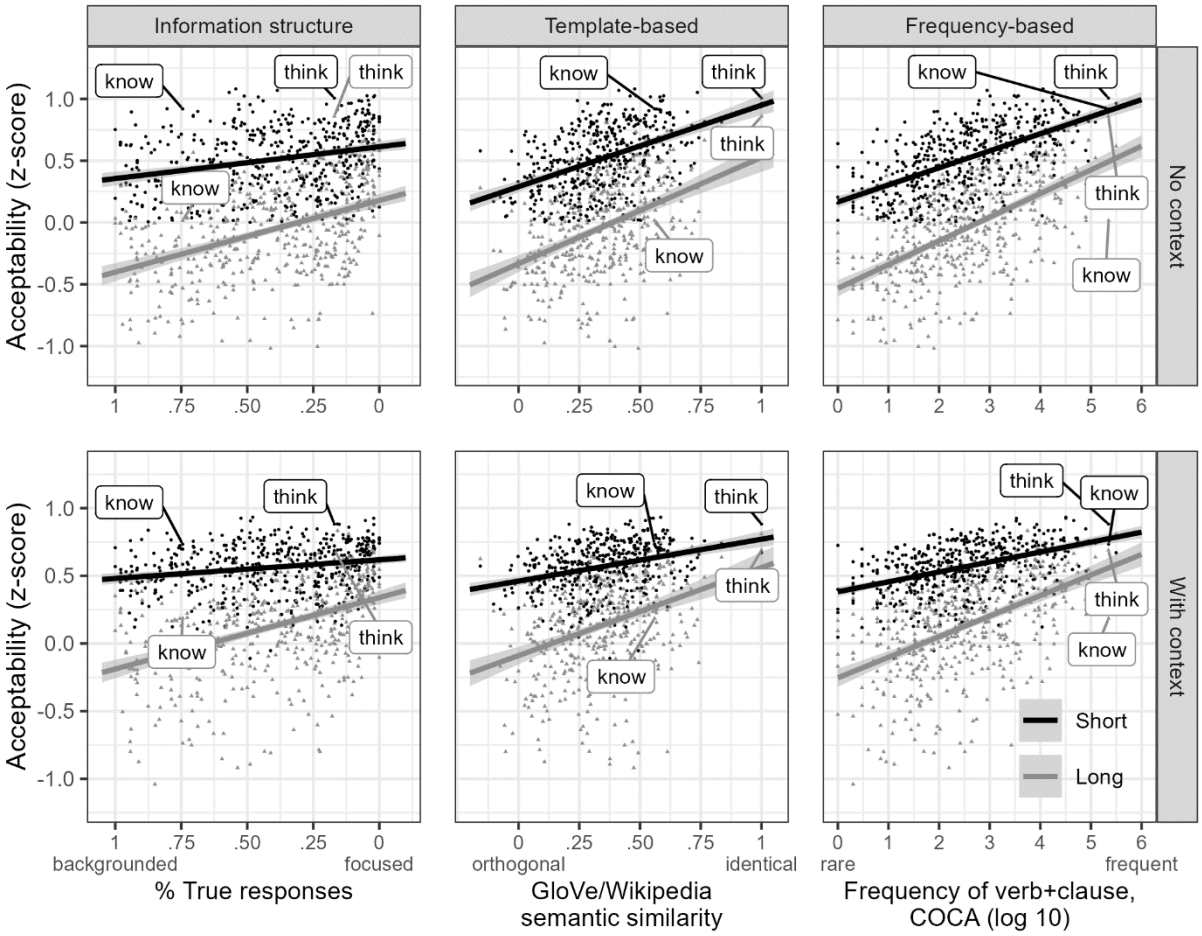
[2] Liu and colleagues (2022) created ordinal mixed effects models on acceptability ratings, whereas we ran linear mixed effects models on z-transformed Likert ratings. Out of an abundance of caution, we also fitted three ordinal mixed effects models of 'no context' raw ratings for the same set of verbs, with wh-dependency length as one fixed effect and a representative predictor for each of the theories—%True responses, GloVe/Wikipedia semantic similarity, and log frame frequency— as the other fixed effect. Except for the template-based model, these models detected significant interaction effects in the predicted direction and Bayes factors above 3. The relatively poor performance of the template-based model is

	No prior context					With prior context				
	b (s.e.)	*t*	*p*	Eff. Size	BF_{10}	*b (s.e.)*	*t*	*p*	Eff. Size	BF_{10}
Information structure theory										
% True responses	-0.33 (0.02)	-14.02	<.01	0.33	>100	-0.39 (0.03)	-13.72	<.01	0.39	>100
% Not-False responses	-0.74 (0.05)	-13.51	<.01	0.51	>100	-0.77 (0.07)	-11.28	<.01	0.53	>100
Template-based processing theory (hybrid score)										
LSA/Wikipedia	-0.12 (0.04)	-3.01	<.01	0.11	0.5	-0.09 (0.05)	-1.68	.09	0.07	<0.1
GloVe/Wikipedia	0.19 (0.04)	5.24	<.01	0.22	>100	0.34 (0.05)	7.51	<.01	0.39	>100
GloVe/Gigaword	0.17 (0.04)	4.84	<.01	0.19	>100	0.30 (0.04)	6.69	<.01	0.33	>100
Wordnet (log)	0.15 (0.03)	4.52	<.01	0.17	>100	0.16 (0.04)	3.90	<.01	0.19	13.5
Frequency-based processing theory										
Frame frequency (log)	0.06 (0.01)	9.67	<.01	0.30	>100	0.08 (0.01)	11.48	<.01	0.45	>100
Verb bias (log)	0.12 (0.01)	11.37	<.01	0.37	>100	0.19 (0.01)	14.10	<.01	0.56	>100

Column labels: 'b', 's.e.', 't', 'p' = coefficient, standard error, *t*-value, and *p*-value of the interaction term, according to lmerTest (with a floor of .01); 'Eff(ect) size' = cumulative size of the interaction effect over the full range of each independent measure (in z-units); 'BF_{10}' = probability of the data under the model with an interaction term relative to a model without this term (with a ceiling of 100 and a floor of 0.1). Analyses with untransformed values were also run but produced worse model fits; these are not reported here for space reasons. Similarly, for the template-based processing theory, the only results reported are for the hybrid similarity score; model fits for *say* and *think* similarity scores are roughly equal (see Appendix C).

TABLE 1. Estimates for interaction effect for models of acceptability of wh-questions presented with or without prior context. Each model crosses wh-dependency length with one predictor, with each predictor corresponding to a theory of bridge effects.

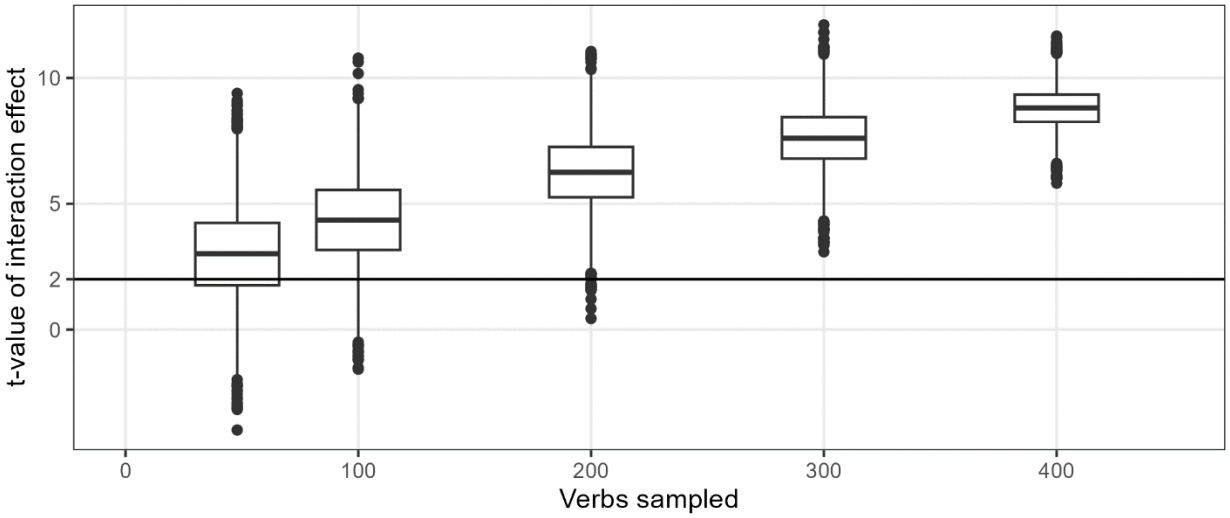
perhaps unexpected given Table 1, but recalls Ambridge and Goldberg's findings as well as anticipates those of Section 6.



Note: Each verb is represented by a dot. In this figure and subsequent scatterplots, *think* and *know* are labeled, since they are frequently used to illustrate bridge effects.

FIGURE 2. Interaction plots of acceptability of wh-questions for selected predictors of information structure, template-based processing, and frequency.

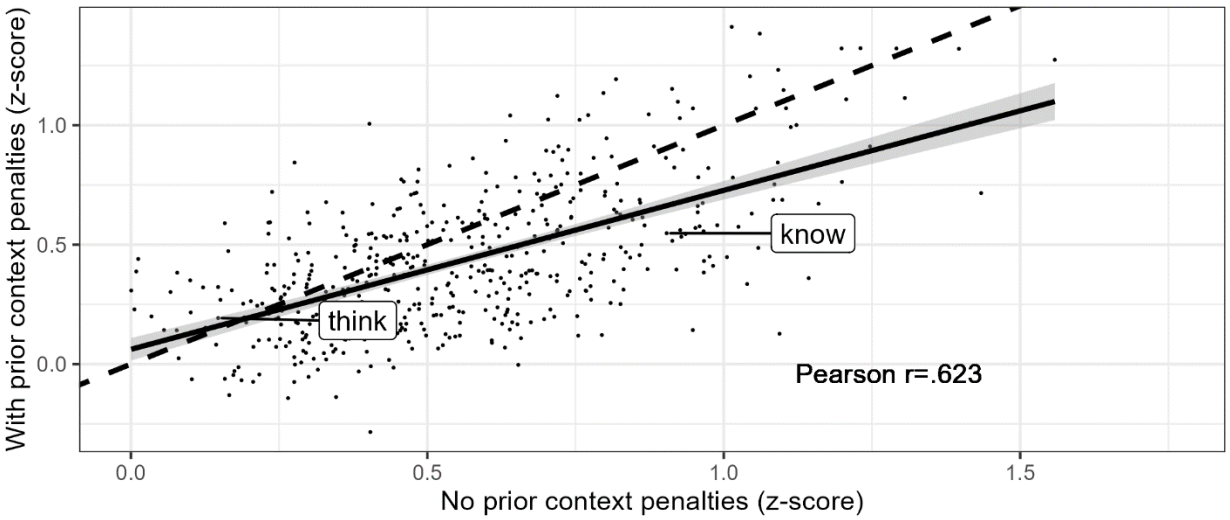
We suspect Liu and colleagues' null result (compared to our positive result) reflects a sample size difference: 48 verbs vs 484 verbs. To test this, we ran resampling simulations. We created 5,000 random samples of 48, 100, 200, 300, and 400 verbs, and fitted interaction models of z-scored 'no context' acceptability ratings for each set, crossing dependency length and log frame frequency. Each model contained by-participant and by-verb random intercepts (so that almost all models would converge). Figure 3 is a box-and-whisker plot of the *t*-values of the interaction term for each sample size (showing the median, 1st, and 3rd quartiles). What we see is that smaller sample sizes lead to smaller *t*-values: for a sample size of 48, 30% of the simulations had *t*-values below 2, which is often taken as a minimum threshold for statistical significance (e.g. Baayen et al. 2008).



