Contents lists available at ScienceDirect

Journal of Memory and Language

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

Journal of
Memory and
Language

Do readers maintain word-level uncertainty during reading? A pre-registered replication study

Michael G. Cutter [a,*], Ruth Filik [a], Kevin B. Paterson [b,1]

[a] University of Nottingham, United Kingdom
[b] University of Leicester, United Kingdom

ARTICLE INFO

Keywords:
Noisy-channel processing
Word-level uncertainty
Eye-movements
Reading
Sentence processing

ABSTRACT

We present a replication of Levy, Bicknell, Slattery, and Rayner (2009). In this prior study participants read sentences in which a perceptually confusable preposition (*at*; confusable with *as*) or non-confusable preposition (*toward*) was followed by a verb more likely to appear in the syntactic structure formed by replacing *at* with *as* (e.g. *tossed*) or a verb that was not more likely to appear in this structure (e.g. *thrown*). Readers experienced processing difficulty upon fixating verbs like *tossed* following *at*, but not *toward*. Levy et al. argued that this suggests readers maintained uncertainty about previously fixated words' identities. We argue that this finding has wide-ranging implications for language processing theories, and that a replication is required. On the basis of a Bayes Factor Design Analysis we conducted a replication study with 56 items and 72 participants in order to determine whether Levy et al.'s effects are replicable. Using Bayesian statistical techniques we show that in our dataset there is evidence against the existence of the interaction Levy et al. found, and thus conclude that this study is non-replicable.

According to noisy-channel accounts of human sentence comprehension, readers maintain a level of uncertainty about the perceptual input they have received from a sentence (Gibson et al., 2013; Levy, 2008; Levy et al., 2009). The basic idea behind this theoretical approach is that the language processing system takes account of the noise and inaccuracy inherent in perceptual processing, and therefore operates on the assumption that encoded information is unlikely to perfectly represent the text on the page. One consequence of this noisy-channel processing is that readers do not always identify words in an all-or-nothing manner, and that instead they hold a level of uncertainty about the identity of each individual word in the sentence. For example, upon encountering the word *at* readers may only assign a probability of 0.85 to the possibility of the word actually being *at*, with probabilities of 0.10 and 0.05 being assigned to the possibility of the word being the perceptually similar *as* or *and*, respectively. These beliefs are refined as readers gather more information from later in the sentence, such that one possibility for the earlier word may be ruled out by later information, while other possibilities are assigned an increasing level of certainty. Within this account of sentence processing, processing difficulty at any one point in time is the product of the extent to which a new piece of information alters the level of belief about the contents of the sentence. In the present paper, we attempt to replicate a study which represents the main body of evidence for this proposal.

Before outlining the study that we replicate, it is worthwhile considering some wider theoretical implications of Levy's (2008) proposal that readers do not identify words with full certainty, at least on first-pass reading. This will help establish the importance of determining whether the key evidence for this proposal is replicable in a highly powered eye movement experiment. The possibility that readers can remain uncertain about the identity of a word that they have just read would seem to have far-reaching consequences for the way in which various component processes of reading are assumed to operate. For example, most models of visual word identification very much depend upon the idea that a word is definitively recognized. In particular, models based around an interactive-activation approach (McClelland & Rumelhart, 1981) assume that orthographic information extracted from a word initially activates a range of candidate lexical items in the mental lexicon, with these items competing for selection in a winner-takes-all manner. Within the interactive-activation framework, only one lexical item ultimately is selected, triggering the retrieval of information about its meaning and syntactic class from memory. It is unclear how models based on such an architecture could account for ongoing uncertainty

[*] Corresponding author at: School of Psychology, University of Nottingham, University Park Campus, Nottingham, NG7 2RD, United Kingdom.
E-mail address: michael_cutter@hotmail.com (M.G. Cutter).
[1] Present address: Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, United Kingdom.

https://doi.org/10.1016/j.jml.2022.104336
Received 31 December 2020; Received in revised form 29 April 2022; Accepted 1 May 2022
Available online 13 May 2022

about the identity of a word, since this presumably would require continued competition between the potential lexical items while readers are processing subsequent words in the text. Models based upon such an architecture include the Dual-Route Cascaded model (Coltheart et al., 2001), the Multiple Read-Out model (Grainger & Jacobs, 1996), the Spatial Coding Model (Davis, 2010), and the CDP+ model (Perry et al., 2007; for a recent discussion of the dependency of lexical processing models on interactive-activation frameworks see Reichle & Schotter, 2020). Consequently, it is unclear how these models of word identification could be reconciled with a theory of sentence comprehension in which word identity remains uncertain. Even a model of word recognition which has similar theoretical underpinnings to the noisy-channel account – known as the Bayesian Reader model (Norris, 2006) with letter identity and position encoded via a noisy channel (Norris & Kinoshita, 2012) – assumes that in the case of normal reading a single candidate for word identity is settled upon. Thus, this model would also struggle to explain word identity remaining uncertain beyond first-pass reading, although assumptions about optimal behaviour within this Bayesian model could be modified to allow subsequent contextual information to trigger the revision of a lexical selection decision downstream of the word in a text.

Beyond the issue of word identification itself, many models of syntactic parsing would seem to leave little room for readers to maintain uncertainty about word identity. Indeed, many such models not only assume that the input to the parser is a fully identified word, but also that even when multiple grammatical analyses of portions of text containing this word are possible, readers will rapidly commit to one of these analyses rather than maintain uncertainty (e.g. see van Gompel et al., 2001). Finally, a highly influential model of eye-movement control during reading, the E-Z Reader model (Reichle et al., 2003; 2009), assumes that attention is deployed towards only one word at a time during reading, with attention not progressing to the next word before the current word is fully lexically identified, with the consequence that words are identified definitively, by retrieving a single item from lexical memory during reading. Furthermore, this word is then instantly integrated into the unfolding sentence representation, with a failure to integrate this word prior to fully identifying the following word leading to processing difficulty. It is unclear how such a model could be reconciled with a system in which words are rarely 'fully' processed, and instead processed with some residual uncertainty.

Given the challenge that word-level uncertainty poses to a range of key theoretical assumptions about foundational aspects of the reading process, it is important to ascertain the strength of the evidence for this phenomenon. The key evidence for word-level uncertainty during sentence reading comes from a study by Levy et al. (2009). In their experiment, participants read sentences such as 1a-1d below.

1a) The coach smiled at the player tossed the frisbee by the opposing team.
1b) The coach smiled at the player thrown the frisbee by the opposing team.
1c) The coach smiled toward the player tossed the frisbee by the opposing team.
1d) The coach smiled toward the player thrown the frisbee by the opposing team.

