Journal of Memory and Language 125 (2022) 104335

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Referencing context in sentence processing: A failure to replicate the strong interactive mental models hypothesis

Jack Dempsey *, Kiel Christianson

University of Illinois, Urbana-Champaign, United States

ARTICLE INFO

Keywords:
discourse processing
sentence processing
replication
psycholinguistics

ABSTRACT

The role of contextual influences on sentence processing remains underdeveloped and is often assumed to play some vaguely defined role in sentence processing. In 2005, Grodner, Gibson, and Watson (GGW) used self-paced reading to test the extent to which preceding discourse structure and complexity influence the interpretation of restrictive relative clauses. Their findings suggest that discourse can influence unambiguous syntactic selection processes and does so rapidly, such that discourse can project syntactic structures onto subsequent text during reading. Although being moderately cited since its publication, there has been no replication attempt to this date, all the while the field has developed more sophisticated methods for crossed designs and has highlighted the need for robust replication attempts to mitigate an ever-growing replication crisis. Bayesian modeling yielded considerable evidence against the interaction effect that supports GGW's Strong Interactive Mental Models Hypothesis, suggesting discourse information does not proactively facilitate or project syntactic structures.

Introduction

Linguistic output and input, although continuous, are often broken down into discrete units by producers, comprehenders, and psycholinguists. There exists some contention between prescriptive definitions of what constitutes a sentence versus the average speaker's intuition of how to group meaningful utterances together, but sentence wrap-up effects (e.g. Caplan, 1972; Gernsbacher, Hargreaves, & Beeman, 1989) suggest that language is likely 'packaged' into discrete units, easing the processing required on the part of the comprehender as well as the producer (MacDonald, 2013). As researchers, these units make it attractive to focus on different levels of processing (e.g. word/letter, phrasal, sentence, discourse), whereas the direct connections between these levels are less well represented in the field of psycholinguistics (c.f., Ferreira & Yang, 2019). However, despite the discreteness of linguistic units, written words often occur in sentential contexts and written sentences often occur in discourse contexts. This presents a troublesome issue for the field of sentence processing: an overreliance on single-sentence stimulus sets may result in the inability to generalize to natural reading contexts. In particular, many models of sentence processing remain agnostic at best as to how information from discourse affects the interpretation of sentence structures.

Sentence processing research has used discourse designs to test certain theories of final interpretations derived from the sentence in question; for example, Slattery and colleagues (2013) used a two-sentence discourse paradigm to show that both the initial misinterpretation and the correct interpretation of garden path sentences exist simultaneously, suggesting lingering misinterpretations occur due to a failure to inhibit the initial misparse rather than a failure to eventually arrive at the correct interpretation (see also Cutter et al., 2021; Dempsey et al., 2021). While Sturt (2003) used this discourse design to establish the gender mismatch effect upon which Slattery et al.'s design depended, other researchers have also benefited from the use of discourse stimuli to investigate downstream effects of sentence processing (e.g. Christianson & Luke, 2011) to examine how context reinforces or protects against garden-path misanalysis. Importantly, discourse processing is not necessarily the main focus of these research designs; rather, discourse is used as a means to influence and infer downstream processes of interpretations derived from a given target structure earlier in the stimulus. These studies therefore make the assumption that previous linguistic input in discourse will affect downstream sentence processing, but this assumption is made without any specific claims on the time-course or quality of information that is carried in memory throughout discourse comprehension.

* Corresponding author at: Department of Educational Psychology, Education Building, Rm. 226A, University of Illinois, 1310 S. 6th St., Champaign, IL 61820, United States.
E-mail address: jkdemps2@illinois.edu (J. Dempsey).

https://doi.org/10.1016/j.jml.2022.104335
Received 30 December 2020; Received in revised form 11 April 2022; Accepted 28 April 2022
Available online 12 May 2022

Garden-path processing literature has shown evidence for an initial syntactic processing mechanism relatively immune to preceding contextual information (Christianson & Luke, 2011; Dempsey & Brehm, 2020; but see Binder et al., 2001 for mixed evidence). This finding contrasts with studies like those mentioned above, which rely on context influencing processing behaviors in downstream structures, in that garden path structures need not rely on previous information in the discourse for successful resolution. Indeed, earlier work from the 1980s investigating the role of discourse on syntactic processing tended to focus on temporarily ambiguous (i.e. garden path) structures like in 1a-1b below. These two sentences are identical and ambiguous until the disambiguating region after 'trouble with': in 1a the resolution 'her husband' leads to an interpretation of 'that' being a complementizer while the resolution 'to leave' in 1b leads to a relativizer interpretation. These kinds of stimuli, used by Crain and Steedman (1985), display a reading pattern where disambiguating regions of sentences like 1a are usually easier to read than their counterparts in sentences like 1b.

1a) A psychologist told the woman that he was having trouble with her husband.
1b) A psychologist told the woman that he was having trouble with to leave.

Referential theory, as proposed by Crain and Steedman (1985), Altmann and Steedman (1988), and Ni and Crain (1989), explains this apparent structural preference for resolutions like 1a compared with resolutions like 1b via the principle of parsimony, or that the easiest interpretation given the current discourse state should be chosen among competitors upon encountering ambiguity. For example, when readers encounter 'the woman' in a null context (i.e. no preceding discourse information), the subsequent 'that' is less likely to be interpreted as a relativizer because this would imply a competitor set of 'woman' referents. Since there exists only one referent for 'woman' in the given discourse at its current state (i.e., null context), it would be a more complex proposition to treat 'that' as a relativizer rather than as a complementizer because the latter does not imply additional referents. For this reason, referential theory claims that readers prefer structural interpretations like 1a, the ambiguity resolution that allows for the most parsominious continuation of the sentence given the current discourse state. However, given an appropriate supportive context, such as 1c taken from Crain & Steedman (1985), the preference shifts to interpretations like 1b since there are now multiple 'woman' referents to allow a contrastive set.

1c) A psychologist was counseling two women. He was worried about one of them, but not the other.

An important drawback of referential theory as proposed by these authors was that it only accounted for how discourse influences sentence processing under ambiguous conditions. As Grodner, Gibson, and Watson (2005; GGW) point out, "it is hard to conceive of a naturalistic sentence that does not exhibit a referential function" (p. 277); in other words, even unambiguous sentences likely reference discourse. GGW thus posited three hypotheses for the existence and thereupon contingent temporal nature of discourse influences on structural parsing. These hypotheses as put forth by GGW are found in Table 1. The Ambiguity Only Hypothesis contrasts with the two Interactive Mental Models Hypotheses in that the latter two assume discourse plays a role in sentence processing in both ambiguous and unambiguous settings. What sets these two hypotheses apart is the temporal nature of contextual influences: the strong version holds that information from discourse is available at early stages of syntactic processing and can even project expected structures onto subsequent sentences, while the weak version limits discourse interactivity to later stages of processing, initial interpretations being later checked against the current discourse model for integration.

Table 1
Mental Models Hypotheses from GGW, 2005. Predictions laid out with respect to GGW's Experiment 2 predictions and findings.