Note: Only t-values of convergent models are included for analysis.
FIGURE 3. Distribution of t-values of interaction effect of frame frequency and dependency length on acceptability of wh-questions presented without prior context, for random samples of verbs of various sizes.

5.3. HOW DOES CONTEXT AFFECT BRIDGE EFFECTS? The finding that bridge effects (i.e., variation in penalties between verbs) exist for both the 'no context' and 'with context' data sets suggests that context, at least as operationalized in a dialogue format, does not eliminate bridge effects, contrary to what one might expect given reports in information structure accounts about context improving wh-extraction. That said, there is a small increase in the acceptability of long wh-dependencies and correspondingly a small decrease in the mean size of penalties between 'no context' and 'with context' (long wh-dependencies: -0.05 z-units vs. 0.13 z-units; penalties: 0.56 z-units vs. 0.44 z-units). This can be seen as consistent with previous reports that context improves extraction from the complement clauses of non-bridge verbs (e.g. Erteschik-Shir 1973, Ambridge & Goldberg 2008, Chaves & Putnam 2020), although the magnitudes suggest that the amelioration is much more modest for a full set of verbs.

Figure 4 corroborates these conclusions, by plotting penalty scores for each of the 484 verbs of interest, along with a line of best fit (solid) for the correlation between the two data sets, and a dashed line showing a hypothetical slope of 1.



Note: Each dot represents one verb; the solid line is the line of best fit and the dashed line has a slope of 1.

FIGURE 4. Correlation between long-distance penalties for wh-questions presented without and with prior context penalties.

We find a strong correlation ($r(482)=.623$, $p<.01$); if context eliminated bridge effects, the 'with context' penalties would be centered around the same mean (plus random variation) and therefore the correlation should be closer to 0. The fact that the slope of the line of best fit is below 1 indicates that 'with context' penalties tend to be smaller than 'no context' penalties.

Taken together, these observations suggest that context reduces penalties to some degree, but this reduction applies either across-the-board such that the overall variation in penalties between verbs remains unchanged, or disproportionately on certain verbs so that overall variation increases.

5.4. THE SIZE OF BRIDGE EFFECTS. Finally, though our results confirm that there are in fact bridge effects in English, and that they do not decrease with context, we note that the mean size of bridge effects is relatively small for all of the measures tested here: about 0.2 to 0.5 z-scores over the entire range of the measures for 'no context' penalties (Table 1). These effect sizes are smaller than typical island effects in English, which range from 0.6 to 1.2 z-scores in a recent review (Sprouse & Villata 2021). Given that bridge effects and island effects are sometimes analyzed as related locality phenomena (e.g. Erteschik-Shir 1973, Ambridge & Goldberg 2008, also Goldberg 2006, Chaves & Putnam 2020, and references cited therein), this effect size difference may be meaningful. We explore this possibility in Section 9.1.

6. EVALUATING THE SOURCE OF BRIDGE EFFECTS. Having established that bridge effects exist both without and with context, we can move to the central theoretical debate in the literature: What is the source of bridge effects? In order to do so, we constructed a set of linear regression models for each existing theory of bridge effects based on the approach in Ambridge & Goldberg 2008: the dependent variable is the mean penalty for each verb and the sole predictor is a measure of interest from each theory. As above, we repeat the analyses for 'no context' and 'with context' penalties.

We will report following aspects of each regression model:

1. The **regression slope**, which indicates the size and direction of the effect of the measure of interest. The regression slope should closely track the interaction coefficients reported in Section 5.
2. The **Bayes factor of each regression model** (BF_{10}). BF_{10} indicates the ratio of the likelihood of the data under the experimental hypothesis to the likelihood of the data under the null hypothesis. BFs greater than 3 are conventionally interpreted as meaningful evidence for the experimental hypothesis.
3. **R^2**, which indicates the proportion of the variance in penalties that is explained by the measure of interest (ranging from 0 to 1).
4. **R^2, corrected for attenuation**. The strength of the relationship between two variables is constrained by the reliability of each variable (Spearman 1904, Muchinsky 1996, etc.). This means that the lower the reliability, the lower the R^2s. We report both raw R^2s and R^2s corrected for this attenuation (using Spearman's method). We estimated reliability for acceptability judgments and backgroundedness measures using a bootstrap-based resampling simulation (see Appendix B for details). Because this resampling method is not feasible for the frequency and semantic similarity measures, we assume perfect reliability for them.

6.1. EVALUATING THE THREE THEORIES. Figure 5 plots the relationship between penalties and one representative predictor per theory, and Table 2 reports the regression results for the full set of predictors.

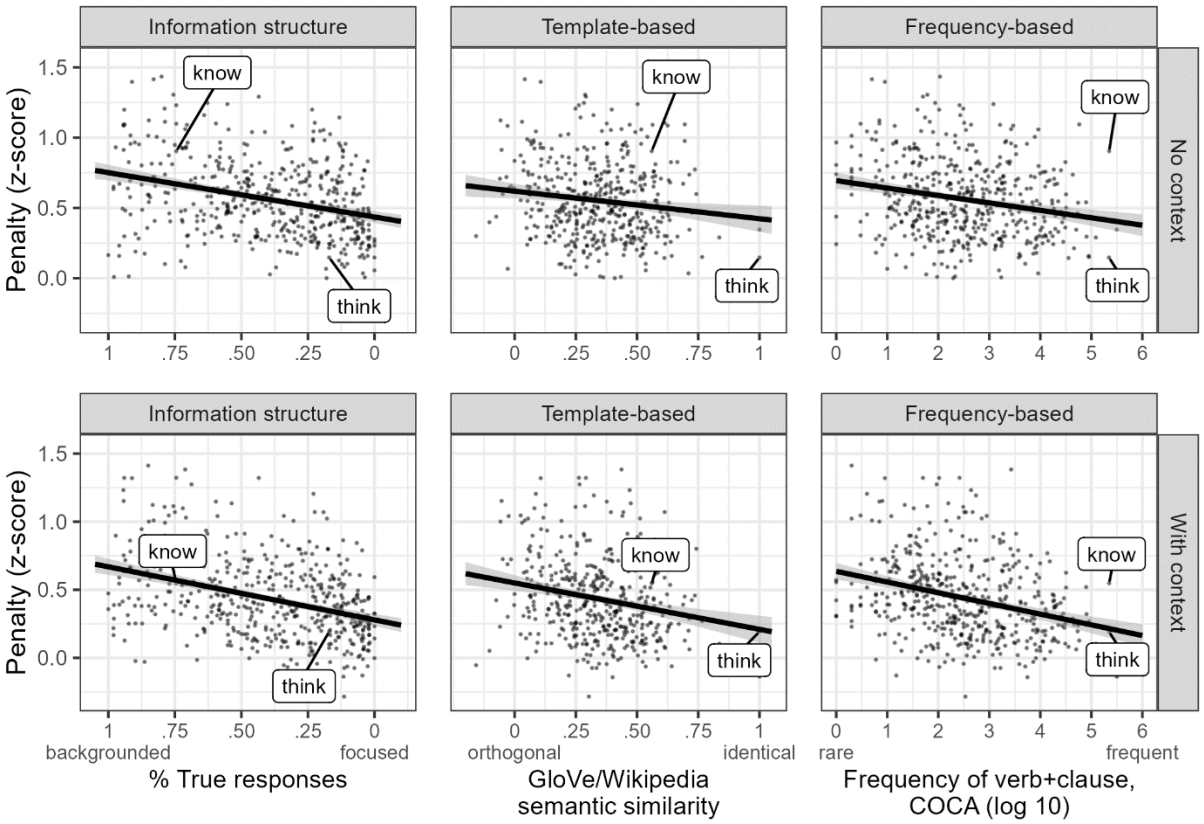


FIGURE 5. Correlations of long-distance penalties with selected predictors of information structure, template-based processing, and frequency.

	No prior context						With prior context					
	b (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2	*b* (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2
Information structure theory												
% True	0.32 (0.04)	7.41	0.32	>100	.103	.133	0.39 (0.05)	8.62	0.39	>100	.134	.186
% Not-False	0.74 (0.10)	7.22	0.51	>100	.098	.143	0.75 (0.11)	6.95	0.52	>100	.091	.143
Template-based processing theory (hybrid score)												
LSA/Wikipedia	0.11 (0.08)	1.44	0.10	0.1	.005	.006	0.08 (0.09)	0.95	0.07	0.1	.002	.003
GloVe/Wikipedia	-0.20 (0.07)	-2.88	0.23	2.9	.018	.023	-0.34 (0.07)	-4.58	0.39	>100	.045	.060
GloVe/Gigaword	-0.18 (0.07)	-2.65	0.20	1.6	.015	.019	-0.30 (0.07)	-4.06	0.33	>100	.036	.047
WordNet path-similarity (log)	-0.15 (0.07)	-2.32	0.17	0.7	.011	.014	-0.17 (0.07)	-2.38	0.19	0.8	.012	.016
Frequency-based processing theory												
Frame frequency (log)	-0.05 (0.01)	-4.92	0.30	>100	.048	.060	-0.08 (0.01)	-6.92	0.43	>100	.091	.121
Verb bias (log)	-0.12 (0.02)	-5.78	0.35	>100	.065	.081	-0.18 (0.02)	-8.70	0.55	>100	.137	.181

Note: The independent variable for each model is listed in the table; the dependent variable is long-distance penalty.

TABLE 2. Comparison of regression models of various theories of the source of bridge effects.

We first discuss the 'no context' data set, as previous experiments on bridge effects have focused on wh-questions presented without context.

Effect sizes are comparable to the results in Section 5 (as expected). Bayes Factors are high for all information structure and frequency predictors and marginally so for only one of the template-based processing predictors (GloVe/Wikipedia).

Turning to R^2s, we see that the template-based processing theory explains very little variance, about 0.5% to 1.8% before correction, and 0.6% to 2.3% after correction. This accords well with our Bayes Factors results and with Ambridge and Goldberg's (2008) findings for a sample of 12. Here we expand Ambridge and Goldberg's result to four similarity measures and 484 verbs, and confirm that semantic similarity is a relatively poor predictor of bridge effects.

The frequency-based processing theory fares only slightly better. Depending on the measure, frequency accounts for 4.8% or 6.5% of variance before correction, and 6.0% or 8.1% after correction.

The information structure theory explains the most variance—9.8% or 10.3%, depending on the measure, before correction, and 13.3% or 14.3% after correction. However, these values are substantially lower than the uncorrected R^2 of .69 that Ambridge and Goldberg (2008) observed for a sample of 12 verbs (note that this difference would likely be even more extreme if the Ambridge and Goldberg R^2 were corrected, which would likely increase the R^2). Though the interpretation of R^2 is subjective, given that Ambridge and Goldberg argued in favor of the information structure theory on the basis of explaining 69% of the variance, we suspect that 13.3% to 14.3% would be seen as relatively underwhelming evidence for the information structure theory.

We believe that Ambridge and Goldberg's substantially larger R^2 likely reflects the well-known fact that smaller samples with power issues—a possibility that they acknowledge of their study—can overestimate the effect size and hence r/R^2. To test this, we replicated their backgroundedness analysis by analyzing the same 12 verbs from our 'no context' data set. The uncorrected R^2 is .485, substantially higher than the R^2 for our full data set and much closer to Ambridge and Goldberg's R^2.

We turn next to the 'with context' results. Effect sizes and Bayes Factors are comparable to the 'no context' results. Uncorrected R^2s show a small increase, resulting in a similar range of variance accounted for: 0.2% to 13.7%. We see a larger increase in the corrected R^2 (because of slightly lower reliability in the smaller 'with context' data set). The template-based processing theory still performs relatively poorly, accounting for 0.3% to 6.0% of the variance (post-correction). Both the frequency-based and information structure theories perform better: 12.1% and 18.1% of the variance for frequency, and 18.6% and 14.3% for information structure. This remains markedly lower than the Ambridge and Goldberg benchmark, and to our minds, lower than the field would expect for a good theory of bridge effects.