In 1a and 1b, the preposition *at* appears early in the sentence, while in 1c and 1d this preposition is replaced with *toward*. While both words are essentially interchangeable in terms of their role, they differ in terms of how perceptually confusable they are with other words. *At*, as outlined above, could hypothetically be confused for either *as* or *and*, while *toward* does not have any near-lexical neighbours with which it might be confused. As such, the argument goes, readers will maintain a level of uncertainty about the preposition *at* (but not *toward*). Due to this uncertainty, information encountered downstream is able to shift the reader's beliefs about the identity of this word, resulting in a level of processing difficulty determined by the size of the belief change. Specifically, if readers encounter a verb that can be treated finitely (e.g. *tossed*), this should lead to a shift in probability away from *at* being treated as *at*, and towards it being treated as *and* or *as*, due to a finite verb being more likely in the syntactic structures formed by substituting *at* (e.g. *The coach smiled as the player tossed the frisbee…*). In 1a and 1c *tossed* can indeed be treated as a finite verb, while *thrown* in 1b and 1d cannot be treated in this way. As such, within Levy's model, readers should experience greater processing difficulty upon encountering *tossed* rather than *thrown* when these words are preceded by *at* earlier in the sentence. In other words, readers should experience most processing difficulty reading the verb in sentences like 1a compared to the verb in sentences 1b-d. This occurs due to the fact that *tossed* causes a large probability shift away from a context in which *at* is in fact *at*, and towards one in which it was either *as* or *and*, while *thrown* does not. It should be noted that, within this theory, *tossed* also causes a level of processing difficulty relative to *thrown* even when the preposition *toward* is used, since it causes a shift from the level of belief in the actual context towards one in which it is assumed a word was missing from the context (e.g. *The coach smiled toward the player* ***who*** *tossed the frisbee*); however, this shift in belief is smaller than when *at* is used. Consequently, Levy's model predicts processing difficulty when a potentially finite verb (*tossed*) appears rather than a non-finite verb (*thrown*), with this processing difficulty being greater when the verb is preceded by the preposition *at* rather than *toward*.

In their study, Levy et al. (2009) observed evidence in favour of their hypothesis in various measures of eye movement behaviour. For example, when the critical word could be treated finitely (e.g. *tossed*) and was preceded by *at* as in 1a, readers were more likely to make regressive eye movements away from the critical verb (0.21 vs. 0.12, 0.10, and 0.14 regression probability for 1b, 1c, and 1d), took longer between first fixating the verb and moving further to the right in the sentence (476 ms vs. 399, 399, and 409 ms), were more likely to make more regressions back to the preposition upon fixating the verb (.36 vs. .31 for all other conditions), and answered comprehension questions less accurately (62% correct vs. 72%, 70%, and 69%). While there was no evidence of word-level uncertainty in earlier measures of eye-movement behaviour which do not take regressions into account, there was a main effect of verb ambiguity in gaze durations on the verb, showing that there is some difficulty related to processing the ambiguous verb regardless of the preposition's form. Thus, this study does indeed suggest that readers maintain uncertainty about the identity of the preposition, and that when readers encounter information that shifts their beliefs in favour of the less likely option they experience processing difficulty, and attempt to resolve the uncertainty by looking back to earlier parts of the sentence.

Given the implications of the word-level uncertainty aspect of the noisy-channel account of sentence processing outlined above, it is vital to establish that the key evidence for this phenomenon is replicable. While independent evidence exists for the idea that readers make post-perceptual inferences about missing words (e.g. Gibson et al., 2013; 2017), none exists specifically regarding word-level uncertainty. In a recent investigation of aging effects on language processing, Cutter et al. (2022) attempted to replicate Levy et al.'s experiment with a group of 60 young and 60 older adults, using self-paced reading. While Cutter et al. observed a main effect of verb ambiguity on self-paced reading times at the verb, they did not observe any evidence of the preceding preposition affecting the size of this verb ambiguity effect, with a Bayesian analysis suggesting that their data represented evidence in favour of a null effect of the interaction between preposition form and verb ambiguity. Thus, they failed to replicate Levy et al.'s key finding, showing no evidence of readers maintaining word-level uncertainty. Furthermore, they also failed to replicate the effect of sentence type on comprehension rates, with any trend towards an interaction actually involving a greater effect of verb ambiguity when the preceding preposition was *toward* rather than *at*. Broadly speaking, there are two potential explanations for this replication failure. The first is that the effect genuinely does not replicate, which would be highly problematic for the idea that people maintain uncertainty about word identity during reading. Over the last several years an increasing level of attention has been given to the fact that many psychological studies do not replicate (Open Science Collaboration, 2015), with studies that report findings based on small sample sizes being particularly unlikely to replicate and to largely overestimate any effect sizes that are genuinely present in the population (Vasishth et al., 2018). In the case of the Levy et al. study a relatively small sample was used, with 24 items and 40 participants. As such, it may be that the statistically significant effect in this study simply does not replicate, with the original observation being largely driven by noise in the relatively small sample.

A second explanation is that the lack of a critical interaction in the Cutter et al. (2022) study was a consequence of methodological factors. Specifically, Cutter et al. recorded self-paced reading times as opposed to tracking eye movements during reading. As outlined above, Levy et al. found the strongest evidence for their key interaction in measures taking account of regressive eye-movements, such as go-past times on– and regressions out of– the critical word. As a measure of reading time, go-past time includes the amount of time between a reader's eye first landing on the critical verb, up until the eye progresses to a word to its right. Crucially, this means that the measure is influenced by time spent re-reading earlier portions of the sentence. In the case of the non-cumulative self-paced reading method used by Cutter et al., it is not possible for readers to make regressions to re-read earlier parts of the sentence upon encountering the target verb. As such, the lack of interaction in Cutter et al.'s study may simply have been due to the use of a methodology that does not allow for re-inspection of earlier parts of the text, rather than readers not experiencing word-level uncertainty. We will delay a detailed discussion of the theoretical implications of Levy et al.'s effects not appearing in self-paced reading until we are able to determine whether or not these effects can in fact be replicated in a large-scale eye-tracking experiment. Assuming that Levy et al.'s original effects do replicate, this would suggest there is something about self-paced reading which prevents readers from maintaining word-level uncertainty, while a failure to replicate Levy et al.'s findings would suggest that it is simply the case that readers do not maintain word-level uncertainty at all.