Hypothesis	Description	Prediction
Ambiguity Only	"The discourse is consulted only in the face of ambiguity. The processing mechanism interprets an ambiguous utterance so as to make the background assumptions of the utterance consonant with a relevant model of the discourse context."	No Effects
Strong Interactive Mental Models	"The discourse model is constantly updated and accessed in the comprehension of a sentence. Sometimes the sentence causes the construction of discourse structure. Othertimes the discourse model directs interpretive processes and projects syntactic structures."	Main Effect of RC Type Interaction of Context & RC Type
Weak Interactive Mental Models	"Sentences are parsed using intrasentential criteria, such as syntactic knowledge. The resultant analysis (or analyses in the case of ambiguity) is then evaluated against the context, and changes are incrementally made to the current discourse model. These changes can incur costs that interfere with interpretive processes and lead to comprehension difficulty."	Main Effect of RC Type No Interaction of Context & RC Type

To test between these competing models, GGW ran two word-by-word self-paced reading experiments manipulating contextual support and structural discourse complexity by having participants read sentences with supportive or null biases towards either restrictive or non-restrictive relative clause structures. The stimuli from this experiment are presented in Table 2 below. They operationalized processing difficulty as reading times in the embedded verb phrase (e.g. 'dog bit') across restrictive and non-restrictive relative clause structures (RCs) and also manipulated whether these sentences were presented in a null context or a supportive context. Restrictive RCs are so named because they 'restrict' the potential pool of referents in the discourse state: for example, 'that a dog bit on the leg' restricts the competitor set of potential 'postman' referents to which the subject of the sentence can refer. Even in null contexts, this purportedly confers discourse complexity; however, this discourse complexity should be more easily interpreted in supportive contexts where a competitor set of 'postman' referents is

Table 2
Stimuli from GGW Experiment 2. The critical region is shown in bold, but participants saw all words in plain text.

Context Type	Relative Clause Type	Stimulus
Null	Restrictive	The postman that a **dog bit** on the leg needed seventeen stitches and had a permanent scar from the injury. It took a few hours for the police to find the dog that was terrorizing the neighborhood.
Null	Non-Restrictive	The postman, who a **dog bit** on the leg, needed seventeen stitches and had a permanent scar from the injury. It took a few hours for the police to find the dog that was terrorizing the neighborhood.
Supportive	Restrictive	A vicious guard dog bit a postman on the leg and another postman on the arm. The postman that a **dog bit** on the leg needed seventeen stitches and had a permanent scar from the injury.
Supportive	Non-Restrictive	A vicious guard dog bit a postman and a garbage man. The postman, who a **dog bit** on the leg, needed seventeen stitches and had a permanent scar from the injury.
Question:		Did the postman need seven stitches?

explicitly made. This would be evidence for the Strong Interactive Mental Models Hypothesis because it would suggest readers were expecting a restrictive structure based on the supportive discourse context.

This 2x2 design allows for a test between the three hypotheses. If context is only referenced in cases of ambiguity, then there should be no difference in reading times between restrictive and non-restrictive RC structures in null contexts; however, if context does inform parsing preferences, the inherent discourse complexity of restrictive RCs should incur a processing cost in the form of longer reading times at the critical region compared with non-restrictive RCs. Therefore, the Ambiguity Only Hypothesis predicts no main effect of RC Type or interaction because neither discourse complexity nor contextual support should influence the processing of unambiguous structures. The Weak Interactive Mental Models Hypothesis predicts no interaction of Context Type and RC Type on reading times because the parser should not use context until after initial algorithmic processing of the syntax; however, the Strong Interactive Mental Models Hypothesis predicts an interaction wherein null contexts lead to slower reading of restrictive RCs while supportive contexts lead to faster reading of restrictive RCs.

GGW found a main effect of Context Type and an interaction between Context Type and RC Type, as predicted by the Strong Interactive Mental Models Hypothesis, across two experiments (Experiment 2 fixed a caveat where the first experiment presented null context sentences with an indefinite subject). From these findings, they conclude that the Strong Mental Models Hypothesis is supported by the data, suggesting that "the current discourse model can guide structure-building processes within the sentence" (p. 287, Grodner et al., 2005). In summary, their data show that discourse properties like the availability of explicit, contrastive referents can change expectations for upcoming structures. This finding is critical for models of sentence processing because it shows that discourse information needs to be accounted for to explain sentence-level comprehension processes, which more naturally appear embedded within discourse contexts. It also contrasts starkly with findings from the garden path literature, which depicts a rather stubborn sentence processing mechanism that refuses to switch initial preferences/expectations away from the *a priori* more expected structure even in supportive contexts (Binder et al., 2001; Christianson & Luke, 2011; Dempsey & Brehm, 2020) or after repeated exposure to the structure (Dempsey, Liu, & Christianson, 2020; Harrington-Stack et al., 2018). However, work by Fedorenko, Piantadosi, and Gibson (2012) showed that supportive contexts can inflate the preference for subject-extracted over object-extracted relative clause structures; therefore, it could be the case that initial expectations for misparses in temporarily ambiguous structures are too strong for contextual effects to overturn early in processing. Whether or not this is true, it seems that context has sometimes been shown to play a proactive role in guiding syntactic selection processes.

Motivation for replication

Perhaps surprisingly, GGW has only been cited 87 times since its 2005 publication according to Google Scholar, which may be interpreted as a low number given its strong claims about the role of discourse on syntactic processing. Nearly all these citations simply cite the article in a cursory fashion as a means of providing evidence for contextual effects – there has not to our knowledge been a direct replication or a study specifically designed to further test the claims made by GGW (although Fedorenko and colleagues (2012) do highlight the importance of GGW's findings more so than other citing authors). As the call for sentence processing models to include influences from discourse grows (e.g., Ferreira & Yang, 2019; Luke & Christianson, 2011) and studies begin to use discourse stimuli to investigate sentence-level phenomena (e.g., Cutter et al., 2021; Dempsey et al., 2021; Slattery et al., 2013), it is more important than ever to revisit this work to first test for replicability before investigating exactly what determines whether, when, and how discourse contexts influence syntactic expectations and interpretations. The first thing GGW showed was that previous context modulated the subsequent interpretation of text for unambiguous conditions. It is important to clarify that neither structure in GGW's design was ambiguous. Although their design was successful in testing specifically between the interactive mental models hypotheses, their findings reveal no evidence of if or how discourse differentially modulates subsequent syntactic processing of ambiguous and unambiguous structures. An implication of this work when viewed together with the ambiguity literature is that discourse processes seem to drive language comprehension as a general function of the parser (i.e., not specific to ambiguity). So, the most apparent line of work that should follow a replication of GGW's findings would be to directly compare the processing of ambiguous and unambiguous structures with null and supportive contexts to see how specific structures interact with this general function. This may, for example, help mediate between findings showing facilitatory discourse influences on parsing and findings showing no such effect.

The claim that language is not processed in a purely bottom-up fashion is certainly not new, being a central claim in the field of predictive processing. A plethora of work over the past several decades has shown that highly constrained contexts can lead to anticipatory processing of various linguistic elements (for a review, see Kuperberg & Jaeger, 2016). In many ways, GGW's findings add support to the ever-growing body of evidence suggesting prediction plays a central role in the comprehension of language. On the other hand, predictive processing has more recently been shown to be selective in its role; for example, Luke and Christianson (2016) showed that most words in a given text are actually not predictable, but that syntactic categories were comparatively more predictable. Additionally, readers often engage in shallow processing whereby they abstain from making hasty decisions given ambiguous input (Traxler et al., 1998; Swets et al., 2008). Together, these findings suggest there remains a gap in knowledge for exactly when, why, and how predictive processing is triggered. To this end, the Interactive Mental Models Hypotheses seek to explain predictive processing on a discourse-syntax interface level – a replication of the Strong Interactive Mental Models Hypothesis would suggest a strong role of discourse in predicting upcoming structures. Alternatively, it may be found that there is more evidence for the Weak Interactive Mental Models Hypothesis, in which case discourse-driven prediction will seem to have a weaker role than previously thought in sentence processing, being relegated to post-sentential integration processes. In sum, this replication could pave the way for future studies investigating more closely the universality of the role of discourse across structures, language processing tasks, and individual participants.