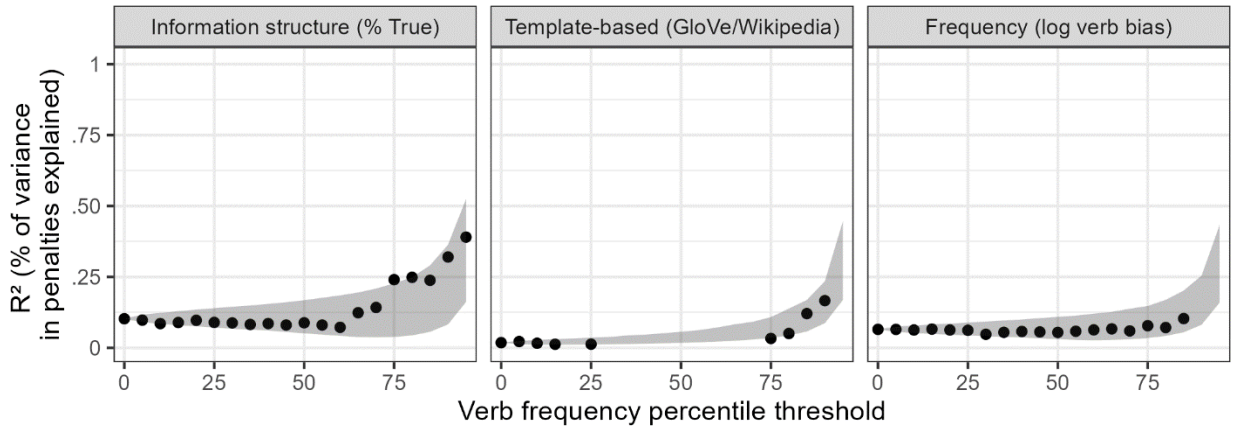
6.2. ARE R^2S LOW BECAUSE OF NOISE FROM INFREQUENT VERBS? In testing a nearly exhaustive set of verbs, we necessarily included many low-frequency verbs. One possibility is that these verbs were unfamiliar to participants, and so penalty scores for these verbs would be unreliable, which could distort model outcomes and reduce R^2.

We think such a scenario is unlikely. First, as noted in Section 3, we only analyzed responses from participants who demonstrated familiarity with the verbs on the synonym post-test.

Second, we find no evidence that penalty scores for infrequent verbs are more variable (unreliable) than frequent verbs. To test this, we calculated the standard deviation of 'no context' penalties for each verb, obtained the frequency of each verb in COCA (regardless of the presence of complement clauses), and calculated a correlation. We did not see a significant (negative) correlation between standard deviation and frequency ($r(482)=-.036$, $p=.43$).

Third, we explored how much R^2 would improve if we excluded infrequent verbs. We sorted the verbs into twenty bins based on their COCA frequency. For the best-performing predictors of each theory—%True responses (information structure), GloVe/Wikipedia similarity (template-based), and log verb bias (frequency-based)—we ran twenty simple linear regression models, incrementally leaving out the low-frequency bins each time, i.e. first fitting data for all verbs, and then for verbs above the 5th, 10th, 15th, etc., percentiles. Figure 6 plots how R^2 changes with this frequency threshold. These values are the points in the plot. There are marked improvements in R^2 but only when we analyze the most frequent 40% of verbs (or higher).

However, there is also a general tendency for R^2 ranges to widen (often producing larger estimates) as sample sizes shrink. To account for this, we ran a Monte Carlo simulation in which verbs were randomly sorted into 20 bins to mimic the percentile analysis above, repeated 5,000 times. We calculated 95% intervals of R^2 based on this simulation, and plotted them as the gray bands in Figure 6. As this figure illustrates, the R^2 increase observed when we restrict the analysis to more frequent verbs almost always lies within these 95% intervals. This suggests that the increase is not uniquely attributable to verb frequency, but also attributable to sample size.



Note: Plots show R^2s for subsets of clause-embedding verbs above selected frequency percentiles, for best-performing predictors of 'no context' bridge effects, with 95% intervals based on the random assignment of verbs into 20 bins. A model's R^2s is excluded if the estimate of the effect is insignificant or in the wrong direction.

FIGURE 6. Results of simulations to see how much better bridge effects can be accounted for by each theory, if less frequent verbs were excluded from analysis.

These analyses suggest that the underperformance of existing theories cannot be meaningfully attributed to the many low-frequency verbs in our data set.

6.3. TAKEAWAYS FOR THE DEBATE ON THE SOURCE OF BRIDGE EFFECTS. In this section, we leveraged our new data sets to attempt to resolve the central debate in this literature: what is the source of bridge effects? We found that the information structure-based theory of bridge effects performs slightly better than the frequency-based and template-based processing theories. However, we also identified two challenges. First, none of these theories provide particularly strong fits to the full set of verbs. Second, adding context had little impact on effect sizes and slightly increased model fits for existing theories. This second outcome presents complications for all three theories. Frequency-based and template-based processing theories are typically silent about the role of context, implying that prior context should cause no change in bridge effects, and therefore no change in the correlation with predictor measures or R^2s. Information structure theories, on the other hand, report that context can decrease penalties for at least some verbs, inviting the inference that bridge effects might decrease or even disappear given a supportive context.

These findings suggest that research on bridge effects could benefit by considering a wider range of theories beyond these three single-predictor theories. In the next section, we attempt to do just that.

7. A THEORETICAL PATH FORWARD: MORPHOSYNTACTIC LICENSING COMBINED WITH SEMANTIC/PRAGMATIC CONSTRAINTS AND PROCESSING COSTS. Our initial goal for this study was confirmatory: to test the empirical predictions of existing theories of bridge effects on a nearly exhaustive set of verbs. However, our results suggest that none of the theories perform particularly well. Therefore, in this section we add an exploratory goal: to find a new theoretical path forward based on the evidence made available in the new data sets. We do this in two steps. The first is to leverage our data sets to identify a new theoretically-relevant predictor for bridge effects. The second step is to integrate this new predictor with existing predictors in a theoretically-consistent way. More specifically, we draw upon suggestions from leading work on bridge effects (in particular Erteschik-Shir 1973, Richter & Chaves 2020, and Chaves & Putnam 2020; cf. a similar approach by Bresnan et al. 2007 for the dative alternation), and propose a multivariate, layered account in which certain clause-embedding verbs can license long-distance wh-extraction, via subcategorization; in addition, the information structure, semantic, and frequency properties of verbs can further influence the acceptability of a wh-dependency. We show that this layered view of bridge effects delivers a substantially improved fit of our data compared to the various single-predictor theories that we have been considering so far, even after accounting for complexity. This finding provides new empirical evidence in favor of a multivariate approach toward bridge effects and also suggests that syntax has an important role to play in this multivariate approach.

7.1. MORPHOSYNTACTIC LICENSING. Our new data sets are available for all researchers to search for new potential predictors for bridge effects. In our own inspection of penalty scores, we noticed that verbs with low penalty scores tend to be verbs that allow both a finite clausal complement and some kind of nonfinite complement that has a close paraphrase that is syntactically finite. Examples of nonfinite frames are listed in 18.

(18) a. Jo claimed to have left. (cf. *Jo claimed that she had left.*)
b. Jo decided to leave. (cf. *Jo decided that she would leave*.)
c. Jo required them to leave. (cf. *Jo required that they leave*.)
d. Jo believed/expected them to have left. (cf. *Jo believed/expected that they left.*)
e. They were said/thought to have left. (cf. *It was said/thought that they left.*)
f. Jo saw them leave. (cf. *Jo saw that they left*.)
g. Jo declared them the winners. (cf. *Jo declared that they were the winners*.)
h. Jo announced them as the winners. (cf. *Jo announced that they were the winners*.)

To our knowledge, this correlation has not previously been noted. To explore this further, we annotated our full list of verbs with subcategorization information (based on Levin 1993 and our own judgments). Figure 7 shows the distribution of penalties based on subcategorization for nonfinite complements.

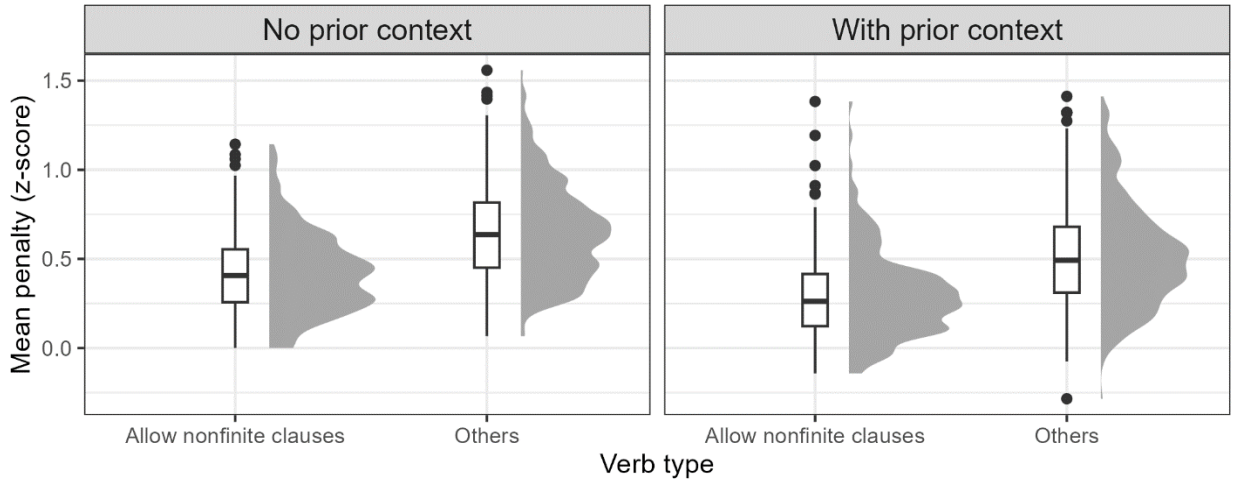


FIGURE 7. Boxplots and density plots depicting the distribution of long-distance penalties of verbs allowing/not allowing nonfinite complement clauses.

We first calculated a simple linear regression to predict penalties based on nonfinite complementation. Though nonfinite complementation is a categorical predictor, the model fit as indicated by R^2 is higher (.159) than the existing theories for the 'no context' penalty scores and relatively similar (.127) to the best-performing existing theories for the 'with context' penalty scores. Figure 8 illustrates this by plotting the uncorrected R^2 values for the best-performing predictors of the existing theories and nonfinite complementation.

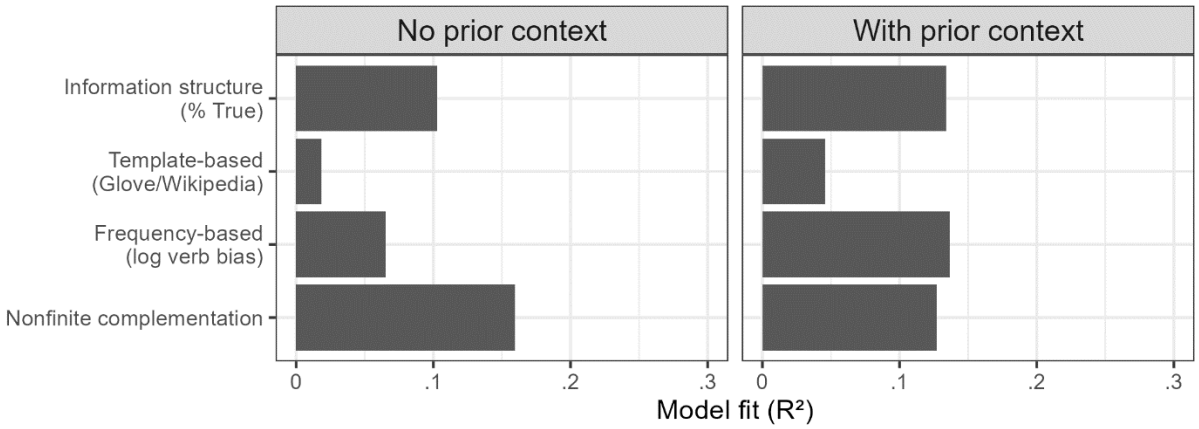


FIGURE 8. Uncorrected model fits for best-performing predictors of current theories of bridge effects and nonfinite complementation.