In addition to our own self-paced reading data, an eye-tracking study conducted by Christianson et al. (2017) is interesting to consider in relation to Levy et al.'s study. In this study, participants read sentences such as "The other team interfered with the player [who was] tossed the ball." Much like the items used by Levy et al., these sentences contain a sequence of words (i.e. *the player tossed the ball*) which can be treated as a main clause in certain syntactic structures, but in the context of this sentence should only be interpreted as a reduced relative clause. Where Christianson et al.'s stimuli differ is that the vast majority of their items did not contain a perceptually confusable preposition such as *at*,[2] and that the verb ambiguity effect was assessed by comparing sentences including vs. excluding the relativizer *who was*, rather than through comparison to an alternative unambiguous verb (e.g. *thrown*). Given that Levy et al. observed no effect of verb ambiguity on go-past times in the absence of a perceptually confusable preposition, one might expect that no, or at least very little, effect of ambiguity should have been observed by Christianson et al. However, Christianson et al. did observe such an effect, with an effect of 508 ms in one experiment and 188 ms in another experiment. While the existence of such an effect is not in and of itself problematic for noisy-channel models of sentence comprehension, it does raise questions about whether the lack of effect in sentences not featuring the word *at* in Levy et al.'s study is replicable, which leads us to further question whether their interactive pattern of results will replicate in a large-scale eye-tracking study.

In the present replication study we aimed to determine whether Levy et al.'s critical interaction does replicate in a large-scale eye-tracking study. We present a replication study in which we followed an identical design to Levy et al., while increasing the number of experimental sentences from 24 to 56, and the number of participants from 40 to a minimum of 72. The key effect which we aimed to establish the replicability of was the interaction between preposition form and verb type observed by Levy et al. in go-past time. We use Bayesian statistical methods to determine whether, within this sample, we have evidence in favour of or evidence against this interaction, through the calculation of Bayes factors (see Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012). As will be shown below, simulations within the framework of a Bayes Factor Design Analysis (Schönbrodt & Wagenmakers, 2018) suggest that our sample size is large enough to avoid observing misleading evidence. Furthermore, we determine an estimate of the size of any effect which is observed, alongside 95% credible intervals around this estimate and the probability of the effect being greater than certain values. While our study was designed around detecting any potential effects in go-past times, we also examine a number of other dependent variables, outlined in our proposed analysis.

Data availibility

A registered, permanent version of our data, materials, and analysis code can be found at https://osf.io/vd32e.

Method

Design & stimuli

We presented participants with stimuli in a 2 (Preposition Form: *at* vs. *toward*) × 2 (Verb Ambiguity: e.g. *tossed* vs. *thrown*) design, using sentence stimuli similar to 1a-d above. In their original study, Levy et al. presented participants with 24 items, making for a total of six per condition. In our study we used these original 24 items, alongside 32 that we created ourselves, making for a total of 56, with 14 items per condition. It should be noted that six of the items used by Levy et al. have been altered slightly for the current experiment, due to concerns that the events described in the original items may seem unusual to the British-English speaking participants we recruited. Leaving such oddities in the items may add unnecessary noise to reading time data. As an extra level of control over our new stimulus items, which was not included in items designed by Levy et al., we ensured that for each item the ambiguous and unambiguous word were matched for length. While this is not technically necessary for investigating the interaction effect, we took this step to reduce extra sources of variance in our own experiment. The complete set of items we used are available at https://osf.io/87fg2//, alongside comprehension questions and a list of changes we made to some of the Levy et al. items. Subsequent to the acceptance of our Stage 1 report we made some slight changes to our stimuli, approved by the editor on 21/10/2021. These changes were made in order to shorten four of our experimental items and one filler item, so that they could fit on a single line of our computer monitor while allowing us to use text of an appropriate size. Details of the changes made can be found at https://osf.io/2zm9k/.

It was important to ensure that the target verbs in our new sentences were similar to those used by Levy et al., in terms of a) the ambiguity of the verbs treated as ambiguous and b) the unambiguity of the verbs treated as unambiguous. To this end, we retrieved the instances per million from the British National Corpus (BNC Consortium, 2007) of how often each verb was used in a simple past tense form and how often each verb was used as a past participle. We then determined out of the total of these two counts, the proportion of the time that the verb was used as a past participle. For the ambiguous verbs from Levy et al., the

[2] Specifically, 67.5% of items included no preposition at all, and only 10% included the preposition *at*, with the remaining items featuring prepositions such as *about* (2.5%), *on* (2.5%), *up* (2.5%), *with* (5%), and *to* (10%).

mean proportion of use as a past participle was 0.375 (range 0.037 – 0.855) while for our new items it was 0.396 (range 0.026 – 0.996). Thus, on average the ambiguous verbs from the original study and the new ambiguous verbs in our own items were similar. We also queried the British National Corpus to ensure that the unambiguous verbs we used were actually unambiguous, in that they are never used as a simple past tense verb. Of our 32 new items, none were ever used as a simple past tense verb. Surprisingly, one of the items used in the Levy et al. study did very occasionally (i.e. 0.7% of the time) appear in the simple past tense. This was the word *sung,* the simple past tense form of which is actually *sang*. Despite this issue we chose to keep this word in, due to it forming part of Levy et al.'s stimuli set rather than being one of our own new items.

These sentences were presented amongst 84 filler items. This maintained the ratio of items to fillers used by Levy et al., while more than doubling the number of experimental items. Of these filler items, 44 were the same as those used by Levy et al. in their study. Eight of these were presented at the start of the experiment to familiarise participants with the experimental procedure. The remaining 40 filler items were used as a form of positive control, in order to demonstrate that even if our participants do not show the effects first demonstrated by Levy et al. (2009) there is at least evidence in their eye-movement behaviour to indicate that they did exhibit a psycholinguistic effect which is highly replicable. Specifically, these filler items included a frequency manipulation, in which a word within the sentence was manipulated to either be high- (e.g. *town*) or low-frequency (e.g. *cove*), with the eyes typically spending longer on the low-frequency as opposed to high-frequency words (e.g. Rayner & Duffy, 1986). Each participant read 20 items in the low-frequency condition and 20 items in the high-frequency condition. It should be noted that all of these filler items had relatively simple syntactic structures, as did those in the Levy et al. study. Thirty-nine of these filler items came from a study by White (2008) which showed reliable frequency effects, with the remaining item being designed for the current study. We assess the effect of this positive control manipulation using the same statistical methods as for our main experiment, outlined below, testing for effects in first fixation duration and gaze duration on these words. Twenty-six of the 84 filler items were followed by comprehension questions. All fillers are available alongside the main experimental items.

The order of stimulus presentation was randomised, with the constraint that no two experimental items were presented consecutively. Items were rotated between the four experimental conditions across four stimulus lists within a Latin Square, such that no one participant saw an item in more than one condition.