More broadly, Good-Enough (GE) (Christianson 2016; Ferreira et al., 2002; Karimi & Ferreira, 2016) and Noisy Channel (NC) language processing accounts (Levy, 2008; Levy et al., 2009; Futrell & Levy, 2017) have grown increasingly central to the field of sentence comprehension over recent years and could also benefit from a replication attempt of GGW. Although they differ on their primary focus and proposed mechanisms of deriving nonliteral interpretations of the input, both GE and NC accounts rely on the assumption that language comprehension is achieved through more than an initial algorithmic processing system. NC accounts of processing argue that comprehenders arrive at interpretations of the input that are guided by experience-based expectations coupled with a desire to mitigate noise (e.g., production errors, 'lossy' memory [Futrell et al., 2020]). For example, studies investigating predictions of NC accounts have shown that comprehenders automatically correct errors in real-time (Brehm et al., 2021) and quickly adapt to the types of errors they need to correct (Ryskin et al., 2018). Similar in its focus on optimizing language comprehension with less than optimal input and processing capabilities (i.e., humans are prone to error), the GE account argues for two separate streams of processing: a bottom-up algorithmic processing stream and an experience-based heuristic processing stream. Although studies in this framework have shown

evidence for misinterpretation of noncanonical sentences due to the heuristic stream finishing its processing before the algorithmic stream (Christianson et al., 2010; Ferreira, 2003), recent work has argued that many of these effects are instead task-effects that are not representative of real-time interpretive processing (Bader & Meng, 2018; Meng & Bader, 2021).

One important source of data that could elucidate this debate on the scope of algorithmic processing in structural comprehension would be a more concise picture of how discourse informs the starting state of any such processing mechanism. For example, GGW's support for the Strong Interactive Mental Models Hypothesis would suggest that structures are projected onto upcoming input from previously comprehended discourse contexts. This could be seen as a heuristic process whereby bottom-up analysis of the syntax is pre-empted with top-down expectations, fitting very nicely in the framework of GE's dual stream model. One potential study could use supportive and null contexts with noncanonical target sentences to investigate if misinterpretation rates are exacerbated by previous discourse. If so, that would be evidence against the claim that interpretations are built via an algorithmic stream alone. A very recent study by Cutter and colleagues (2021) investigated downstream comprehension following noncanonical structures and found no evidence against an algorithmic-only processing system; however, these stimuli occur in a null context with respect to the noncanonical structure. NC accounts would benefit from clarifying the extent to which previous information from discourse affects decisions made upon encountering noisy input. For example, agreement attraction errors may be resolved in comprehension via different noisy channel updates (e.g., corrections to the verb or corrections to the mismatching target noun) depending on information from the previous context. While these two accounts are illustrative of the need to replicate our understanding of discourse influences on sentence processing, virtually any account of language comprehension needs to be able to account for how interpretive processes are modulated by preceding context if researchers hope to generalize such accounts to naturally occurring written language.

In addition to the widely acknowledged replication crisis in cognitive psychology (e.g. DeLong, Urbach, & Kutas, 2017), the field has also progressed in various ways since 2005. For instance, ANOVAs like those used by GGW and many others in the field prior to the 2010s have since been sidelined for mixed effects models when dealing with crossed experimental designs where item and participant random effects can be expected to contribute a large amount of variance to the data (Baayen et al., 2008). Additionally, GGW does not report an effect size; however, a simple post-hoc power analysis with G*Power software (Erdfelder et al., 1996) was conducted to assess the initial power achieved by GGW's Experiment 2 (see supplementary materials). In their design, they presented participants with 20 items, 5 per condition, and recruited 51 participants (50 participants after comprehension question accuracy cleaning), resulting in 250 observations per cell. This results in a post-hoc power estimate of 59.96% power given an alpha level of .05, medium effect size of.25 Cohen's F, and correlation among measures being set to .5. If the Cohen's F effect size, not reported by GGW, were .4, the power would be better at a near ceiling 99.03%; however, if the effect size were small at a Cohen's F of .1, the power would only have been 10.04%, meaning the range of small to large effect sizes possible yields a power range from 10.04% to 99.03%. Because no effect size measure was reported by the authors, any replication attempt should account for a wide range of possible effect sizes in its *a priori* power analysis.

Bayesian statistics have also become more popular among psycholinguists in the past fifteen years, with recent guides being published to encourage their use in journals (e.g. Schad et al., 2020; Wagenmakers et al., 2016). Bayesian methods are particularly useful in the gradual accumulation of evidence for effects and also in their ability to directly compare probabilities of models, whereas frequentist models should not be used in such a way that *p*-values are contrasted as a measure of probabilistic magnitude (Wagenmakers, 2007). Bayesian models allow for a more informed process in the accumulation of converging evidence because they take into account prior expectations for effects, which influences a posterior distribution derived from the data in a way proportional to the sample size (i.e., priors have a larger effect on the posterior distribution when sample sizes are smaller). Although highly informative priors should be avoided in direct replication attempts because they may bias the results in favor of what is trying to be replicated (Liu, 2019), mildly informative priors reflecting expectations for reading patterns (e.g., log-transformed reading times being fit roughly to a Gaussian distribution) can still help avoid sampling errors and allow the data to reveal robust effects (Gelman et al., 2008). Thus, the use of Bayesian methods allows us to use prior expectations, even if only mildly informed, to generate data-driven beliefs of true effects.

Additionally, Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018) presents a convenient method of ensuring adequate power is reached by allowing researchers to set a required minimum sample size that can then be checked at pre-determined intervals until evidentiary support for or against the alternative hypothesis is met (see Methods for more information on Bayesian modeling and power analyses). Unlike more traditional approaches to power, this allows researchers to base sample sizes on observed, quantifiable evidence either for or against the null hypothesis rather than on expected outcomes of repeated experiments. These updates to how the field of psycholinguistics approaches statistical modeling add to the importance of replicating past findings, the implications of which are critical for understanding how the parser uses the discourse state to inform its interpretive preferences.

The current work aims to replicate the main findings from GGW's Experiment 2: specifically, the main effect of RC Type and its interaction with Context Type such that 1) supportive contexts lead to faster critical region reading times overall, 2) the critical region in null contexts is read faster for non-restrictive RCs, and 3) the critical region in supportive contexts is read faster for restrictive RCs. To our knowledge, there has not yet been any attempt to directly replicate these findings despite dozens of citations benefitting from GGW's strong claim that discourse context projects syntactic expectations onto subsequent input. The current replication attempt uses Bayesian inference to obtain adequate evidentiary support for or against the hypotheses in question, allowing also for a maximally fit random effects structure. Our hope is that this study will reinvigorate the field's interest in researching how sentence processing works when embedded within discourse to more faithfully model how humans comprehend language.