We take this as a potentially interesting research direction. As reviewed in Section 2, existing theories of bridge effects have mostly focused on non-syntactic factors under the assumption that the syntactic structure of all finite-clause embedding verbs is identical (with the exception of manner-of-speaking and/or factive verbs, e.g., Stowell 1981, Snyder 1992, Kastner 2015, Stoica 2016, de Cuba 2018, among others; and also see Kiparsky & Kiparsky 1970 for a complex NP analysis of factive verbs, although they explicitly reject this analysis). Our analysis here suggests it may be profitable to explore a syntactic difference across a wider range of verbs.

The question, of course, is how subcategorization for nonfinite complements could be relevant for extraction from finite complements. As one can see from 18, there is substantial diversity in the syntax of nonfinite complements and verb semantics: verbs in 18a-18c are typically labeled as control verbs; 18d as exceptional case marking (ECM) verbs, which mark the clause's subject in the accusative case; 18e can be seen as a variant of ECM, except that the verb can only mark the clause's subject in the nominative when the verb is passivized;[3] 18f are perception verbs; 18g and 18h are what Levin (1993:180) calls 'verbs with predicative complements,' which are used to 'characterize … properties of entities.' Despite this diversity, these various frames share common morphosyntactic properties. As noted above, the complement is nonfinite. Additionally, the subject of the complement stands in a structural relation with an element outside of the complement: in 18a-18c the (null) subject is bound (controlled) by an argument of the matrix verb, while in 18d-18h the subject is in the accusative (or nominative), in effect the object (or subject) of the matrix clause.

One promising direction, therefore, is to connect these properties to recent cross-linguistic research on A-dependencies crossing finite clause boundaries, like exceptional case marking (ECM) and indexical shift (e.g., Wurmbrand 2019, also 2018). Briefly, in these dependencies, the subject of a verb's (finite) complement clause behaves as if it were syntactically related to the main clause. Based on a cross-linguistic survey, Wurmbrand suggests

[3] These verbs are sometimes labeled as *wager*-class verbs (Postal 1974, among others). We do not use this label here. Reed (2023) notes that the membership of *wager*-class verbs is poorly understood and further argues that this class can be treated as special cases of ECM verbs.

that this is because the subject comes to occupy a special position in the complement clause's left periphery, which is high enough for the subject to enter into dependencies with elements in the main clause. In ECM, for instance, the subject is structurally high enough to get case-marked by the main clause's verb.

While Wurmbrand (2019) does not give an analysis of English ECM or other nonfinite clauses, extending her proposal to them seems quite feasible. Specifically, suppose in English, these nonfinite complements also have the same special position in the left periphery; for convenience, we label these complements as 'XP' in 19, instead of identifying them with any particular syntactic projection. For the case-marking examples 18d-18h, we can basically adopt Wurmbrand's analysis: the subject moves to this left periphery position for case-marking, as illustrated in 19a. In the control cases in 18a-18c, one option, following Landau 2015, is that the null subject PRO moves to the left periphery position for binding purposes, as shown in 19b. This movement might be semantically motivated, as the correlate of λ-abstraction, which allows the complement to be interpreted as a predicate (but see Landau 2015 for a more nuanced analysis).

To explain bridge effects, suppose that the same verbs allowing these nonfinite complements also require their finite clausal complements to have a special position in the left periphery. However, in this case, instead of allowing subjects to enter cross-clausal binding or case dependencies, this position is instead exploited for cross-clausal wh-dependencies, i.e., license further extraction of a wh-phrase to the matrix clause, as in 19c.

(19) a. Jo expects [$_{XP}$ them$_1$ [$_{TP}$ __$_1$ to win]].
b. Jo expects [$_{XP}$ PRO$_1$ [$_{TP}$ __$_1$ to win]].
c. What$_1$ did Jo expect [$_{XP}$ __$_1$ that [$_{TP}$ they would win __$_1$]]?

Put differently, what we are suggesting here can be seen as an adaptation and refinement of the classic 'escape hatch' analysis of Chomsky 1973 (among many others). In our analysis, some, but not all, English clause-embedding verbs allow their complement clauses to contain escape hatches, which are necessary for licensing cross-clausal case or binding dependencies (if non-finite) or wh-dependencies (if finite) (see Kim & Goodall 2022 for a recent proposal where long-distance extraction is also seen as a special case). Our adaptation of Wurmbrand's analysis also treats wh-dependencies on par with exceptional case marking and indexical shift, suggesting that languages might vary as to which of these dependencies (if any) are allowed to cross finite complement clauses. There are undoubtedly additional predictions and implications that could be explored in future work. But for the next subsection, we will focus on embedding this morphosyntactic licensing property within a theory that can better predict bridge effects.

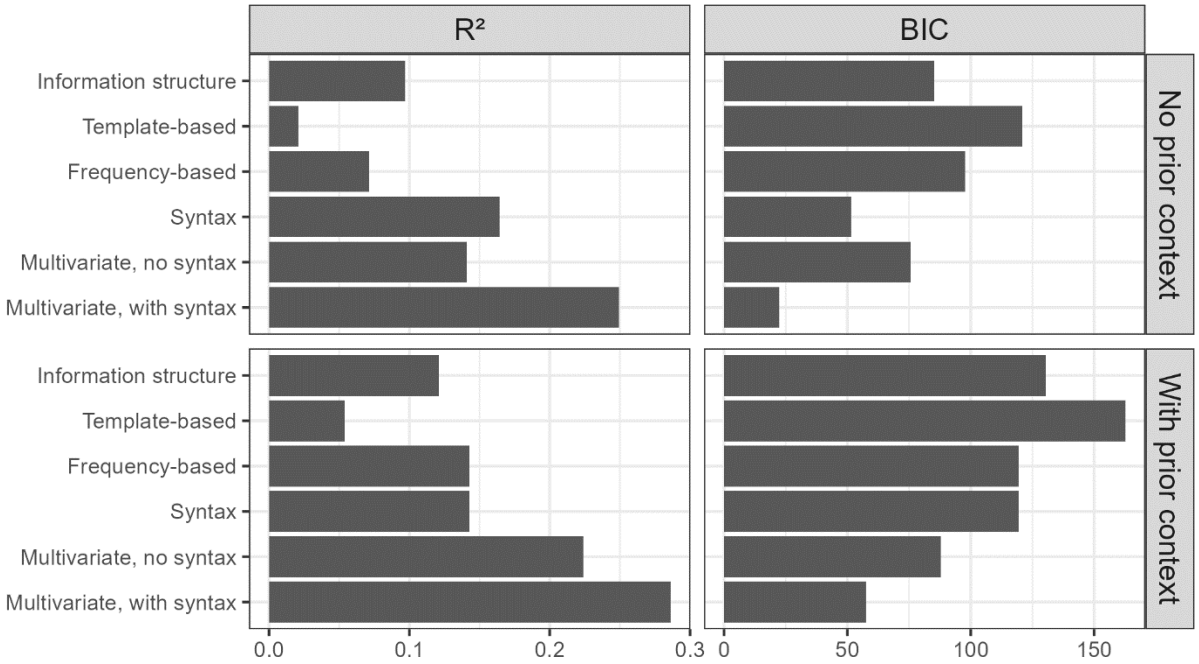
7.2. ADDING SEMANTIC/PRAGMATIC CONSTRAINTS AND PROCESSING COSTS. Nonfinite complementation alone provides at best a slightly improved fit for bridge effects over the other single-source theories, and so it is unlikely to constitute a complete theory of bridge effects by itself. But it could be part of a layered theory in which extraction is licensed by morphosyntax (e.g., the availability of escape hatches), but is additionally constrained by the information structure constraints and processing costs proposed in existing theories. In other words, a layered theory can help explain exceptions to our generalization involving nonfinite complementation and gradience in our data: because of differences in information structure properties, processing costs, etc., penalty scores for some verbs allowing nonfinite complements are higher than expected, while those for some verbs that disallow them are lower than expected. Such a theory is also uncontroversial, in that theories that propose a syntactic component to licensing long-distance dependencies are usually compatible with the assumption that other factors, such as information structure and processing complexity, continue to impact acceptability (though, the reverse may not be true—some theories of information structure or processing complexity may eliminate the need for a syntactic component to explain acceptability). To illustrate this, consider *know*. This is a verb that allows nonfinite complements, as illustrated in 20, but has a relatively high penalty, as is well-documented in the literature. In our layered theory, we would claim that extraction from *know* is licensed from the angle of morphosyntax, but not information structure, since *know* is factive and hence backgrounds its complement clause.

(20) Jo knows <u>there to be several problems</u>.

Furthermore, we see this approach as drawing on several prior suggestions from the bridge effects literature. First, in order to explain cross-linguistic variation in extraction, Erteschik-Shir (1973) proposes that there is a class of 'potential bridges' based on information structure, and that a subset of those become actual bridges through the acquisition process. A theory that combines morphosyntactic licensing and other constraints shares the same hierarchical arrangement (albeit working with different theoretical primitives), and provides similar flexibility to capture cross-linguistic variation (though, perhaps moving it to the syntactic component). Second, after concluding that the frequency-based theory is inadequate, Richter and Chaves (2020) suggest that combining semantic and pragmatic factors could provide a better explanation for bridge effects. Here we, too, are suggesting combining multiple types of factors. Finally, Chaves and Putnam (2020) argue that locality in general (encompassing both bridge and island effects) might best be explained with an 'eclectic' theory that draws on syntactic, semantic/pragmatic, and processing factors. What we do here is to develop a specific version of this kind of approach.

To empirically evaluate this layered theory, we created six regression models that instantiate different theoretically-driven possibilities, comparing both R^2s (uncorrected for attenuation) and Bayesian Information Criterion (BIC), which balances data fit with a penalty for increased complexity (the lower the BIC, the better) (Figure 9). We include four single-predictor theories as baselines: information structure, template-based, frequency, and syntax (nonfinite complementation). To facilitate comparison, we selected the best-performing predictor for each non-syntactic theory and a set of 439 verbs that have complete predictor information for these theories. We then consider two theories that combine predictors: a model that combines only the three non-syntactic predictors, which we call *multivariate, no syntax*, and a model that combines all four predictors (including nonfinite complementation), which we call *multivariate, with*

syntax. Both multivariate models are additive, meaning there are no interaction terms in the models. We believe this is consistent with existing theories: each component is independent of the others.[4]



Note: The higher the R^2 and the lower the BIC, the better the fit.
FIGURE 9. Model fits for single-predictor and multivariate models of bridge effects.

For both 'no context' and 'with context' penalties, the 'multivariate, with syntax' model has the highest R^2s of all the models under consideration, including the 'multivariate, no syntax' model. Crucially, the R^2s are substantially higher, in the .25-.3 range, or about 2.5 times the R^2s for the best-performing single-predictor theory in the literature (namely, information structure). Furthermore, even though the 'multivariate, with syntax' model is the most complex of the various models, it has the lowest BIC values, confirming that it achieves the best coverage of the data, even after controlling for complexity.