In addition, each sentence in the original Levy et al. study was followed by a yes/no comprehension question, with these questions probing various aspects of the sentence's meaning. Our sentences were also followed by comprehension questions, which were designed to probe comprehension in the same ways as those used by Levy et al. Specifically, Levy et al. used eight different types of comprehension question (see Supporting Appendix to Levy et al. for details), with four of these directly probing understanding of the reduced relative verb and the remaining four types probing other aspects of the sentence. For example, for the sample sentence above the possible question types would have been.

1) Did the player toss/throw a frisbee?
2) Did someone toss/throw the player a frisbee?
3) Did the player toss/throw the opposing team a frisbee?
4) Did the opposing team toss/throw the player a frisbee?

There were four instances of each of these types of questions in the Levy et al. study. We used these same questions for the Levy et al. stimuli in the current study, and we used six of each type for our own new stimuli. The remaining four types of question were.

5) Did the coach smile?
6) Did someone smile at the player?
7) Was the coach tossed/thrown a frisbee?
8) Was the player smiled at?

There were two instances of each of these types of question in Levy et al.'s stimuli. We used the same question types for the items from the Levy et al. study, and used each of these four types of question twice in our new stimuli. The questions can be viewed alongside our stimuli at https://osf.io/87fg2//. To clarify, each experimental item was only ever presented with a single comprehension question type, with the question type varying across items.

Apparatus

Sentences were presented on a single line on a BenQ XL2430 24″ monitor in Courier New font, using font size 23 at a viewing distance of 97 cm, with 1 degree of visual angle containing approximately 3.2 characters. Eye-movements were monitored using an SR-Research EyeLink 1000 running at 1000 Hz.

Procedure

Prior to arriving for the experiment, participants were sent an information sheet, and a web-link at which to provide informed consent for the experiment. Participants took part in the eye movement experiment individually. Upon arrival, each participant was seated in front of a desktop mount EyeLink 1000 and the monitor on which we presented our stimuli. Participants were calibrated using a three-point horizontal calibration grid; if a participant had an average error greater than .30 or any individual error above .50 during a validation procedure, they were re-calibrated until the error dropped below these values. During the experiment, each trial consisted of 1) a drift check in the centre of the screen, 2) a drift check in the position of the first character of the sentence, 3) a gaze contingent box in the same position as the second drift check, 4) the experimental sentence itself and 5) a comprehension question. If a participant returned an error above .40 on one of these drift checks on two consecutive trials they were recalibrated. Participants were asked to read the sentences silently for comprehension, and asked to minimise blinking while reading each sentence. They used a computer mouse with two buttons to answer yes–no comprehension questions; clicking the right button to answer *no* and the left button to answer *yes*.

Participants

We collected and analyzed data from 72 native speakers of English with normal or corrected to normal vision. These participants were young adults (mean age = 18.9; minimum age = 18; maximum age = 31; 65 female), and naïve as to the purpose of the experiment.

Sample size and planned analysis

We analyse our data using Bayesian statistical techniques. There are two main aspects of this analysis. One was to estimate the size of any effects we may observe, by fitting Bayesian mixed models to our data. We examine the estimates from these models, as well as 95% Credible Intervals, and the probability of the effect of each variable being above certain values. More detail on this aspect of our analysis can be found further below.

The second aspect of our analysis, and the aspect around which we based our sampling plans, was to calculate Bayes factors (see Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012) in order to determine whether we have a) evidence in favour of the interaction between preposition form and verb ambiguity originally observed by Levy et al. in go-past time, or b) evidence against the existence of this interaction.

Table 1
A Table Displaying the Outcome of our Bayes Factor Design Analysis, in Terms of the Proportion of Iterations on which we would Observe a Certain Level of Evidence in each Hypothetical Population.

	Large interaction	Small interaction	Null Effect
Misleading	.002	.035	.004
Uninformative (2.99–1/3)	.011	.149	.059
Moderate Evidence	.027	.125	.159
Strong Evidence	.032	.103	.777
Very Strong Evidence	.044	.134	.000
Extreme Evidence	.884	.455	.001

In brief, a Bayes factor provides a ratio of a dataset's marginal likelihood under two differing statistical models, such that it is possible to infer which statistical model is more likely to describe the processes that generated the data. The value of the Bayes factor represents the ratio of evidence for one model relative to the other, with a Bayes factor comparing Model A to Model B ranging from zero (representing evidence in favour of Model B) all the way to infinity (representing evidence in favour of Model A). Some scholars (e.g. Jeffreys, 1961; Lee & Wagenmakers, 2013) divide this possible range of Bayes factors into distinct evidential categories, with each category representing a different strength of evidence for each model. Specifically, values between 1/3 and 3 are treated as evidence in favour of neither hypothesis. Values from 3–10, 10–30, 30–100, and greater than 100 are treated as moderate, strong, very strong, and extreme evidence in favour of $H_{1,}$ while values of 1/3–1/10, 1/10–1/30, 1/30–1/100, and smaller than 1/100 are treated as the same categories but in favour of H_0. Thus, this technique will allow us to determine if our data supports Levy et al.'s original finding (i.e. H_1; evidence for a model including an interaction in go-past time), or the absence of their original effect (i.e. H_0; evidence for a model without an interaction in go-past time).

In order to determine our sample size, we performed a Bayes Factor Design Analysis (Schönbrodt & Wagenmakers, 2018). The purpose of this was to determine a minimal sample size at which we could be confident of obtaining a Bayes factor that was not misleading, in that it would not represent evidence for H_0 if Levy et al.'s effects truly exist in the population, or that it would not represent evidence for H_1 if readers turn out not to maintain word-level uncertainty during reading. While our proposed minimal sample size may still return a Bayes factor that is uninformative and represents evidence in favour of neither model (i.e. 1/3–3), this is not problematic – using Bayesian methods, we can simply increase the sample size until our model provides conclusive evidence in either direction.

To perform the design analysis, we repeatedly simulated data from several hypothetical populations, each representing a different effect of word-level uncertainty that could feasibly be present in the wider population. The effect we focussed upon in these simulations was the interaction between preposition form and verb ambiguity observed in go-past times by Levy et al., with this arguably being the most important aspect of their study to replicate.