Methods

Participants & Bayes Factor design analysis (BFDA)

Because we are taking an agnostic approach to the effect size expected from this experiment, the proposed power analysis randomly samples from a distribution of effect sizes ranging from a Cohen's F of .1 to .4 (which corresponds to a Cohen's D of .2 to .8). Using the BFDA package (Schönbrodt & Stefan, 2019) in R version 4.0.3 (R Core Team, 2020), we simulated 10,000 iterations sampling from a similarly sized distribution of effect sizes from the range outlined above with a minimum sample size of 160. BFDA is used here to simulate probabilities that a Bayes Factor (BF) will fall outside the pre-specified range of inadequate evidentiary support. For example, if we decide we want to see a BF of either 10 or 1/10 for the alternative model, meaning that it is ten times more or less likely respectively than the null model, we can observe what percentage of simulations reach the necessary BF level and what BF the simulations arrive at if they reach a pre-determined stopping point. This stopping point can be arbitrarily chosen, chosen with regards to practical issues such as recruitment or funding capacity, or left out to let each iteration arrive at a BF outside the predetermined threshold (although this may be computationally, and literally, costly). We chose a stopping point of 260 for the current replication proposal

because it yielded a power level of 80% and seemed feasible given resource and time constraints. Two simulations were conducted: one that assumes the true effect size exists for a given range (we specified this as a normal distribution of Cohen's F values, 95% of which range from .1 to .4) and one that assumes the null effect to be true (effect size = 0). For the alternative model simulation, we show the percentage of how many iterations reached the boundary in favor of the alternative model and the BFs of the iterations that reached our stopping point of 260 without crossing either threshold. Importantly, the percentage of iterations that reach a BF smaller than 1, whether by surpassing 1/10 or by reaching the sample size stopping point with that value, reveals the Type II error rate (real effect not found). Similarly, the null model simulation shows us the percentage of how many iterations reached the boundary in favor of the null model and the BFs of the iterations that reached the stopping point; the percentage of iterations that reach a BF greater than 1 would here be the Type I error rate. Therefore, if we want to run a study with 80% power, we would need to find the minimum sample size required for which 80% of iterations in the alternative model simulation result in a BF greater than 1. As an additional check, the Type I error rate should also be below 5% to adhere to normal practices.

The simulations for the alternative and null models can be found in Fig. 1a-b below. The final simulations were based on the lowest minimum number of participants that would allow for less than 5% Type I error rate and less than 20% Type II error rate as interpreted above. Following the results of the BFDA, 160 participants were initially recruited with a planned stopping interval of 20 until either a BF of 10 or 1/10 was reached, or until we reached our stopping point of 260 participants. We reached adequate evidence with our initial recruitment of 160 participants. Importantly, because BFs provide a measure of probabilistic magnitude, even if we had reached our stopping point without obtaining the *a priori* specified BF levels, we still could have interpreted the results. For example, if we had reached the stopping point and observed a BF of 8, we still could have concluded that there was substantial evidence in favor of the alternative hypothesis (criterion as proposed by Jeffreys, 1939/1998). We ran Bayesian mixed models to check the resultant BF value. If the BF had exceeded the 1/10–10 boundary, we would have stopped collection at that point. If not, then another 20 participants would have been recruited until the next required sample size was reached. This process would have been repeated until we had either achieved a BF value that exceeded the boundary or had reached our stopping point.

The stopping point due to resource and time restrictions was set to 260 paricipants; however, a few clarifications are warranted here. First, the power analysis was very conservative in that it sampled effect sizes from a very wide distribution – if the effect size were medium or large, we would very likely not need to reach the stopping point. Second, inconclusive evidence (i.e., BFs ranging from 1/3 to 3) would still be a rather telling result from a highly powered study – the burden of proof in replication attempts should be placed on the original study rather than on the replication, and finding inconclusive evidence is meaningful in that it still constitutes a failure to replicate the original effects. Third, Bayesian inference allows for the data collected in this study to inform future related work. Casting theories from psychological inquiry in the form of probabilities helps overcome the 'paradox' of converging evidence (Davis-Stober & Regenwetter, 2019) and moves the field towards more robust methods of utilizing the accumulation of scientific evidence in a meaningful way (i.e., the use of informative priors) (Vanpaemel, 2010). In other words, were the results of this replication attempt inconclusive, they would still be impactful.

Participants were recruited with monetary compensation or course credit from a midwestern university community in the United States. Some participants were also recruited via Amazon Mechanical Turk given the large sample size required. Participants needed to be native English speakers with corrected to normal vision to take part in the research. Data from participants were removed prior to analysis if their comprehension question accuracy (see Materials section) fell below 70% accuracy for all items, the criterion used by GGW. To stay faithful to the BFDA's protocol, if a participant being removed from analysis were to result in a sample size lower than the threshold needed to conduct either the first or nth analysis iteration, the analysis would have been postponed until another participant with adequate accuracy was run.

Materials

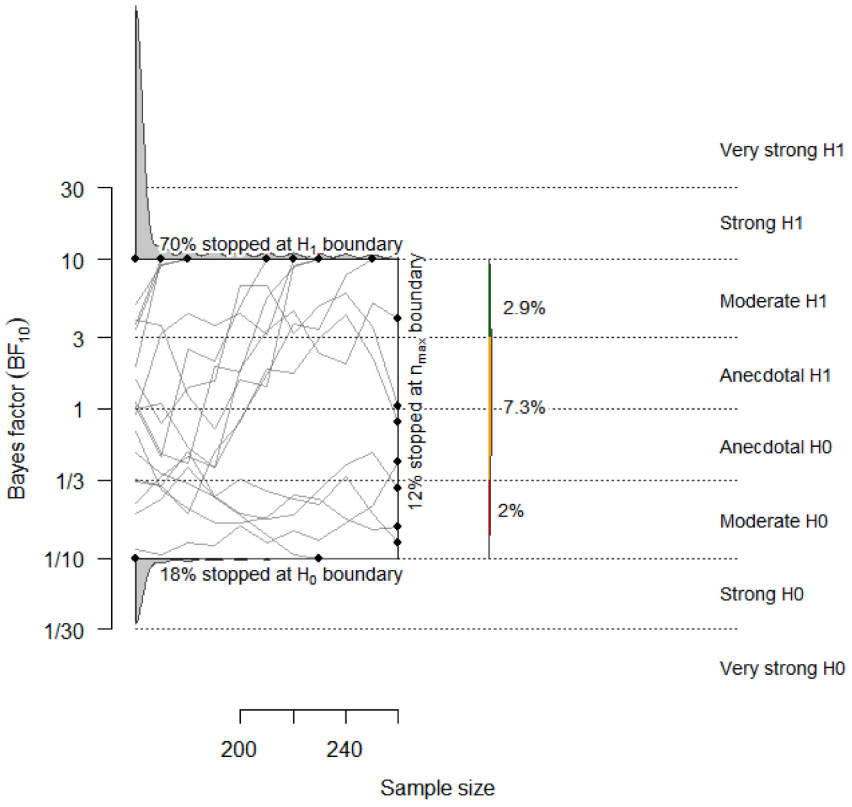
The target sentences were exactly those provided by GGW's Appendix A with indefinite subjects being replaced with definite subjects as was done for GGW's Experiment 2. The full list of stimuli can be found in the supplemental materials, and an example is provided in Table 2. Twenty experimental items were used, each participant only seeing one of four conditions per item for a total of five items per condition per participant throughout the course of the experiment. Additionally, thirty-five filler items were pseudo-randomly placed into the presentation order of stimuli. Because GGW do not provide details about the filler items other than the fact that many of them included definite NPs without explicit antecedents, we wrote new filler items (thirty-five like in GGW) to be used in this replication. All filler items consisted of two sentences logically connected in a discourse, mimicking the experimental items. Crucially, none of these filler items contained relative clauses and varied in terms of complexity. The full list of filler items can be found along with the experimental items in the supplemental materials. All items also contained a comprehension question asking about information not related to the relative clause or discourse reference (i.e., not related to the restrictive nature of the relative clauses). These comprehension questions were meant solely to be a task to keep participants engaged and were used as a criterion for data cleaning.