There are potentially interesting orderings among the other theories, such as the fact that the syntax-only model outperforms the 'multivariate, no syntax' model for 'no context', and that R^2s for the 'multivariate, no syntax' model increases substantially for 'with context' (a point we consider in section 9.2). But we take the primary result of this analysis to be that a layered theory that features syntactic licensing in addition to information structure constraints and processing costs, as proposed here, yields substantially better model fits than single-predictor theories or a multivariate non-syntactic theory, even after taking into consideration its relative complexity. That said, it is an open question whether our results satisfy the field's conception of a good theory of bridge effects (cf. Ambridge and Goldberg's R^2 of .69, albeit for an underpowered 12-

[4] Out of an abundance of caution, we also ran the full interaction models. Interaction models always have higher BICs than the additive models, suggesting that the increased fit of the interaction models (between .01 to .05 in R^2) is outweighed by their greater complexity.

verb sample). We note that our R^2 values are an underestimate; actual R^2s will almost certainly be higher after correcting for attenuation (but we do not do that here, because to our knowledge, there is no correction formula applicable for models with multiple predictors). It may also well be the case that there are other components in a layered theory that we have not yet uncovered. But we hope that this illustrates a promising path forward based on the new information in our data sets.

8. POTENTIAL CONCERNS. In the review process, two anonymous reviewers provided helpful comments on our experiments and analyses that might be of interest to readers. We present a brief discussion here.

8.1. A COMPETITION-BASED APPROACH BASED ON NUMBER OF SUBCATEGORIZATION FRAMES. One reviewer suggested an interesting counterproposal that could be potentially evaluated with our data sets: perhaps bridge verbs are those with fewer competing (non-clausal) subcategorization frames, and non-bridge verbs are those with more competing (non-clausal) subcategorication frames. In other words, the more subcategorization frames allowed, the larger the bridge penalty.

To test this proposal, we used White and Rawlins' (2020) publicly available MegaAcceptability data set. White and Rawlins combined clause-embedding verbs each with 50 different subcategorization frames (including clauses and others) and collected acceptability ratings for each verb-frame combination. For each verb, we counted the number of frames whose normalized acceptability, as computed by White and Rawlins, is greater than a given acceptability threshold, as a proxy of the number of subcategorization frames the verb allows. For ease of reference, we will call this the verb's 'frame diversity'.

As a first analysis, we only considered non-clausal frames (frames without an embedded S or VP, in MegaAcceptability terms). Of our 484 verbs of interest, 415 were present in the MegaAcceptability data set. Our analyses suggest that the greater the frame diversity, the smaller the penalty. Setting the acceptability threshold to 0, the Pearson correlation between non-clausal frame diversity and no-context penalties is -.18 (p<.01), implying a very low R^2 of .03. Raising the acceptability threshold to 0.5 produces a correlation of -.10 (p=.04), and raising the threshold to 1 produces a nonsignificant correlation of -.01 (p=.80).

For comprehensiveness, we tested another version that considers all frames, including clausal ones. This yielded very similar results. Setting the acceptability threshold to 0 produces a Pearson correlation of -.11 (p=.02), or an R^2 of .01. Raising the threshold to 0.5 produces a correlation of -.10 (p=.04), as does raising the threshold to 1. Therefore, although this is a potentially interesting approach to bridge effects, we conclude that there is no clear evidence for it in our data sets. But it again illustrates the potential value of our data sets for additional theorizing.

8.2. PRIMING IN THE CONTEXT EXPERIMENTS. One reviewer raises a potential concern about the context experiment, where a context sentence precedes the target sentence to be judged for acceptability. In the short wh-dependency condition, illustrated in 21, the embedded clauses in both context and target sentences are identical, but that is not the case in the long wh-dependency condition, illustrated in 22, because of wh-extraction. The context sentence therefore might have primed the target sentence, and boosted acceptability ratings, more in the short condition than in the long condition (see e.g. Luka & Barsalou 2005 for further discussion of this priming effect). If so, this would have increased estimates of penalty sizes (defined as short ratings – long ratings) for the 'with context' data set.

(21) A: Someone thought that the duchess would invite the arrogant knight.
B: Really? Who thought that the duchess would invite the arrogant knight?

(22) A: The princess thought that the duchess would invite a certain person.
B: Really? Who did the princess think that the duchess would invite?

An across-the-board increase in penalties for all verbs would not be a problem, because our analysis defines bridge effects as a difference in penalties between verbs (i.e., an interaction of dependency length and verb). It would only be a problem if the increase targeted only a subset of verbs, such as verbs that independently disprefer clause-embedding, because such an increase would inflate the average size of 'with context' bridge effects. Such an increase could arise either from a priming mechanism that targets low-acceptability constructions, or from ceiling effects: the short condition for verbs that are highly compatible with clausal complements might be judged as highly acceptable even without prior context, and therefore cannot benefit as much from priming.

We agree with the reviewer that the differences between the 'no context' and 'with context' data sets need to be kept in mind while testing theories using both data sets. However, we believe that the second priming scenario, in which context increased penalties for a subset of verbs, is unlikely.

First, the analysis in section 5.3 presents potential evidence against this priming scenario: 'with context' penalty sizes were on average slightly smaller than 'no context' penalty sizes. The analysis in section 7 also presents potential evidence against this: we found qualitatively similar model fits for the two data sets. That said, we still believe it is worth checking for this effect given that we wish both the 'no context' and 'with context' data sets to be useful for theorists exploring new theories of bridge effects.

To evaluate this possible concern, for each of our 484 verbs of interest, we sorted the 'no context' short condition acceptability (z-scored) from lowest to highest into deciles, in order to reflect how much the verbs disprefer (low deciles) or prefer (high deciles) clausal complements. For each of these deciles, we calculated the effect of context as the median difference between the two data sets, so that we can see whether context boosted acceptability ratings more in the short condition than in the long condition, as expected under this priming scenario. The median boost due to context for each decile are presented in Figure 10.

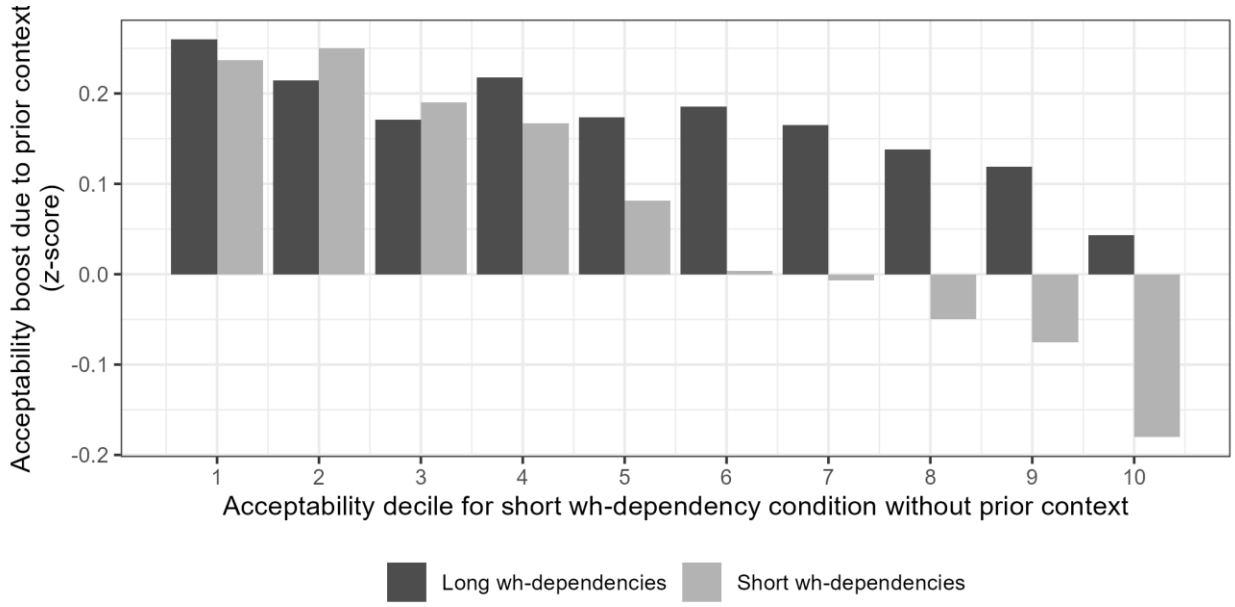


FIGURE 10. Median boost in z-scored acceptability due to prior context for short and long wh-dependencies (i.e. 'with context' ratings – 'no context' ratings).

As Figure 10 shows, adding context did boost acceptability for short wh-dependencies, especially for lower-decile verbs (1-5), which disprefer clausal complements, as suggested by the reviewer. But, contrary to the concern, we find very similar boosts for long wh-dependencies in the same lower deciles.

As for verbs that are more compatible with clausal complement (deciles 6-10), context boosted the acceptability of the long condition more than the short condition, entailing smaller penalties; in fact, context tended to lower acceptability of the short condition for these deciles. The net effect aligns with the result that we discussed in section 5.3, wherein 'with context' penalties are on-average slightly smaller than 'no context' ones. One possibility suggested by this finding is that this could be a ceiling effect: short wh-dependencies were more acceptable in the first place, so the effect of context had more room to improve the long wh-dependencies. Crucially, though, this effect is again inconsistent with the priming scenario, which predicts a larger boost for the short condition than for the long condition.

8.3. USING SHORT WH-DEPENDENCIES WHEN CALCULATING THE BASELINE ACCEPTABILITY OF CLAUSAL COMPLEMENTS FOR EACH VERB. One reviewer raises a potential concern about the use of short (matrix) wh-dependencies in the baseline condition and how penalties are calculated (see Section 2.1). We welcome the opportunity to discuss this because it was an intentional choice that we made that departs from Liu et al. 2022 and Ambridge & Goldberg 2008 (among others), which used declaratives in their baseline condition (versus long wh-dependencies in the target condition). We chose to use short wh-dependencies in the baseline condition to guard against the possibility that certain verbs might be resistant to wh-questions in general, regardless of whether it is matrix extraction or extraction from the complement clause. If that were the case, then a penalty calculated using a declarative baseline would be confounded with the wh-question-resistance effect, potentially inflating penalties for some verbs. The result would look just like bridge effects in our analysis, but would not be true bridge effects. By using wh-dependencies in both conditions, any wh-question-resistance effect will be subtracted out. (We are assuming that the wh-question-resistance effect impacts both short and long wh-dependencies uniformly. If it affects long wh-dependencies more, there is no way to disentangle that from a bridge effect, regardless of the baseline condition.)

The same reviewer also notes that using short wh-dependencies in the baseline condition means that the short condition might not be a clean measure of how acceptable each verb is with clausal complements, with potential complications for which verbs to exclude from analysis (Section 3.2). At the very least, there is a well-known acceptability decrease for wh-questions compared to declaratives. It is possible that this effect due to wh-questions might linearly sum with the presence of a clausal complement, in which case acceptability ratings for the short condition do not truly reflect whether a verb allows clausal complements. If so, it would not be ideal for us to have used short condition ratings (specifically, whether a verb has a negative z-scored rating in that condition) to determine which verbs to exclude from analysis as not allowing such complements. However, we believe that the benefit of having a clean penalty score outweighs the cost of not having a clean measure of clause embedding. That is why we chose this condition.

In retrospect, we could have attempted to quantify a wh-resistance effect, and had an independent measure of the acceptability of clausal complements, if we tested all three conditions—a declarative, a short wh-dependency, and a long wh-dependency. Unfortunately, we cannot re-run these experiments for financial reasons. But we can compare our short condition results to White and Rawlins' publicly available MegaAcceptability data set to determine if our short condition ratings are good estimates of whether a verb allows finite clausal complements. As mentioned earlier, White and Rawlins collected acceptability judgments for a large set of clause-embedding verbs occurring with finite clausal complements (and other complements), but importantly, their materials did not involve any wh-extraction of the verb's subject. If there is a relatively uniform effect of wh-questions across verbs, we expect to find a relatively large positive correlation between their ratings and ours. That will not preclude that our ratings might be artificially low, and therefore we excluded more verbs from our analyses than we should have. But it at least would show that our use of wh-questions did not confound the penalty calculations.