The simulation of our data was based on an approach and R script used by Von der Malsburg and Angele (2017) in order to simulate eye movement measures in an investigation of false positives in reading research. In this script, the user provides a set of parameter values, including the number of participants, number of items, standard deviations for random intercepts for first fixation durations for subjects and items, and the mean and standard deviation of the first fixation duration. On the basis of just these parameters it is possible to simulate first fixation durations. In order to calculate gaze durations the user sets the proportion of trials on which participants should make a first-pass refixation on the target word, and the mean duration and standard deviation of any first-pass refixations that are made. These first-pass refixations are added to a proportion of the simulated first fixation durations to derive gaze durations. Finally, in order to calculate go-past times the user sets the proportion of trials on which participants should make a regression out of the target word, and the mean and standard deviation of the amount of time between a regression being made and a fixation being made to the right of the target word (i.e. post-regression re-reading time). These values are added to the simulated gaze duration data in order to derive go-past times. We took this approach to simulating later measures since it more accurately represents the characteristics of these measures (i.e. trials in which multiple fixations/regressions were made adding to the tail of the distribution) than simply simulating a normal distribution using a particular mean and standard deviation.

For each reading time measure, values were sampled from a log-normal distribution with the geometric mean and standard deviation set for that particular measure. In each of our simulations, we separately set the parameter values for each of the four conditions with the goal being to obtain a pattern of simulated means representing the means that should be present in the hypothetical population. The number of sentence stimuli in all simulations was set to 56, with this being the total number we have available to present. The standard deviation of random intercepts was set to 1.129 for subjects and 1.0765 for items. The standard deviation for first-fixation durations, first-pass refixation durations, and post-regression re-reading times were set to 1.36, 1.54, and 1.75, respectively. The probability of participants making a second first-pass fixation was set to .245. These values were all set somewhat arbitrarily, since this information was not readily available from the study we are attempting to replicate. However, they are within the range of reasonable parameter values used by Von der Malsburg and Angele (2017) in their previous simulations of eye-tracking data. Our scripts are available at https://osf.io/87fg2/.

The first hypothetical population that we simulated was one in which the effects observed by Levy et al. were real and accurately represented the size of the effect at the population level. The parameter values that we used to obtain a similar pattern of means to Levy et al. are shown in Table A1.[3] Having set these parameter values we then used them to generate a hypothetical dataset with a particular number of participants. Once the data set was generated we used the lmBF function from the BayesFactor (Morey & Rouder, 2018) package in R (R Core Team, 2022) with default prior values to obtain a Bayes Factor for 1) A model in which preposition form and verb ambiguity interacted, and 2) A model in which there were only additive effects between these variables. These models also included random intercepts for participants and items. The dependent variable in these models were log-transformed go-past times. We then divided the interactive model by the additive model in order to determine whether there was evidence in favour of H_1 or H_0 in the simulated data set. We repeated this process 1000 times to produce unique data sets based on the population parameters defined in Table A1, and determine the proportion of samples for which we gained a) evidence correctly in favour of an interaction (i.e. a Bayes factor above 3), b) uninformative evidence (i.e. a Bayes factor between 1/3 and 3), or c) misleading evidence (i.e. a Bayes factor below 1/3). Table 1 below displays the proportion of Bayes factors obtained in each evidential category outlined above for each hypothetical population, when assuming 72 participants. While we performed simulations with fewer participants, we choose to only present the results for the number of participants we decided to test on the basis of these simulations. In addition, Table 2 displays the mean value produced in each condition across all of our individual simulations, to demonstrate that our selected parameter values did produce data in line with the hypothetical population.

[3] Note that it was not simply the case of setting the means in the parameter value to be identical to those observed by Levy et al., since the mean taken as input in the script we adapted was in fact the geometric mean. As such, it was necessary to engage in a process of trial and error while using multiple potential values until we settled on a value which resulted in a highly similar outcome to that observed by Levy et al.

Table 2
The Average of the Reading Measures Presented Across all Simulations of each Hypothetical Population.

	First Fixation Duration	Gaze Duration	Go-Past Time	Regression probability
		Simulation 1 – Levy et al.'s effects		
A-A	285	357	478	.210
A-U	280	323	400	.120
T-A	285	359	400	.100
T-U	285	343	410	.141
		Simulation 2 – Reduced Interaction		
A-A	285	357	455	.169
A-U	280	323	400	.120
T-A	285	360	400	.100
T-U	285	343	409	.140
		Simulation 3 – Null Effect of Interaction		
A-A	285	359	399	.100
A-U	285	343	408	.140
T-A	285	359	399	.100
T-U	285	343	409	.140

Note. A-A = At-Ambiguous (e.g. *tossed*); A-U = At-Unambiguous (e.g. *Thrown*); T-A = Toward-Ambiguous; T-U = Toward-Unambiguous.

The second hypothetical population we simulated was one in which the interaction observed by Levy et al. exists, but in their study was 50% larger relative to the effect as it exists in the population. The purpose of this hypothetical population was to assess how much data would be needed to detect an effect assuming that Levy et al.'s study represented an overestimate of this effect, since studies finding significant effects using small samples often report inflated effect sizes (Vasishth et al., 2018). As such, our parameters were set to produce a difference of ~ 56 ms in go-past times between the two conditions including the preposition *at*, rather than the 75 ms difference observed by Levy et al. To do so, we reduced the Regression Probability in the At-Ambiguous condition from .21 to .17, while holding the mean length of any post-regression rereading times to be the same as in Simulation 1. The means per condition for this simulation are displayed in Table 2, and the evidential categories into which the Bayes factors fell in Table 1.

The third and final hypothetical population was one in which there is no interaction between preposition form and verb ambiguity. To generate data for this population, we simply changed the parameters for the condition in which *at* was followed by an ambiguous verb to be identical to the values for the condition in which *toward* was followed by an ambiguous verb from Simulation 1, and the values for the condition in which *at* was followed by an unambiguous verb to be identical to the values for *toward* followed by an unambiguous verb. The outcomes of these simulations can be seen in Tables 1 and 2.

On the basis of these simulations we planned to initially acquire data from 72 participants. To start with the simulation of a null effect, our simulations indicate that we should very rarely (i.e. 0.4% of the time) observe misleading evidence in favour of H_1 with this dataset, and that we are only likely to obtain uninformative data on 5.9% of iterations of the experiment with this sample size. In all other cases, we would expect to correctly observe evidence in favour of a null effect. If, on the other hand, we assume there is an interaction between preposition form and verb ambiguity in the wider population, then even if this effect is smaller than that observed by Levy et al., we should only observe misleading evidence on 3.5% of iterations of our study, uninformative evidence in 14.9% of iterations, and evidence correctly in favour of an effect in all remaining iterations. If Levy et al.'s study accurately reflects the size of the effect in the population, we would obtain misleading evidence on a mere 0.6% of iterations. Thus, 72 participants with 56 items was an appropriate initial sample size for minimizing the probability of observing misleading evidence in our study.