Procedure

Participants were recruited via word of mouth, flyers, class announcements, and through online subject pools like SONA or Amazon Mechanical Turk. The experimenter was in the same room as them during in-person data collection, but a cubicle wall from floor to ceiling separated the researcher and the participant during the actual experiment to align with COVID-19 protocols. The researcher and participant also wore a face mask to maximize precautions, and there was extra time added between participants for sanitation purposes. Although these steps created an environment that differs very slightly from the normal experimental procedure, we felt that these precautions had become the norm in society and therefore should not have imposed a substantial difference in data collection. 80 participants were initially recruited via Mechanical Turk, but only 40 were kept after data cleaning (see Results section), resulting in a final breakdown of 25% MTurk and 75% in-person participants. This rather surprising data loss encouraged us to switch the experiment to in-person data collection, thus the remainder of the participants were recruited for in-person participation. The MTurk experiment was identical once participants clicked to participate on the link – no instructions about the task other than that it was a "reading task for native speakers of English" were given. We also looked for differences between data collection sites and found that both sites produced comparable results (see Results section).

After signing a consent form, participants were seated at a computer and the researcher opened up the experiment online hosted by the Ibex Farm server (Drummond et al., 2016). Participants read the instructions and were then asked to complete eight practice items (also in the supplementary materials) before going on to the main task. The practice items gave accuracy feedback, but items in the main task did not. For both practice items and the main task, participants read through sentences in a word-by-word moving window self-paced reading paradigm (for a review, see Jegerski, 2013). Following a fixation cross appearing where the first word appeared, the full sentence appeared on the screen with underscores replacing all the words. As participants pressed the spacebar, the first word appeared on the screen, that word disappearing with the appearance of the next word. This process continued until the

a) Alternative Model



b) Null Model

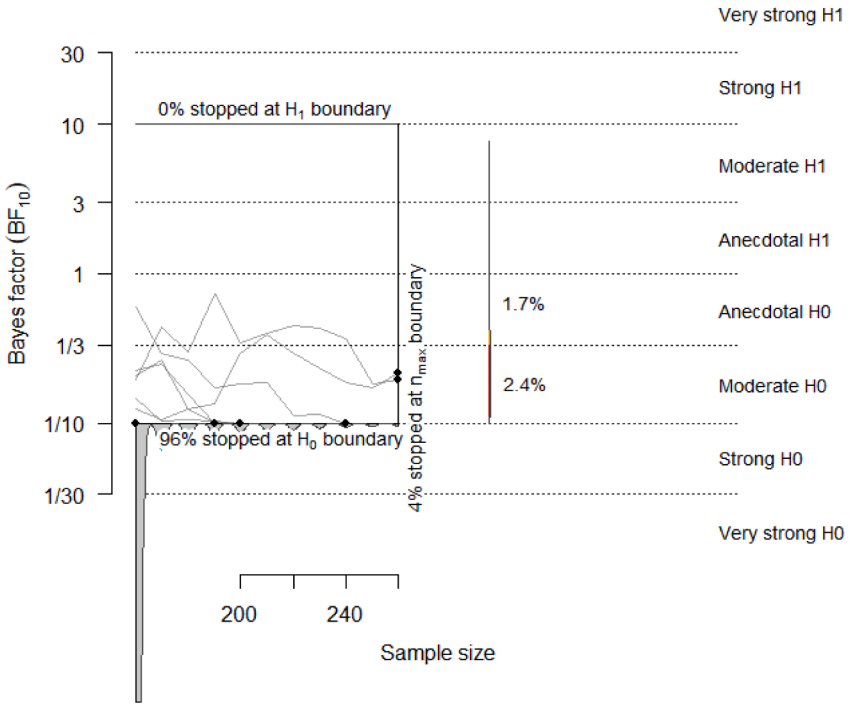


Fig. 1. a & 1b. Figures generated via the BDFA package in R. Each point is the ending BF of a simulated iteration with a minimum sample size stopping point of 160. The code can be found in the supplementary materials. Two simulations were run drawing samples from distributions of effect sizes assuming a null effect (Fig. 1b) or an effect size ranging from Cohen's F of .1 to .4 or Cohen's D of .2 to .8 (small to large effect sizes; Fig. 1a). The colored bars at the maximum sample stopping point of 260 reveal the distribution of BFs that had not surpassed the 1/10–10 boundaries. For the alternative model (1a), 88% of iterations stopped before the maximum sample stopping point of 260: 70% correctly stopped at a BF greater than 10 (around 75% greater than 1) while 20% stopped at a BF smaller than 1 (i.e. Type II error rate). For the null model (1b), 96% of iterations correctly stopped at a BF smaller than 1/10 while less than 1% stopped with a BF greater than 1 (i.e. Type I error rate).

sentence had been read. Like GGW, each sentence was presented as a new line to ensure the same physical location on the screen was used for the critical region across conditions. After reading through the sentences in each discourse, a superficial yes/no question about some element in the discourse appeared. This process repeated itself until all items had been read. At the end of the experiment, there was a demographic questionnaire asking for age, gender, educational background, language background, and a free-form optional response asking if participants noticed anything explicit about the items in the experiment. These were asked for either screening or descriptive reasons and we did not plan to incorporate them into inferential models.

Data availability

The data, stimuli, and analysis code for this study are openly available for download on OSF (DOI https://doi.org/10.17605/OSF.IO/E5WKA).

Analysis

Data cleaning procedures

We followed GGW's criterion of cleaning for reading times in that any critical region reading time greater than 5 SDs above the mean by participant was removed from analysis. Although this may be a rather conservative cutoff, and many studies incorporate a lower-bound cutoff for excluding reading time data as well, it is important to replicate the methods used by the original study's authors. For this reason, graphical representations and tables of conditional means and distributions are included in Table 3 and Fig. 2 below to make accessible and easily interpretable the nature of the data collected.

An initial 80 participants were recruited via Amazon Mechanical Turk; however, 40 of them were thrown out due to accuracy below 70%. This is somewhat alarming and is elaborated upon in the Discussion section. The remaining participants were recruited in-person, and no participant was thrown out for low accuracy after that. The reading time cutoff of 5 SDs resulted in a loss of less than 1% of the reading time data. The excluded participants' data are available in supplementary materials.

Reading time analysis

Because this is a replication attempt, the priors used for Bayesian inference were not highly informative. This is because we did not want the findings from the original study to sway the posterior derived from the models generated from the current data (Liu, 2019). Instead, mildly informative priors were used to allow for a wide distribution of possible values *a priori* while still assuming distributional properties of the expected data (e.g., Avetisyan et al., 2020). The priors can be found in Table 4 below. Since all reading time data were log-transformed to approximate a normal distribution, the parameters used in the priors should also be interpreted on a log-odds scale. For example, to allow for a wide range of possible mean reading time values, we set the intercept prior's distribution to the log-equivalent of a range from 0 to 10 s (10,000 msec).