Therefore, for each verb, we identified the frame in MegaAcceptability that was closest in syntax and semantics to ours (whether there is an indirect object, or whether the complement

clause contains a future modal, etc.). The two studies have 503 verbs in common. We then correlated our 'no context' short ratings from the no-context data set with the normalized acceptability as reported by White and Rawlins for that verb and frame. We found large positive Pearson correlations for all 503 verbs ($r(501)=.63$, $p<.01$) and for the 484 verbs of interest, of which 415 are present in MegaAcceptability ($r(413)=.51$, $p<.01$). This implies that if matrix wh-extraction has an impact on the acceptability of clausal embedding, the effect is relatively uniform across verbs. (Because White and Rawlins used semantically bleached materials, and because their normalized ratings are based on ordinal regression, we can't conclusively determine if our ratings are lower than theirs, as expected given the effect of wh-questions on acceptability. But we suspect that to be the case.)

Yet another way to address the reviewer's concern about the exclusion of verbs is to redo our analyses to include more verbs. As described in more detail in Appendix D, we tried out two different verb exclusion criteria, (i) relaxing the no-context short wh-dependency acceptability threshold to -0.25, yielding a set of 488 verbs, and (ii) eliminating it altogether, yielding a set of 536 verbs. We then repeated the model fit analysis reported in Section 7.2 Figure 9, where we compare single-factor models of bridge effects with multivariate models with or without nonfinite complementation included as a predictor. Results of these analyses, which are reported in Appendix D, are very similar to those in Section 7.2, suggesting that our conclusions for Section 7.2 (and preceding sections) are not sensitive to our verb exclusion criterion.

9. IMPLICATIONS BEYOND BRIDGE EFFECTS.

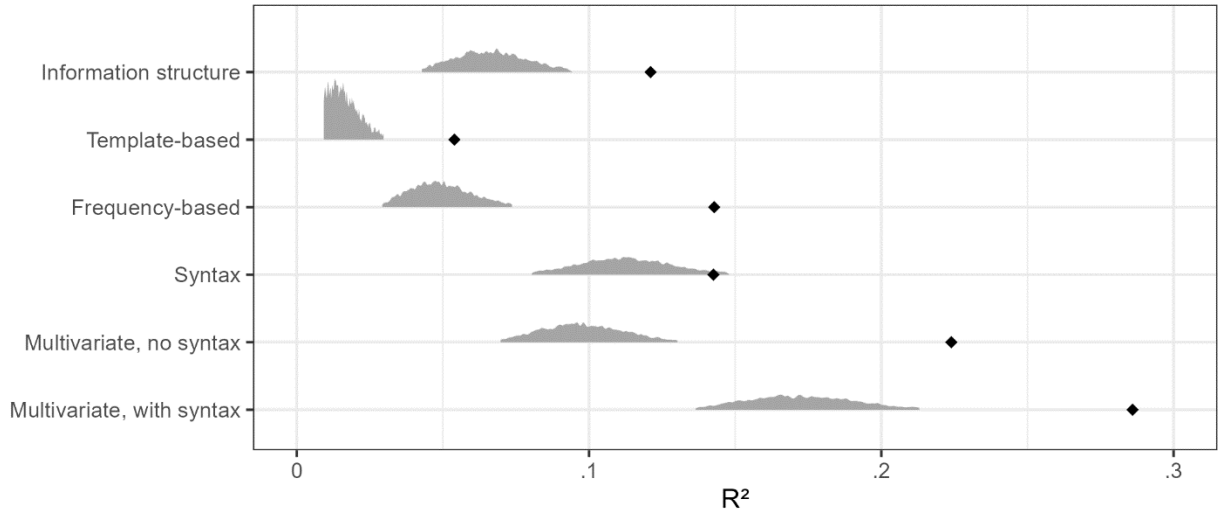
9.1. THE UNIFICATION OF BRIDGE AND ISLAND EFFECTS IN THE THEORY OF LOCALITY. Given that bridge effects are a type of locality phenomenon, one question that arises in the literature is whether bridge effects and islands effects can be unified under a single analysis, as often proposed in information structure theories (e.g. Erteschik-Shir 1973, Ambridge & Goldberg 2008, and references cited therein; cf. Chaves & Putnam 2020). However, our results raise a novel empirical challenge for the unification of bridge effects and island effects: as Sections 5 and 6 noted, the effect sizes of bridge effects are substantially smaller than the effect sizes of island effects. For example, the maximum impact of backgroundedness on penalties—going from not backgrounded at all to totally backgrounded—is about 0.2-0.5 z-units depending on the measure and the data set. In contrast, various island effects in English that have been tested using similar experimental methods appear to have effect sizes of 0.6-1.2 z-units (see Sprouse & Villata 2021 for a review). This effect size difference suggests either that bridge effects should be treated as distinct from island effects, or that theories seeking to unify them, whether rooted in information structure or otherwise, need to include an additional layer of complexity to explain the differing effect sizes.

We are personally inclined to interpret the difference in effect sizes between bridge and island effects as reflecting different sources. That said, there is a three-way distinction in Erteschik-Shir's (1973) theory that could potentially be adopted for a unified non-syntactic account in order to explain the effect sizes we observe. As alluded to in section 7.2, to explain cross-linguistic differences between Danish and English, Erteschik-Shir (1973:125ff) suggests that there is a set of 'potential bridges', which are presumably universal, while usage within a language determines which potential bridges become actual bridges. This analysis implies at

least three different classes of constituents—non-bridges, potential bridges that do not become actual bridges, and potential bridges that do. We note that this three-way distinction could be exploited to explain the observed effect sizes (although not necessarily consistent with Erteschik-Shir's information structure proposal): island effects might correspond to extraction from non-bridges, while the variation found in bridge effects might correspond to extraction from potential bridges that do not become actual bridges (higher penalties) as well as extraction from actual bridges (lower penalties). Researchers interested in this approach could use our backgroundedness and acceptability data sets, perhaps by combining them with island effects data sets, or by collecting similar data sets in other languages to quantify the cross-linguistic variation in bridge effects.

9.2. THE EFFECT OF CONTEXT. We originally collected 'with context' penalties in response to reports in information structure theories that supportive context can make long-distance wh-extraction more acceptable, which raise the possibility that context can decrease or eliminate bridge effects. Though we found no such impact on bridge effects in Section 5, we did observe a novel effect of context in the analyses in Sections 6 and 7: model fits (R^2s) are higher for 'with context' penalties for information structure, template-based, frequency-based, and multivariate theories (but not for nonfinite complementation when it is the sole predictor). These higher R^2s warrant further exploration, both for what they could mean for theories of bridge effects, and for what they could mean for a more general theory of the effect of context on the acceptability of long-distance dependencies.

The first question we can ask is whether the pattern that we observed—the higher R^2s—is meaningful: whether it is beyond what we would expect due to sampling error between the two data sets. Sampling error is a particularly plausible explanation because our 'with context' data set is about half the size of the 'no context' data set in terms of the number of observations per verb (due to the addition of catch trials to ensure the context was read). To rule out this possibility, we ran a bootstrap analysis to determine if the 'with context' R^2s are more extreme than we would expect based on the 'no context' data set. We randomly sampled participant responses with replacement from the 'no context' data set such that the size of the random sample for each verb is equal to the actual set of 'with context' observations for that verb. We then calculated penalty scores from these random samples and fitted the same regression models as in Section 7.2. We repeated the process 5,000 times to generate expected distributions of R^2 values under repeated sampling. We then constructed a confidence interval with the 2.5^{th} and 97.5^{th} percentiles of each distribution. Observed 'with context' R^2s are higher than the upper bound of this interval for all but the syntax-only model (Figure 11), indicating that higher R^2s for 'with context' data are unlikely a sampling artifact.



Note: Distributions are those of simulated R^2s based on 'no context' penalties (excluding values beyond 2.5th and 97.5th percentiles); diamonds represent observed 'with context' R^2s.
FIGURE 11. Results of simulations to see whether R^2s for 'with context' penalties are more extreme than what is expected from 'no context' penalties.

Given that the increase in R^2 appears to be meaningful, we can next ask why context improves the fit of several (non-syntactic) predictors. One possibility is that there are actually two effects combined in the 'no context' penalties: a non-syntactic (information structure, semantic, or frequency) effect, as hypothesized in existing theories, and some additional pragmatic effect. Adding supportive context eliminates or reduces this second effect, providing a better estimate of bridge effects, and thus allowing the non-syntactic factors to perform better. Identifying this second effect is beyond the scope of this paper, as it will require a general theory of how context affects acceptability judgments. Though no such theory currently exists, the suggestions in information structure-based theories about the role of context may be a useful place to start (see works cited above). Furthermore, the data sets and predictors that we compiled here could be used to test theories about which semantic/pragmatic properties (of verbs or even the specific sentence frames that we constructed) are more likely to be affected by (dialogue) context.

10. CONCLUSION. Bridge effects have been variously attributed to information structure constraints or processing factors related to template-based processing or frequency effects (e.g. Erteschik-Shir 1973, Ambridge & Goldberg 2008, Dąbrowska 2008, Richter & Chaves 2020, Kothari 2008, Liu et al. 2022, among others). A recent study (Liu et al. 2022), pursuing an extreme version of a frequency-based processing account, has even suggested that bridge effects do not exist, once the frequency of a verb co-occurring with a finite complement clause is taken into account. The lack of a consensus on such basic questions around bridge effects, as we argued and as suggested in recent work, partly reflects the fact that experimental studies on which these claims are based have studied relatively few clause-embedding verbs, which makes sampling errors more likely.

Our solution to this empirical and theoretical stalemate was to create two large-scale data sets of English bridge effects as benchmark data sets, which we have made publicly available.

We collected acceptability judgments for sentences presented without and with prior context for a nearly exhaustive set of 640 clause-embedding verbs, and compiled theoretically-relevant measures for each verb, such as backgroundedness judgments, semantic similarity measures, and frequency estimates. Focusing on a subset of 484 verbs for whom finite complement clauses are most likely grammatical for our participants, we addressed three questions about bridge effects: Do they exist at all? Which existing theory best explains the source of bridge effects? And are there new patterns in our data sets that could lead to a better theory?

Across the full range of verbs, we found clear evidence of bridge effects: verbs do vary significantly in whether they allow long-distance wh-extraction relative to short wh-extraction. Bridge effects exist even after accounting for frequency (contra Liu et al. 2022), backgroundedness, and semantic similarity. Bridge effects were also observed in the presence of a dialogue that provided prior context, contrary to what one might expect given prior reports that context can make long-distance wh-extraction more acceptable for certain verbs.

With this confirmation of the existence of bridge effects, we then statistically evaluated the three leading theories about their source. We found that the information structure theory performs the best, but model fits for all three theories are relatively low, contrary to expectations and previously-reported experiment results (although these experiments had much smaller samples).

The underperformance of existing theories suggests that bridge effects could benefit from fresh theorizing. We identified from our data sets a novel morphosyntactic predictor of bridge effects—nonfinite complementation, which we believe potentially connects to a growing literature on cross-clausal A-dependencies in theoretical syntax (Wurmbrand 2019 and references therein). Integrating this new predictor with existing ideas in the bridge effects literature, we presented a multivariate layered theory of bridge effects in which wh-extraction is licensed by (morpho)syntax and is further subject to information structure constraints and processing costs. Our evaluation of this multivariate theory shows that it explains the largest share of observed variation (R^2), relative to a multivariate theory without a syntactic licensing component as well as single-predictor theories, in which bridge effects are attributable to a single factor (e.g. information structure, some processing factor, morphosyntax). The multivariate theory with a syntactic licensing component also has the lowest BIC, indicating that it achieves the best fit of the data (among the various models evaluated), after controlling for model complexity.