Proposed analysis steps

Prior to analysing data, we will determine if any participant's data needs to be excluded, and replaced. The only planned criteria used for this will be to assess participants' performance on simple comprehension questions that appear following our filler items. If a participant achieves <75% accuracy on these questions, we will not use their data. It should be noted that we will not use comprehension accuracy on our experimental items to determine performance, as the comprehension questions are designed to be hard to answer. Once we have obtained usable data from 72 participants, we will begin formal analysis. The first step in this process will be to use SR-Research DataViewer to perform standard cleaning operations on our data, with this including the removal of fixations lasting longer than 800 ms, the merging of fixations below 80 ms with fixations less than 0.5 degrees away and fixations below 40 ms with fixations less than 1.25 degrees away, and finally the removal of any remaining fixations below 80 ms. These values represent DataViewer's default settings for fixation cleaning for reading studies. We will then use DataViewer to produce interest area reports and fixation reports. Using these reports, we will then clean our data for instances in which participants blinked during a fixation on our target word, and any values for each measure that are over 4 standard deviations above the mean for that measure. At this point, our data are ready for formal analysis.

As mentioned above, there are two main aspects of our data analysis, with the first of these being to calculate a Bayes factor to assess whether we have evidence for or against the critical interaction observed by Levy et al. in go-past times on the critical target verb. These Bayes factors are estimated in the same way as in the simulations for our Bayes Factor Design Analysis, detailed above. While the key measure from Levy et al.'s study to replicate was the effect in go-past time, we also calculate Bayes factors to assess the presence of the interaction in first fixation duration, gaze duration, and total viewing times on this target word. Levy et al. also examined these measures in their study, and, while they did not find significant interactions, it may be that small effects in these measures emerge with our larger sample size. The generated Bayes factor will be used to state whether an effect definitively exists or does not exist in each measure. If the Bayes factor for go-past time is uninformative with 72 participants we will continue testing participants in multiples of four (to respect counterbalancing) and adding their data to our analysis until the evidence from the Bayes factor is able to support either H_1 or $H_{0.}$

As well as calculating Bayes factors to determine which hypothesis our data supports, we also construct Bayesian mixed models using the brms package (Bürkner, 2020) in R in order to estimate the size of any effects that we observe, in addition to 95% credible intervals around these estimates and the probability of the effect being greater than 0 ms, 10 ms, 20 ms, 30 ms, 40 ms, and 50 ms. We construct models for go-past

Table 3
Conditional Means (and Standard Errors) for Each of our Dependent Variables in our Overall Dataset.

	At		Toward	
	Ambiguous	Unambiguous	Ambiguous	Unambiguous
First Fixation Duration	274 (4)	270 (4)	286 (4)	272 (4)
Gaze Duration	338 (5)	316 (5)	352 (6)	328 (6)
Go-Past Time	457 (12)	410 (10)	465 (13)	441 (12)
Total Time	880 (20)	811 (19)	893 (21)	796 (20)
Regression Out	0.20 (0.01)	0.17 (0.01)	0.18 (0.01)	0.19 (0.01)
Regressions In	0.32 (0.01)	0.27 (0.01)	0.49 (0.02)	0.42 (0.02)
Comprehension (All)	0.80 (0.01)	0.82 (0.01)	0.79 (0.01)	0.82 (0.01)
Comprehension (RRV)	0.78 (0.02)	0.80 (0.01)	0.76 (0.02)	0.80 (0.01)

time in addition to first fixation duration, gaze duration, total viewing time, regression probability out of the verb, regression probability into the preposition, and comprehension performance for both the full set of items and then only items with questions querying interpretation of the relative clause. The final three measures are included here, but not in the Bayes factor analysis, due to the fact that to the best of our knowledge there is no method for constructing logistic mixed models built into the BayesFactor package, while there is in the brms package. For the reading time measures, we input untransformed reading times as dependent variables, setting the family function of the model to lognormal. For the three models using binomial variables, we set the model distribution to Bernoulli. Each model includes fixed main effects for preposition form and verb type, as well as the interaction between these two factors. We also included by-participant and by-item random intercepts and slopes for each fixed effect in our model. Priors for all models will be weakly informative, with priors of *Normal*($\mu = 0, \sigma = 10$) for the model intercept and *Normal*(0, 1) for each fixed effect and standard deviation parameter, and a regularisation of 2 on the covariance matrix of random effects. Each model was run with four chains of 5000 iterations each, with 1000 iterations being treated as warmup. If any model returned parameters with an Rhat above 1, further iterations were added.

Finally, we perform several further analyses, suggested by reviewers of an earlier version of the current manuscript. Specifically, we analyse our data for go-past time, regressions out of the target region, and comprehension performance using more typical frequentist linear mixed models, in order to ensure that any divergence from the findings of Levy et al. is not due to the use of Bayesian statistics. It should be noted that this is still not a direct replication of their analysis, due to the fact that separate by-subject and by-item ANOVAs were used to analyse their data. However, since the publication of Levy et al.'s paper, linear-mixed models have become strongly established as the standard technique in psycholinguistic research, and so we chose to adopt this method.

We also examine the data for just the items that were used in Levy et al.'s study, in order to ascertain that any differences between our findings and those of Levy et al. are not simply due to our new items differing from those of Levy et al. in some unforeseen manner. If this was the case, we might expect to see no critical interaction in the overall analysis, while this effect should still be present in just the items used by Levy et al.

A final issue we examine, also looked at by Levy et al., is whether readers show greater evidence of word-level uncertainty in trials in which they skip rather than fixate the preposition during first-pass reading. In such instances, readers will only have sampled perceptual information from the preposition in lower acuity areas of the visual field, and thus may maintain greater uncertainty. Levy et al. did actually examine this issue themselves, finding little effect of whether the preposition was initially skipped. However, it could be the case that such an effect would come across more clearly in our own larger data set.

Results

Before analysing our data, we determined if any participants' data needed to be excluded. Five additional participants beyond the 72 used in analysis were tested, but their data was not included. The decision to exclude three of these was made during the testing session due to poor calibration and at times an inability to track the eye, with these participants failing to complete all experimental trials. Another was excluded due to comprehension below 75% on our filler items, a criteria determined in our Stage 1 report. The final excluded participant was tested prior to statistical analysis of our sample of 72 participants confirming that we had reached our stopping criteria (i.e. a decisive Bayes factor for the interaction in go-past time).

Before formal data analysis we used SR-Research Data Viewer to clean our data. The merging operation specified above affected 0.95% of fixations across the whole experimental session. Within Data Viewer interest areas were setup to consist of each relevant word within the sentence alongside the space preceding it. Blinks and extreme reading times were removed, as specified above. The reports output by Data-Viewer and all R Scripts used in data analysis can be found on the OSF.