Bayesian hierarchical models were run using the brms package (Bürkner, 2017) using eight Monte-Carlo chains over 5000 iterations (2500 burn-in/warmup). Log-transformed reading times were summed at the two words in the critical region and regressed onto fixed effects of Context Type and RC Type (sum coded: Null Context & Restrictive RCs = -.5, Supportive Context & Non-Restrictive RCs =.5), their interaction, and a continuous fixed effect of word length as a control. Random intercepts by participants and items along with random slopes of Context Type, RC Type, and their interaction by participants and by items were included as well. For clarity, the 'interaction' always refers to an interaction between RC Type and Context Type. The decision to log-transform differed from the original study but allows a better approximation of a normal distribution. All null effects found in the log-transformed model were also null in the raw model (omitted from this article for brevity). After deriving posterior estimates from these Bayesian models, BFs were computed via the bridge sampling method (Gronau et al., 2017) with 50,000 iterations (10,000 warmup) using the bridgesampling package (Gronau, Singmann, & Wagenmakers, 2020). This process involved comparing less complex models to more complex ones to determine the BF of each parameter. For example, to assess the BF of the interaction term, a null model was also fit to the data without this interaction as a fixed effect; the BF of the alternative model compared with the null model is thus a magnitudinal measure of probability for the interaction effect (i.e., how much more likely is the model with the interaction term to better account for the data than a model without the interaction term). Visualizations of the posterior distributions and their 95% credible intervals (CrIs) are reported in Table 5 and Fig. 3; however, these are meant only as a descriptive measure of the posterior estimates for each effect and do not reveal any probabilistic information in terms of model comparison. Interpretations of the crucial interaction effect instead rely on the BF value obtained from bridge sampling.

The model revealed a main effect of Context Type such that supportive contexts elicited faster reading times in the critical region. No main effect of RC Type was observed in the CrIs, and the interaction also included 0 in its CrI. The bridge sampling procedures resulted in very strong evidence in favor of the null hypothesis for the main effect of RC Type, very strong evidence against the null hypothesis for the main effect of Context Type, and strong evidence in favor of the null hypothesis for the interaction effect. The failure to replicate this interaction presents a major failure to replicate the central finding of GGW, suggesting that discourse does not project specific structures onto the parser.

Exploratory analysis

To explore whether in-person and MTurk participants differed in their reading behaviors, we ran a model testing for any such differences. Fig. 4 shows the distribution of data across conditions for in-person and MTurk participants after screening for accuracy. As mentioned in the Data Cleaning section, half of the participants who took part in the study via MTurk were excluded from analysis due to accuracy below 70%. This is striking for several reasons. First, screening measures were used, such as a 95% HIT approval rate and at least 50 HITs completed, to ensure that the respondents were seasoned workers. Additionally, participants were paid $7 for their time, which roughly translates to about $21 an hour given that many finished the task in roughly 20 min. So it appears that even when paying experienced workers a higher rate than is customary, the quality of data is much less robust than when collected in person. The field should reconsider which platforms to use based on performance metrics such as comprehension question accuracy, among others (see Eyal et al., 2021 for further discussion). After screening for accuracy, participants showed the same descriptive pattern of results, which was corroborated by a model run with fixed effects of RC Type, Context Type, Platform (sum coded: In-Person = -.5, MTurk =.5), and all interactions. The model, reported in Table 6, yielded no main effects or interactions other than the main effect of Context Type found in the previous model, suggesting the datasets are comparable.

Table 3
Means by condition.

RC Type	Context Type	Mean RT	SD
Restrictive	Null	838.114	779.837
Non-Restrictive	Null	827.857	813.121
Restrictive	Supportive	711.761	467.918
Non-Restrictive	Supportive	715.388	402.688

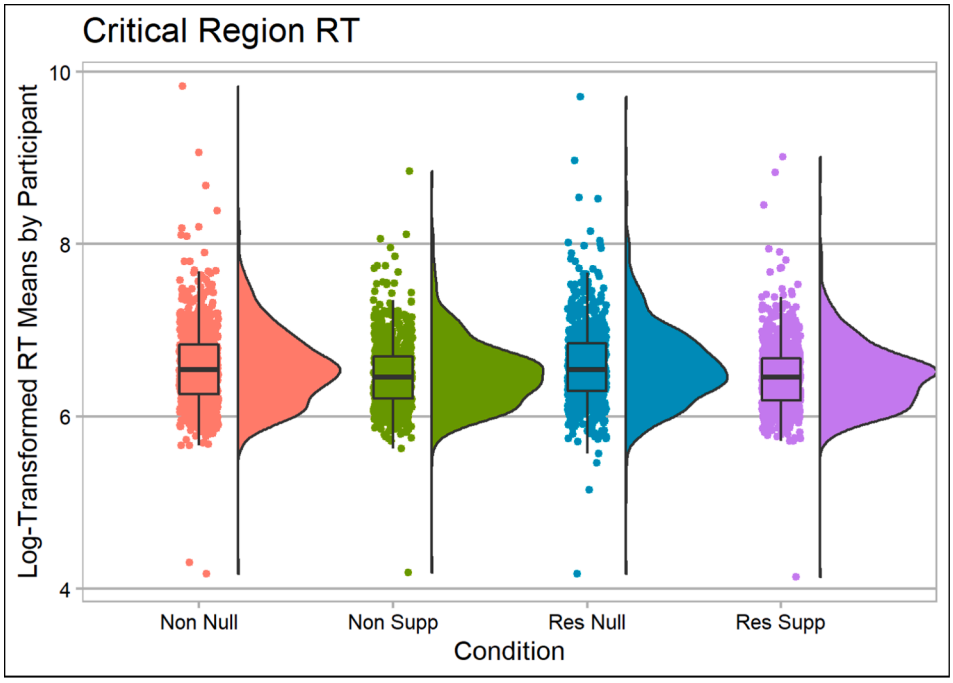


Fig. 2. Raincloud plots (Allen et al., 2019) depicting participant means of log-transformed reading times by condition (RC Type & Context Type).

Table 4
Priors for Bayesian Hierarchical Models.

Intercept	N(0, 9.21)
Betas	N(0, 1)
SDs	N(0, 1) *Half-Normal
LKJ (Random Effects Correlations)	2

Table 5
Model Output.

Effect	Estimate	SE	95% CrI	BF_{01}
RC Type	-.01	.01	[-.03,.02]	72.21
Context Type	-.10	.02	[-.14, -.07] *	.0006
Interaction	-.03	.02	[-.07,.02]	23.98

Discussion

Grodner and colleagues put forth three separate hypotheses concerning the interactivity of discourse sentence-level parsing: the Ambiguity Only Hypothesis, the Strong Interactive Mental Models Hypothesis, and the Weak Interactive Mental Models Hypothesis. The Ambiguity Only hypothesis states that context is only consulted in cases of sentence-level ambiguity. The Weak Interactive Mental Models Hypothesis states that discourse is consulted after sentence-level interpretations have been formed, whereas the Strong Interactive Mental Models Hypothesis predicts that discourse plays an early, anticipatory role in sentence-level parsing. An interaction between RC Type and Context Type such that restrictive RCs were read more slowly in null contexts and faster in supportive contexts, compared to non-restrictive RCs, is the crucial effect needed to support a strong version of the Interactive Mental Models Hypothesis. This would demonstrate that supportive discourse content can project specific structures onto the parser. The current study fails to replicate this interaction, suggesting discourse-level information does not play an early, dynamic role in syntactic parsing. On the other hand, the main effect of context suggests there is general facilitation from supportive contexts; however, it seems that the difficulty associated with a structure's implicit discourse structure does not affect the interactivity of discourse-level information on sentence-level processing. Thus, although the data show no evidence of inherently more difficult processing of restrictive RCs compared to non-restrictive RCs, it does suggest that discourse contexts may facilitate processing more generally. For this reason, our results fall more in line with the Weak Interactive Mental Models Hypothesis. To clear any confusion, the original hypothesis for the Ambiguity Only Hypothesis was that there would be no main effect of RC Type due to restrictive RCs inherent discourse complexity. Although this was not borne out in the data, we still found a main effect of context type; so we still find evidence for a weakly interactive account.