In sum, our investigation of a comprehensive set of English clause-embedding verbs provided novel evidence for bridge effects as a phenomenon to be studied, but showed that existing theories, elegant as they may be, provide only a limited explanation of overall variation. On a more positive note, our study here has pointed to a clear path forward. We demonstrated how insights from existing theories and a novel morphosyntactic predictor identified from our data sets can be combined fruitfully to develop a layered theory that offers better empirical coverage. In the longer term, we hope our data sets and findings will support efforts to develop more robust theories of bridge effects and wh-dependencies.

REFERENCES

AMBRIDGE, BEN, and ADELE GOLDBERG. 2008. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics* 19.357–89.

AMBRIDGE, BEN; JULIAN M. PINE; and ELENA V. M. LIEVEN. 2015. Explanatory adequacy is not enough: Response to commentators on 'Child language acquisition: Why universal grammar doesn't help'. *Language* 91.e116–e126.

ANAND, PRANAV; JANE GRIMSHAW; and VALENTINE HACQUARD. 2019. Sentence embedding predicates, factivity and subjects. *Tokens of Meaning: Papers in Honor of Lauri Karttunen*, ed. by Cleo Condoravdi and Tracy Holloway King. Stanford, CA: CSLI Publications.

BAAYEN, R. HARALD; DOUGLAS J. DAVIDSON; and DOUGLAS M. BATES. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

BATES, DOUGLAS; MARTIN MÄCHLER; BEN BOLKER; AND STEVE WALKER. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67.1–48.

BIBER, DOUGLAS. 1999. A register perspective on grammar and discourse. *Discourse Studies* 1.131–50.

BIRD, STEVEN; EWAN KLEIN; and EDWARD LOPER. 2009. *Natural language processing with Python: analyzing text with the Natural Language Toolkit*. O'Reilly Media.

BRESNAN, JOAN; ANNA CUENI; TATIANA NIKITINA; and HARALD BAAYEN. 2007. Predicting the dative alternation. *Cognitive Foundations of Interpretation*, ed. by Gerlof Bouma, Irene Krämer, Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

CHAVES, RUI P., and MICHAEL T. PUTNAM. 2020. *Unbounded dependency constructions: Theoretical and experimental perspectives*. Oxford: Oxford University Press.

CHOMSKY, NOAM. 1973. Conditions on transformations. *A festschrift for Morris Halle*, ed. by Stephen Anderson and Paul Kiparsky, 232–86. New York: Holt, Rinehart, and Winston.

FELLBAUM, CHRISTIANE. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

DE CUBA, CARLOS. 2018. Manner-of-speaking *that*-complements as close apposition structures. *Proceedings of the Linguistic Society of America* 3.32–1.

DĄBROWSKA, EWA. 2008. Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics* 19.391–425.

DĄBROWSKA, EWA. 2013. Functional constraints, usage, and mental grammars: A study of speakers' intuitions about questions with long-distance dependencies. *Cognitive Linguistics* 24.633–665.

DAVIES, MARK. 2020. The Corpus of Contemporary American English (March 2020). Online: https://www.english-corpora.org/coca/ (accessed 2019-2020).

DEAN, JANET. 1967. Noun phrase complementation in English and German. MIT MS.

DEERWESTER, SCOTT; SUSAN T. DUMAIS; GEORGE W. FURNAS; THOMAS K. LANDAUER; and RICHARD HARSHAM. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41.391–407.

DRUMMOND, ALEX. 2012. IbexFarm (version 0.3.7). [Software]

ERTESCHIK-SHIR, NOMI. 1973. On the nature of island constraints. Cambridge, MA: MIT dissertation.

ERTESCHIK-SHIR, NOMI. 2017. Bridge Phenomena. *The Wiley Blackwell Companion to Syntax, 2nd Edition*, ed. by Martin Everaert and Henk van Riemsdijk. Oxford: John Wiley and Sons.

FARES, MURHAF; ANDREY KUTUZOV; STEPHAN OEPEN; and ERIK VELLDAL. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 271–276.

FEATHERSTON, SAMUEL. 2004. Bridge verbs and V2 verbs—the same thing in spades? *Zeitschrift für Sprachwissenschaft* 23.181–209.

GOLDBERG, ADELE. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

HALE, JOHN. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of NAACL*, 159–66.

HOFMEISTER, PAUL, and IVAN A. SAG. 2010. Cognitive constraints and island effects. *Language* 86.366–415.

JEFFREYS, HAROLD. 1961. *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

KASS, ROBERT E., and ADRIAN E. RAFTERY. 1995. Bayes Factors. *Journal of the American Statistical Association* 90.773–95.

KASTNER, ITAMAR. 2015. Factivity mirrors interpretation: The selectional requirements of presuppositional verbs. *Lingua* 164.156–88.

KIM, BOYOUNG, and GRANT GOODALL. 2022. The island/non-island distinction in long-distance extraction: Evidence from L2 acceptability. *Glossa* 7.1.1–42.

KIPARSKY, PAUL, and CAROL KIPARSKY. 1970. Fact. *Progress in Linguistics*, ed. by Manfred Bierwisch and Karl Erich Heidolph, 143–73. The Hague: Mouton.

KOTHARI, ANUBHA. 2008. Frequency-based expectations and context influence bridge quality. *Proceedings of WECOL 2008: Western Conference on Linguistics*, ed. by Michael Grosvald and Dionne Soares, 136–49. Fresno, CA: California State University.

KUZNETSOVA, ALEXANDRA; PER B. BROCKHOFF; and RUNE H. B. CHRISTENSEN. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82.1–26.

LANDAU, IDAN. 2015. *A two-tiered theory of control*. Cambridge, MA: MIT Press.

LEVIN, BETH. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

LEVY, ROGER. 2008. Expectation-based syntactic comprehension. *Cognition* 106.1126–77.

LIU, YINGTONG; RACHEL RYSKIN; RICHARD FUTRELL; and EDWARD GIBSON. 2022. A verb-frame frequency account of constraints on long-distance dependencies in English. *Cognition* 222.104902.

LUKA, BARBARA J., and LAWRENCE W. BARSALOU. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52.436–459.

MAKOWSKI, DOMINIQUE; MATTAN S. BEN-SHACHAR; and DANIEL LÜDECKE. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software* 40.1541.

MUCHINSKY, PAUL M. 1996. The correction for attenuation. *Educational and Psychological Measurement* 56.63–75.

MÜLLER, SONJA. 2015. Deriving island constraints with Searle and Grice. a pragmatic account of bridge verbs. *Studia Linguistica* 69.1–57.

PARKER, ROBERT; DAVID GRAFF; JUNBO KONG; KE CHEN; and KAZUAKI MAEDA. 2011. *English Gigaword, fifth edition* (LDC2011T07). Philadelpha, PA: Linguistic Data Consortium.

PENNINGTON, JEFFREY; RICHARD SOCHER; and CHRISTOPHER D. MANNING. 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–43.

POSTAL, PAUL. 1974. *On raising*. Cambridge, MA: MIT Press.

REED, LISA A. 2023. Simplifying the theoretical treatment of *wager* verbs. *The Linguistic Review* 40.461–97.

RICHTER, STEPHANIE, and RUI P. CHAVES. 2020. Investigating the role of verb frequency in factive and manner-of-speaking islands. *Proceedings of CogSci 42*, 1771–7. Cognitive Science Society.

SNYDER, WILLIAM. 1992. Wh-extraction and the lexical representation of verbs. MIT MS. Online: https://william-snyder.uconn.edu/wp-content/uploads/sites/2834/2019/11/Snyder_1992_Wh_V.pdf (accessed May 2023).

STOWELL, TIMOTHY. 1981. Origins of phrase structure. Cambridge, MA: MIT dissertation.

SPEARMAN, CHARLES. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15.72–101.

SPROUSE, JON; CARSON T. SCHÜTZE; and DIOGO ALMEIDA. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001-2010. *Lingua* 134.219–48.

SPROUSE, JON; BERACAH YANKAMA; SAGAR INDURKHYA; SANDIWAY FONG; and ROBERT C. BERWICK. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review* 35.575–99.

SPROUSE, JON; and SANDRA VILLATA. 2021. Island effects. *The Cambridge Handbook of Experimental Syntax*, ed. by Grant Goodall, 227–57. Cambridge: Cambridge University Press.

ŞTEFĂNESCU, DAN; RAJENDRA BANJADE; and VASILE RUS. 2014. Latent Semantic Analysis models on Wikipedia and TASA. *Proceedings of LREC 2014*, 1417–22.

STOICA, IRINA. 2016. Island effects and complementizer omission: the view from manner of speaking verbs. *Constructions of Identity (VIII): Discourses in the English-Speaking World*, ed. by Petronia Petrar and Amelia Precup, 191–200. Cluj-Napoca, România: Presa Universitară Clujeană.

URIAGEREKA, JUAN. 1999. Multiple spell-out. *Working minimalism*, ed. by Samuel D. Epstein and Norbert Hornstein, 251–82. Cambridge, MA: MIT Press.

VASISHTH, SHRAVAN; DANIELA MERTZEN; LENA A. JÄGER; and ANDREW GELMAN. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103.151–75.

WHITE, AARON STEVEN, and KYLE RAWLINS. 2018. The role of veridicality and factivity in clause selection. *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, ed. by Sherry Hucklebridge and Max Nelson, 221–34. Amherst, MA: GLSA Publications.

WHITE, AARON STEVEN, and KYLE RAWLINS. 2020. Frequency, acceptability, and selection: a case study of clause-embedding. *Glossa* 5.105.1–41.

WURMBRAND, SUSI. 2018. Cross-clausal A-dependencies. Lecture handouts and slides for course taught at the ACTL Summer School, London.

WURMBRAND, SUSI. 2019. Cross-clausal A-dependencies. *Proceedings of the Fifty-Fourth Annual Meeting of the Chicago Linguistic Society*, ed. by Eszter Ronai, Laura Stigliano, and Yenan Sun, 585–604. Chicago: Chicago Linguistic Society.

Supplementary material for "A nearly-exhaustive experimental investigation of bridge effects in English"

NICK HUANG
National University of Singapore

DIOGO ALMEIDA
New York University Abu Dhabi

JON SPROUSE
New York University Abu Dhabi

APPENDIX A. VERBS AND DESCRIPTIVE STATISTICS ABOUT ACCEPTABILITY AND PENALTIES (CSV FILE). These figures are derived from our acceptability judgment experiments and data analysis process as reported ino Section 3.

APPENDIX B. CORRECTING FOR THE RELIABILITY-BASED ATTENUATION OF R^2. We correct for attenuation in R^2 values with the formula in (1). This formula is derived from the formula for correcting Pearson correlation coefficients, given in (2) (e.g. Spearman 1904; Muchinsky 1996), and the fact that R^2 in simple linear regressions is equivalent to the square of the correlation coefficient between the dependent and independent variables.

(1) $$\text{Corrected } R^2 = \frac{\text{Observed } R^2}{\text{Reliability}_X \times \text{Reliability}_Y}$$

(2) $$\text{Observed correlation}_{X,Y} = \text{True correlation}_{X,Y} \times \sqrt{\text{Reliability}_X \times \text{Reliability}_Y}$$

We obtained reliability estimates through bootstrapping. For each verb, we calculated the long-distance penalty scores for every participant whose responses met our inclusion criteria. For each of the 484 verbs of interest, we created two sets of penalty scores that match in size the original set of scores, by randomly sampling with replacement from the original set. We calculated a mean penalty score for each verb in each set, producing two lists of 484 penalty scores. The Pearson correlation between the two lists was calculated. We repeated this process 5,000 times, taking the mean correlation as the estimate of reliability of long-distance penalties. The reliability of acceptability penalties, in the absence of a dialogue establishing prior context (Section 3.2) is .81, while the reliability of penalties in the presence of such a dialogue (Section 3.3) is .76.