Main analysis

We begin the presentation of our results with what we consider to be the key findings to replicate from Levy et al. (2009), consisting of an interaction between preposition form and verb ambiguity in go-past time, the probability of making a regression from the verb in first-pass reading, and comprehension rates. These are the three measures in which Levy et al. observed the theoretically important interaction between verb ambiguity and preposition form which could be considered evidence for word level uncertainty. We also examine a number of other measures of reading times. The conditional means and standard errors of all measures are presented in Table 3. Fig. 1 shows the probability of the interaction term in our Bayesian regression model being greater than or equal to a range of values for each measure. These values were converted back to a millisecond or probabilistic scale for interpretability. Plots of the posterior distribution for the interaction effect in log units with 95% credible intervals can be found in the Appendix as Fig. A1. For reading time models, in addition to reporting a Bayes factor for the interactive effect (as planned in our Stage 1 report), we also report Bayes factors for the main effects of ambiguity and preposition form, for the sake of completeness.

In go-past time (Intercept $b = 5.90$), the model constructed using the brms library revealed a main effect of verb ambiguity ($b = -0.07$, CrI [-0.13,-0.01], P($\widehat{b} < 0$) = 0.99) but little evidence for a main effect of preposition form ($b = 0.03$, CrI[-0.01,0.07], P($\widehat{b} > 0$) = 0.93) and most importantly little evidence for an interaction between preposition form and verb ambiguity ($b = 0.04$, CrI[-0.04, 0.12], P($\widehat{b} > 0$) = 0.83). Fig. 1 shows the probability of the interaction between preposition form and verb ambiguity being greater than a range of values. As can be seen, if there is any interaction at all it was most likely a very small effect indeed. In addition, we also calculated a Bayes factor in order to assess the level of evidence for/against an interactive effect within our model. The Bayes factor offered strong evidence for the null hypothesis ($BF_{10} = 0.097$). The Bayes factor analysis also offered very strong evidence for an effect of verb ambiguity ($BF_{10} = 45$), and evidence against an effect of preposition form ($BF_{10} = 0.154$). Finally, we also ran a frequentist linear-mixed model, in order to match the frequentist approach used by Levy et al. (2009). This model returned a significant effect of verb ambiguity ($b = -0.07$, SE = 0.03, t = -2.39) but non-significant effects of both preposition form ($b = 0.03$, SE = 0.02, t = 1.67) and the interaction between verb ambiguity and preposition form ($b = 0.04$, SE = 0.04, t = 1.09).

Turning now to regressions out of the target verb (Intercept $b = -1.67$), we found no evidence for a main effect of verb ambiguity ($b =$

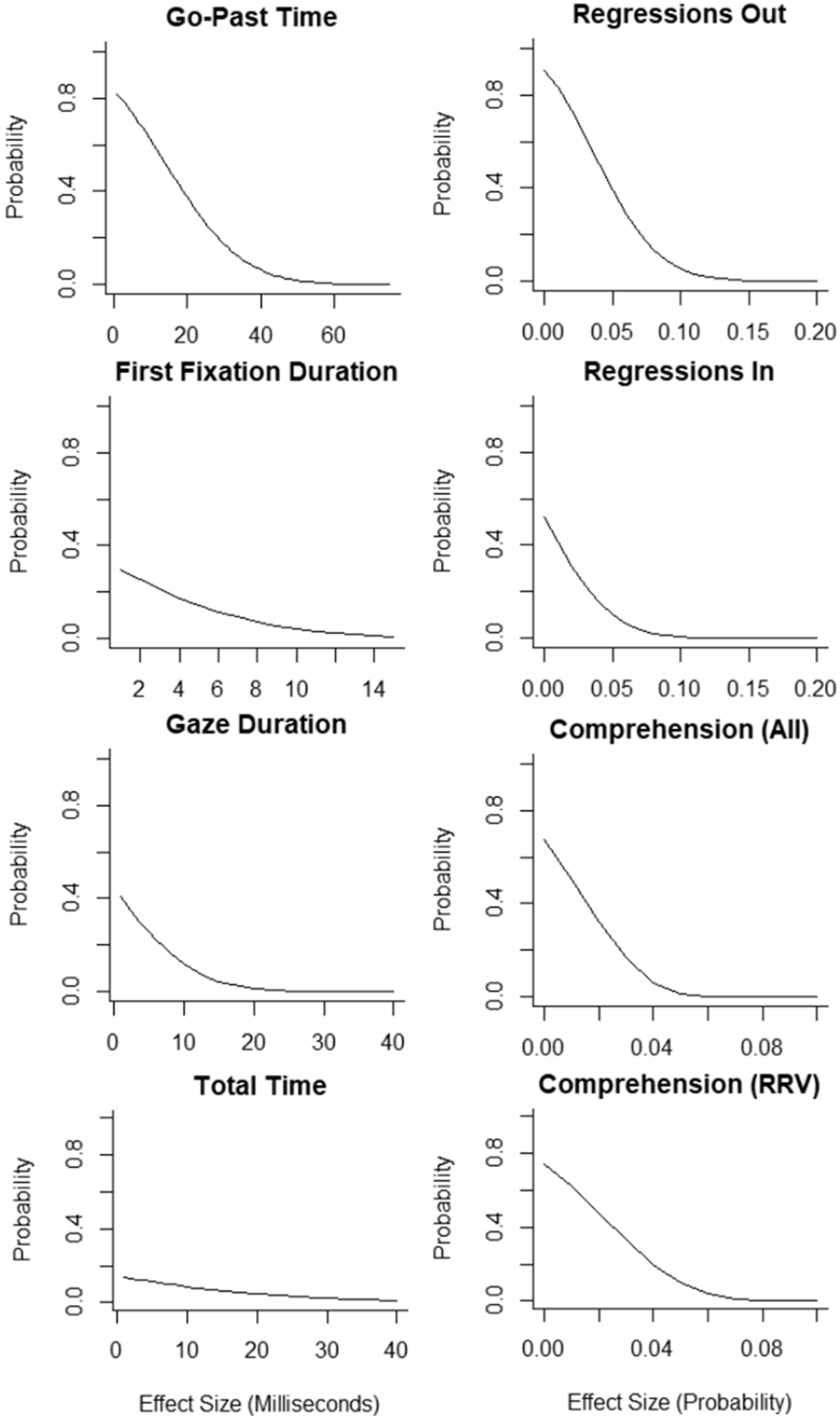

Fig. 1. The probability of the interaction effect being above certain values in each measure.