One issue with self-paced reading is that it proves difficult to extract temporal information since all reading behavior is subsumed in word-by-word reading times. For example, it is unknown whether the main effect of context was due to a general facilitation of having read previous material, or if there was structure-specific support for both types of RCs. On the other hand, since neither of these structures are strictly ambiguous, it seems that context has an effect on structures that are unambiguous, favoring a weakly interactive model over an ambiguity only model. It has been shown that comprehension of temporarily ambiguous sentences is sometimes modulated by discourse models while initial processing is not (Christianson & Luke, 2011; but see Dempsey & Brehm, 2020). Taken together, this might suggest that discourse-checking may occur at a later time, although it is unknown if this happens before or during sentence wrap-up effects. The Weak Interactive Mental Models Hypothesis seems to favor shallow processing models over more proactive, predictive models of sentence comprehension because it suggests the comprehenders' normal state is passive until ambiguous input triggers a search for additional information. Work over the past few years has shown that the majority of words are not predictable by their contexts, although their syntactic categories are moderately predictable (Luke & Christianson, 2016), and the current replication attempt takes this further in suggesting that discourse models do not project syntactic structures onto readers' expectations. In the words of Clifton and Staub (2008): "…parsing is not a matter of choosing between two interpretations. The parser must be able to deal with the arbitrarily large number of possible sentences in the language, constructing syntactic analyses and semantic interpretations on the fly, being prepared for the

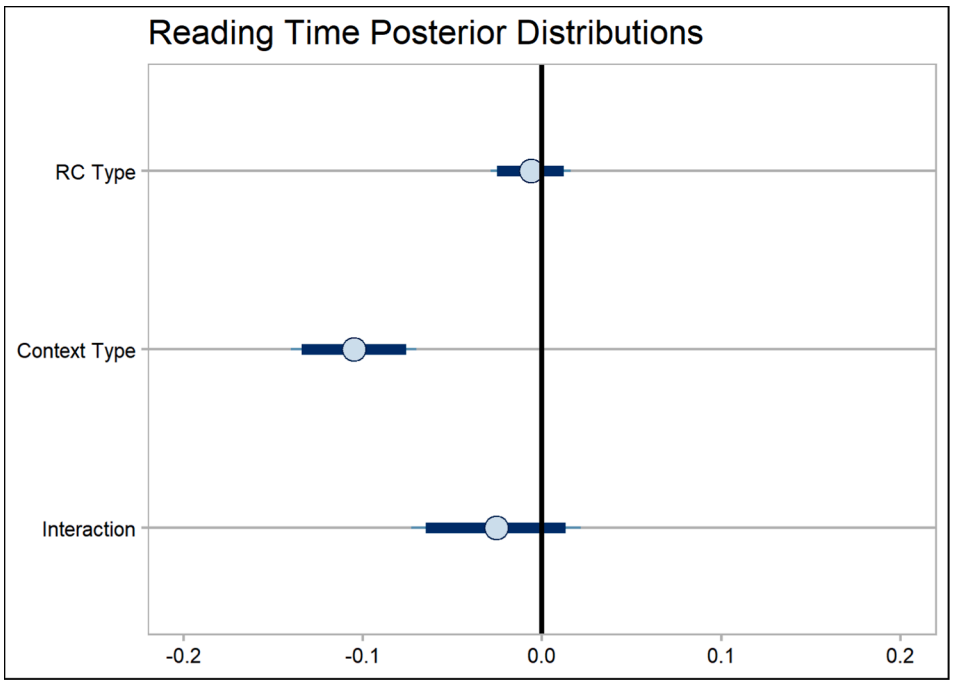


Fig. 3. Posterior distributions with 95% Credible Intervals.

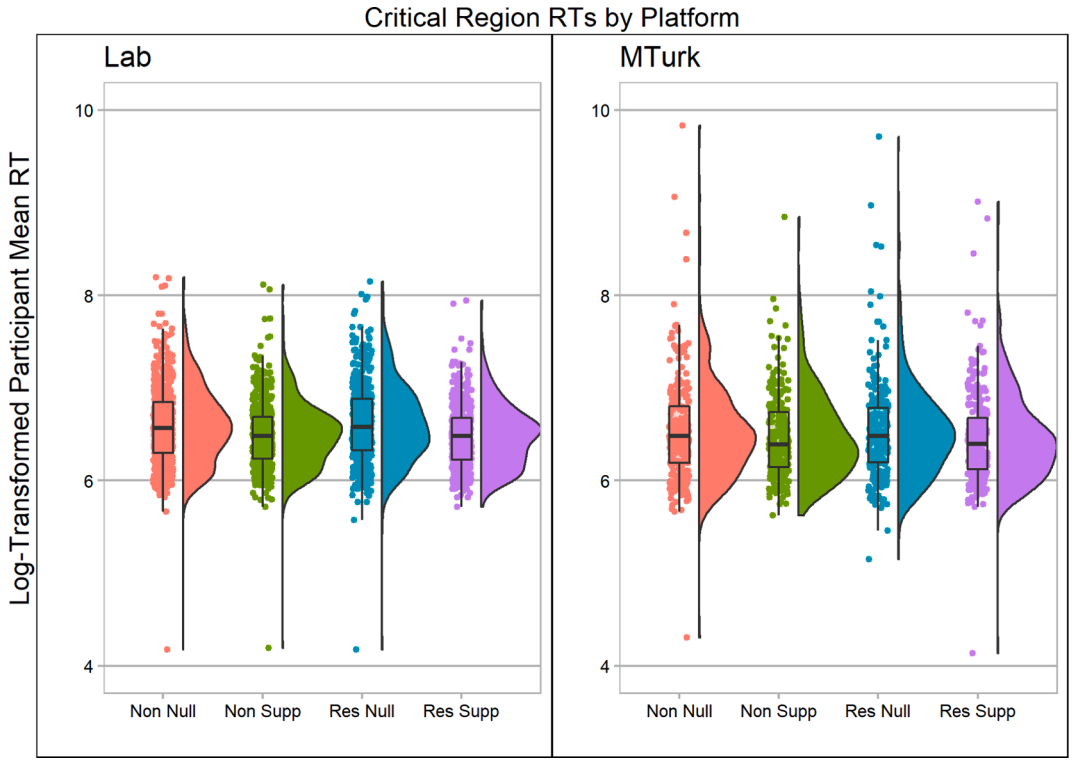


Fig. 4. Raincloud plots (Allen et al., 2019) depicting participant means of log-transformed reading times by condition (RC Type & Context Type) and by platform (in-lab vs. MTurk).

unexpected" (p. 243). Although a restrictive relative clause is certainly a proper continuation of contrastive referent sets, it could be that a contrastive referent set alone is not enough to constrain readers' expectations toward a specific structure, as a hasty commitment could cause processing difficulty downstream.

The failure to replicate here does not categorically reject the notion that readers can sometimes predict upcoming structures in a facilitatory manner; rather, it appears that discourse complexity alone is an insufficient trigger for doing so. As the proponents of Referential Theory argued, ambiguity seems to be sufficient in triggering facilitated reading times at the disambiguating region; however, this presents a dilemma: how can readers rapidly use discourse information to resolve an ambiguity when noticing said ambiguity takes time in and of itself? Taking a look back at sentences 1a and 1b again, context 1c establishes a

Table 6
Platform model output.