We repeat this analysis for the two backgroundedness measures (True and not-False measures; Section 4.1). The reliability of the True measure is .95, while the reliability of the not-False measure is .85. Note that these estimates are for a total of 482 verbs (the 484 verbs less *bear* and *stand*; *forgive* did not meet our inclusion criteria).

Calculating reliability for the other variables is trickier, since the process presupposes that we can easily obtain new estimates for each measure. This is not feasible for semantic similarity measures, which were derived using computationally intensive methods, nor for measures derived from large, tagged corpora, since there are relatively few of these. For the sake of exposition, we assume perfect reliability (=1) for these measures.

APPENDIX C. REGRESSION RESULTS FOR THE TEMPLATE-BASED PROCESSING THEORY. As described in Section 4.2, we used four different data sets to calculate a set of three semantic similarity measures per data set: a similarity score with *say* as the benchmark, a similarity score with *think* as the benchmark, and a hybrid semantic similarity score that takes whichever score is greater between *say* and *think*. We fitted regression models for each of the twelve similarity measures, but only reported results for four of these measures—the hybrid scores—in the main paper, for space reasons. Tables 1 and 2 present the results for the remaining eight measures.

	No prior context					With prior context				
	b (s.e.)	*t*	*p*	Eff. Size	BF_{10}	*b (s.e.)*	*t*	*p*	Eff. Size	BF_{10}
Similarity to *say*										
LSA/Wikipedia	-0.10 (0.05)	-1.88	.06	0.09	<0.1	-0.15 (0.06)	-2.25	.02	0.13	<0.1
GloVe/Wikipedia	0.20 (0.04)	5.13	<.01	0.24	>100	0.29 (0.05)	5.99	<.01	0.34	>100
GloVe/Gigaword	0.27 (0.04)	6.76	<.01	0.32	>100	0.33 (0.05)	6.70	<.01	0.40	>100
WordNet path-similarity (log)	0.06 (0.04)	1.70	.09	0.07	<0.1	0.07 (0.05)	1.46	.14	0.08	<0.1
Similarity to *think*										
LSA/Wikipedia	-0.13 (0.04)	-3.22	<.01	0.12	0.9	-0.10 (0.05)	-1.88	.06	0.09	<0.1
GloVe/Wikipedia	0.17 (0.04)	4.76	<.01	0.20	>100	0.33 (0.04)	7.52	<.01	0.39	>100
GloVe/Gigaword	0.15 (0.04)	4.12	<.01	0.17	24.5	0.29 (0.05)	6.35	<.01	0.33	>100
WordNet path-similarity (log)	0.10 (0.04)	2.45	.01	0.12	<0.1	0.11 (0.05)	2.21	.03	0.13	<0.1

TABLE 1. Interaction effects between wh-dependency length and semantic similarity scores for models of z-scored acceptability.

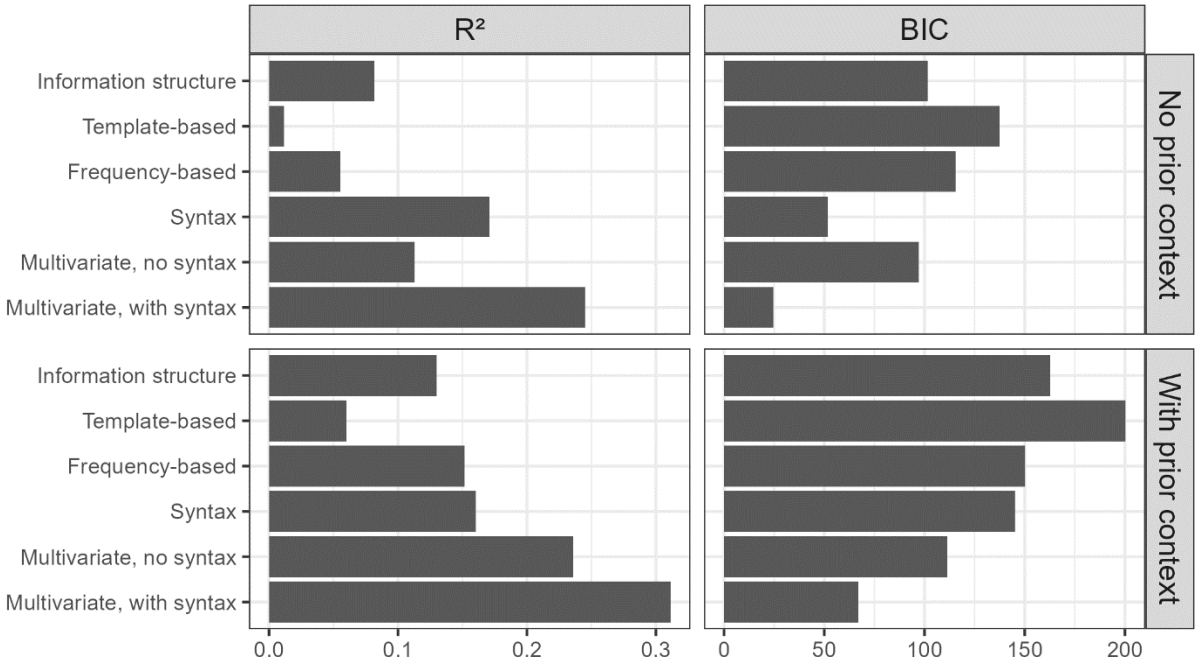
	No prior context						With prior context					
	b (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2	*b* (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2
Similarity to *say*												
LSA/Wikipedia	0.08 (0.10)	0.79	0.07	<0.1	.001	.002	0.14 (0.11)	1.26	0.12	0.1	.004	.005
GloVe/Wikipedia	-0.21 (0.07)	-2.82	0.25	2.5	.018	.022	-0.30 (0.08)	-3.71	0.35	42.3	.030	.040
GloVe/Gigaword	-0.27 (0.08)	-3.62	0.33	30.9	.028	.035	-0.33 (0.08)	-4.12	0.40	>100	.037	.049
WordNet path-similarity (log 10)	-0.06 (0.07)	-0.88	0.07	<0.1	.002	.002	-0.08 (0.08)	-0.99	0.09	<0.1	.002	.003
Similarity to *think*												
LSA/Wikipedia	0.12 (0.08)	1.57	0.11	0.2	.006	.007	0.10 (0.08)	1.13	0.09	<0.1	.003	.004
GloVe/Wikipedia	-0.18 (0.07)	-2.59	0.21	1.3	.015	.019	-0.33 (0.07)	-4.53	0.39	>100	.045	.059
GloVe/Gigaword	-0.16 (0.07)	-2.28	0.18	0.6	.011	.014	-0.29 (0.07)	-3.82	0.33	64.7	.032	.042
WordNet path-similarity (log 10)	-0.11 (0.08)	-1.36	0.12	0.1	.004	.005	-0.11 (0.08)	-1.35	0.13	0.1	.004	.005

TABLE 2. Comparison of models of bridge effects for template-based processing theory.

APPENDIX D. RESULTS FOR LINEAR REGRESSIONS BETWEEN PENALTIES AND BEST-PERFORMING PREDICTORS OF EACH THEORY, FOR ALTERNATIVE VERB EXCLUSION CRITERIA. Sections 5-7 presented analyses for a set of 484 verbs of interest, for which "no context" short wh-dependencies had a z-scored acceptability rating of 0 or greater, on the assumption that verbs with negative ratings do not allow finite clausal complements. As described in Section 8.3, a reviewer expressed concerns over the validity of this criterion, because the presence of the wh-dependency might have also lowered acceptability.

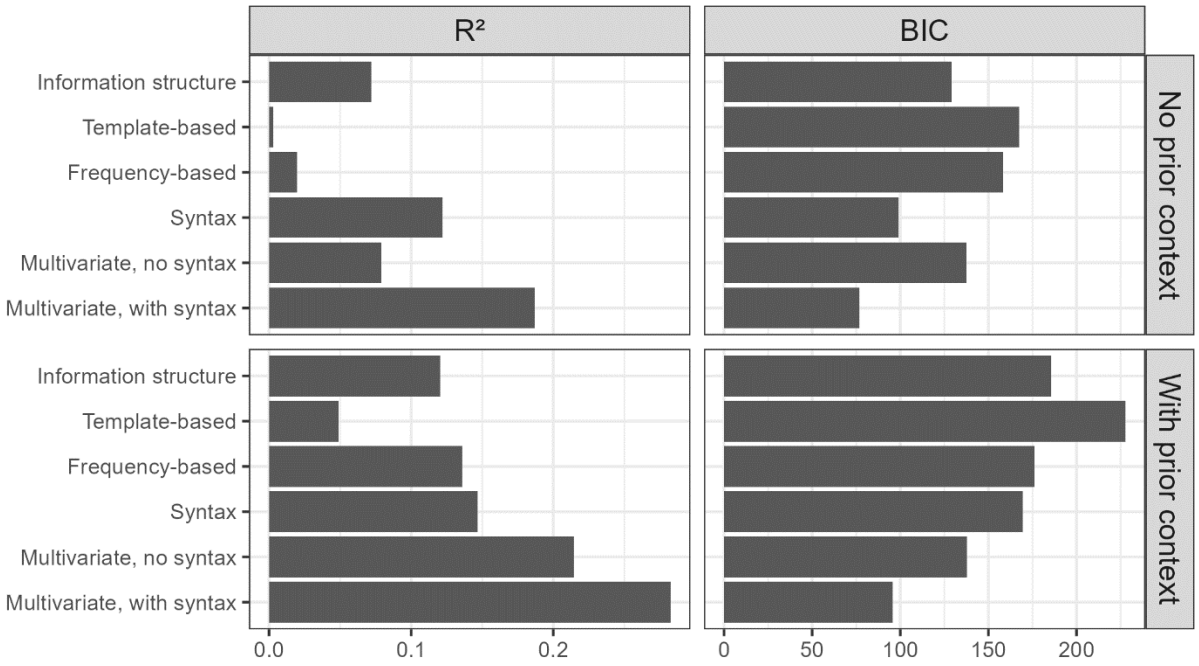
In response to this concern, we re-ran a key analysis—the model fit analysis reported in Figure 9, Section 7.2. This analysis compares model fits for the best-performing predictor for each of the four single-factor theories: %True responses (information structure), GloVe/Wikipedia similarity (template-based), log verb bias (frequency-based), and nonfinite complementation (syntax), and contrasts them with two multivariate models, one that linearly combines all three non-syntactic predictors, and another that linearly combines all four predictors. Crucially, we tried out two different verb exclusion criteria, (i) relaxing the no-context short wh-dependency acceptability threshold to -0.25, and (ii) eliminating it altogether. In both cases, we still required short wh-dependencies to be at least as acceptable as long wh-dependencies. The first criterion (threshold of -.25) yielded a set of 488 verbs with a full set of predictors, and the second (no threshold) a set of 536 verbs with a full set of predictors.

Figures 1 and 2 show model fits (R^2s) and Bayesian Information Criterion (BIC), which balances model fit with a penalty for increased complexity (the lower the BIC, the better). Visually speaking, both figures are very similar to each other (and also to Figure 9 and tables reported in the paper): the syntax-only model is often one of the best single-factor models, with fits comparable to, if not higher than, the information structure-only or frequency-only models. The template-based-only models had the lowest R^2s. The "multivariate, with syntax" models consistently had the highest R^2s and lowest BICs. These results show that our conclusions leading up to and through Section 7 are not sensitive to what verbs were included in our analyses.



Note: The higher the R^2 and the lower the BIC, the better the fit.

FIGURE 1. Model fits for selected models of bridge effects, for the 488 verbs where short wh-dependencies had a z-scored acceptability rating above -0.25 (among other criteria).



Note: The higher the R^2 and the lower the BIC, the better the fit.

FIGURE 2. Model fits for selected models of bridge effects, for a set of 536 verbs where there was no criterion on the acceptability of short wh-dependencies (among other criteria).