Effect	Estimate	SE	95% CrI
RC Type	-.01	.01	[-.03,.01]
Context Type	-.10	.02	[-.13, -.06]*
Platform	-.04	.07	[-.17,.09]
RC Type*Context	-.03	.03	[-.08,.02]
RC Type*Platform	-.02	.02	[-.07,.02]
Context Type*Platform	.05	.03	[-.00,.10]
3-Way Interaction	-.01	.05	[-.10,.09]

competitor set of "woman" referents that triggers a facilitation of the disambiguating region; however, just because the observed facilitation is at the disambiguating word does not mean that that is when the expectation is triggered. One could argue that, upon seeing 'the woman,' readers' preferences increase for interpreting 'that' as a relativizer interpretation of 'that' over its complementizer interpretation. So, it is likely not the case that participants are predicting ambiguity before encountering it; rather, encountering a contrastive referent set in discourse may help resolve the structural ambiguity before the "disambiguating" region. Since the word 'that' presents an ambiguity in and of itself (complementizer vs. relativizer), readers may call on context to arrive at a relativizer continuation because that fits better in the discourse model. Future work should employ natural reading methodologies such as eye-tracking to study the time-course of discourse influences on ambiguity resolution.

We mentioned in the introduction the possibility of using the Strong Interactive Mental Models Hypothesis as a way to test claims made by GE processing accounts and post-interpretive, algorithmic-only accounts. Our failure to replicate informs this debate in meaningful ways. For example, the data here can be reconciled with GE processing accounts in that readers do not use discourse in a proactive manner, preferring instead a "good-enough" (i.e., least-effort) interpretation that is largely algorithmic – there is no need to expend additional resources to facilitate a structure if it will not ultimately aid comprehension. A purely algorithmic argument, on the other hand, would need to explain why discourse information sometimes helps, the conditionality of which suggests it is not a part of the algorithmic parse but rather an experience-based heuristic employed to alleviate the burden of difficult or ambiguous structures.

We turn now to the implications of our failure to replicate at large. Although a Type II error rate is still technically possible, it is not probable given the *a priori* power analysis, so we will not focus on this and instead assume the current findings to be true. Therefore, in conjunction with our very close replication of the original study, it is likely that the failure to replicate is due to a Type I error in the original study. Such failure to replicate may be due to power, the lack of random effects used in the models, or some combination of those and other factors. With the growing popularity of Bayesian modeling, we can directly compare a null model with its alternative, which is particularly important for failures to replicate. For example, this allows us to say with a specific degree of probability that the null effect was true, thus making the probability that the original study was a Type I error probable. In frequentist statistics, a *p*-value above a certain threshold (e.g.,.05) does not provide evidence against the effect; rather, it simply fails to provide evidence for the effect. An important rule to remember is that the absence of evidence is not the same as the evidence of absence. For this reason, Bayes Factors are useful because they can provide probabilistic information for the null hypothesis, as was done for the current study, thereby providing direct evidence of absence. Even with high power, there is still a small chance for a Type II error, so replication attempts without a Bayes Factor will inevitably be scrutinized for this omission. For these reasons, we believe Bayes Factors should be a staple in future replication attempts.

In conclusion, the current failure to replicate suggests that discourse complexity alone does not trigger a proactive facilitation of certain structures, favoring instead an account of discourse influences on sentence comprehension where information from the discourse is relegated to later, integrative processes. Future work is needed to establish the time course of such influences and whether there are certain highly frequent discourse models that might trigger predictive processing.

CRediT authorship contribution statement

Jack Dempsey: Conceptualization, Methodology, Writing – original draft. **Kiel Christianson:** Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2022.104335.

References

Allen, M., Poggiali, D., Whitaker, K., et al. (2019). Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Res, 4*, 63.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191–238.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language, 112*, Article 104087.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(8), 1286.

Binder, K. S., Duffy, S. A., & Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language, 44*(2), 297–324.

Brehm, L., Jackson, C. N., & Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, 1-25.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1–28.

Caplan, D. (1972). Clause boundaries and recognition latencies for words in sentences. *Perception & Psychophysics, 12*(1), 73–76.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology, 69*(5), 817–828.

Christianson, K., & Luke, S. G. (2011). Context strengthens initial misinterpretations of text. *Scientific Studies of Reading, 15*(2), 136–166.

Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 538.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the syntactic processor. *Natural language parsing: Psychological, computational and theoretical perspectives, Cambridge University Press, New York*.

Cutter, M., Paterson, K., & Filik, R. (2021). Online representations of non-canonical sentences are more than good-enough. *The Quarterly Journal of Experimental Psychology*.

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review, 126*(6), 865–879.

Dempsey, J., & Brehm, L. (2020). Can propositional biases modulate syntactic repair processes? Insights from preceding comprehension questions. *Journal of Cognitive Psychology, 32*(5–6), 543–552.

Dempsey, J., Christianson, K., & Tanner, D. (2021). Misretrieval but not misrepresentation: A feature misbinding account of downstream attraction effects in comprehension. *Quarterly Journal of Experimental Psychology*.

Dempsey, J., Liu, Q., & Christianson, K. (2020). Convergent probabilistic cues do not trigger syntactic adaptation: Evidence from self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(10), 1906–1921.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience, 32*(8), 966-973.

Liu, Q. (2019). *Sensitivity analysis of Bayesian priors in replication studies.* (Unpublished master's thesis). University of Illinois, Urbana-Champaign.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28*(1), 1–11.
Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20.
Fedorenko, E., Piantadosi, S., & Gibson, E. (2012). Processing relative clauses in supportive contexts. *Cognitive Science, 36*(3), 471–497.
Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164–203.
Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15.
Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes, 56*(7), 485–495.
Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science, 44*(3), Article e12814.
Futrell, R., & Levy, R. (2017, April). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 1*, long papers (pp. 688-698).
Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360–1383.
Gernsbacher, M. A., Hargreaves, D. J., & Beeman, M. (1989). Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language, 28*(6), 735–755.
Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition, 95*(3), 275–296.
Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97.
Gronau, Q. F., Singmann, H., & Wagenmakers, E. (2020). bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software, 92*(10), 1–29.
Harrington-Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition, 46*(6), 864–877.
Jeffreys, H. (1939/1998). *The theory of probability*. Oxford University Press.
Jegerski, J. (2013). Self-paced reading. In *Research methods in second language psycholinguistics* (pp. 36–65). Routledge.
Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology, 69*(5), 1013–1040.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.
Levy, R. P., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *In Advances in Neural Information Processing Systems* (pp. 937–944).
Liu, Q. (2019). *Sensitivity analysis of Bayesian priors in replication studies.* (Unpublished master's thesis). University of Illinois, Urbana-Champaign.
Luke, S. G., & Christianson, K. (2011). Stem and whole-word frequency effects in the processing of inflected verbs in and out of a sentence context. *Language and Cognitive Processes, 26*(8), 1173–1192.
Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60.
MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology, 4*, 226.
Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology, 74*(1), 1–28.
Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition, 181*, 141–150.
Schad, D. J., Betancourt, M., & Vasishth, S. (2020). June 18). Toward a Principled Bayesian Workflow in Cognitive Science. *Psychological Methods. Advance online publication.*
Schönbrodt, F. D. & Stefan, A. M. (2019). BFDA: An R package for Bayes factor design analysis (version 0.5.0) Retrieved from https://github.com/nicebread/BFDA.
Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review, 25*(1), 128–142.
Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language, 69*(2), 104–120.
Clifton, C., Jr, & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass, 2*(2), 234–250.
Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48*(3), 542–562.
Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition, 36*(1), 201–216.
Traxler, M. J., Pickering, M. J., & Clifton, C., Jr (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language, 39*(4), 558–592.
Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*(6), 491–498.
Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779–804.
Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*(3), 169–176.