-0.07, CrI[-0.31, 0.16], P($\widehat{b} > 0$) = 0.74), no evidence for an effect of preposition form ($b = -0.04$, CrI[-0.26, 0.18], P($\widehat{b} > 0$) =.38), and no evidence for an interaction between preposition form and verb ambiguity ($b = 0.28$, CrI[-0.14, 0.70], P($\widehat{b} > 0$) =.90). A frequentist logistic mixed-effects model revealed no significant effect of verb ambiguity ($b = -0.09$, SE = 0.11, z = -0.86), preposition form ($b = -0.02$, SE = 0.10, z = -0.20), or interaction between these two variables ($b = 0.26$, SE = 0.19, z = 1.40).

In comprehension performance across all items (Intercept $b = 1.87$) we observed no effect of verb ambiguity ($b = 0.16$, CrI[-0.08, 0.39], P($\widehat{b} > 0$) = 0.92), no effect of preposition form ($b = -0.06$, CrI[-0.27, 0.15], P($\widehat{b} > 0$) = 0.72), and no interaction between verb ambiguity and preposition form ($b = 0.09$, CrI[-0.32, 0.49], P($\widehat{b} > 0$) = 0.67). A frequentist model showed a significant effect of verb ambiguity ($b = 0.21$, SE = 0.10, z = 2.17), but no significant effect of preposition form ($b = -0.05$, SE = 0.10, z = -0.53) or interaction ($b = 0.10$, SE = 0.18, z = 0.56). A similar pattern of results was observed when we analyzed comprehension performance on only the items which were followed by a comprehension question which specifically probed understanding of the reduced relative clause (Intercept $b = 1.68$), with no effect of verb ambiguity ($b = 0.17$, CrI[-0.10, 0.42], P($\widehat{b} > 0$) = 0.90), no effect of preposition form ($b = -0.02$, CrI[-0.24, 0.20], P($\widehat{b} > 0$) = 0.41), and no interaction between verb ambiguity and preposition form ($b = 0.15$, CrI [-0.32, 0.61], P($\widehat{b} > 0$) = 0.74). A frequentist model revealed a significant effect of verb ambiguity ($b = 0.23$, SE = 0.10, z = 2.42), but no significant effects of preposition form ($b = -0.03$, SE = 0.10, z = -0.329) or interaction ($b = 0.16$, SE = 0.19, z = 0.847).

Turning briefly away from the target verb, it is also worth considering whether readers were more likely to make a regression into the preposition in the critical condition. Our statistical model (Intercept $b = -0.61$) here suggests this was not the case, with a main effect of ambiguity ($b = -0.29$, CrI[-0.44, -0.14], P($\widehat{b} < 0$) = 1), an effect of preposition form ($b = 0.78$, CrI[0.59, 0.99], P($\widehat{b} > 0$) = 1), and no interaction ($b = 0.01$, CrI[-0.30, 0.33], P($\widehat{b} > 0$) = 0.52).

We will now turn to the remaining measures examined by Levy et al. (2009). It is worth reiterating that in these measures no significant interaction between verb ambiguity and preposition form was observed in the original study. Thus, evidence against these effects in the current

Table 4
Separate Conditional Means (and Standard Errors) for Items used by Levy et al. (2009) and our New Items.

	A-A	A-U	T-A	T-U
		Levy et al. Items		
First Fixation Duration	275 (5)	267 (6)	276 (6)	268 (5)
Gaze Duration	337 (8)	311 (8)	353 (9)	311 (7)
Go-Past Time	469 (19)	397 (15)	484 (20)	402 (16)
Total Time	902 (30)	779 (28)	913 (31)	714 (27)
Regression Out	0.19 (0.02)	0.17 (0.02)	0.19 (0.02)	0.16 (0.02)
Regression In	0.34 (0.02)	0.29 (0.02)	0.50 (0.02)	0.43 (0.02)
Comprehension (All)	0.76 (0.02)	0.82 (0.02)	0.76 (0.02)	0.80 (0.02)
Comprehension (RRV)	0.73 (0.02)	0.78 (0.02)	0.73 (0.02)	0.78 (0.02)
		Our Items		
First Fixation Duration	273 (5)	271 (5)	295 (6)	276 (5)
Gaze Duration	338 (7)	319 (7)	351 (6)	341 (8)
Go-Past Time	448 (15)	419 (13)	449 (16)	470 (17)
Total Time	864 (27)	834 (24)	877 (27)	855 (28)
Regression Out	0.21 (0.02)	0.18 (0.02)	0.17 (0.02)	0.21 (0.02)
Regression In	0.30 (0.02)	0.25 (0.02)	0.48 (0.02)	0.42 (0.02)
Comprehension (All)	0.84 (0.02)	0.83 (0.02)	0.81 (0.02)	0.84 (0.02)
Comprehension (RRV)	0.81 (0.02)	0.81 (0.02)	0.79 (0.02)	0.81 (0.02)

Table 5
Means (and Standard Errors) for Sentence Containing *At*, Conditional on Target Ambiguity and Article Skipping.

	Fixated At		Skipped At	
	Ambiguous	Unambiguous	Ambiguous	Unambiguous
First Fixation Duration	282 (7)	281 (7)	271 (4)	265 (4)
Gaze Duration	345 (10)	326 (10)	335 (6)	311 (6)
Go-Past Time	472 (22)	447 (20)	451 (14)	393 (11)
Total Time	915 (40)	907 (39)	866 (23)	769 (20)
Regressions Out	0.22 (0.03)	0.20 (0.03)	0.20 (0.02)	0.16 (0.02)
Regressions In	0.31 (0.03)	0.26 (0.03)	0.32 (0.02)	0.27 (0.02)
Comprehension (All)	0.80 (0.03)	0.82 (0.03)	0.80 (0.02)	0.83 (0.02)
Comprehension (RRV)	0.77 (0.03)	0.78 (0.03)	0.78 (0.02)	0.81 (0.02)

Table A1
Parameter Values in our Simulations for each Hypothetical Population in each Condition.

	First Fixation Duration	First-Pass Refixation Duration	Post-Regression Re-Reading Time	Regression probability
		Simulation 1 – Levy et al.'s effects		
A-A	269	268	492	.21
A-U	264	160	550	.12
T-A	269	276	343	.10
T-U	269	216	402	.14
		Simulation 2 – Reduced Interaction		
A-A	269	268	492	.17
A-U	264	160	550	.12
T-A	269	276	565	.10
T-U	269	216	402	.14
		Simulation 3 – Null Effect of Interaction		
A-A	269	276	343	.10
A-U	269	216	402	.14
T-A	269	276	343	.10
T-U	269	216	402	.14

Note. A-A = At-Ambiguous (e.g. *tossed*); A-U = At-Unambiguous (e.g. *Thrown*); T-A = Toward-Ambiguous; T-U = Toward-Unambiguous. The values in the two middle columns are not the average amount reading time increased between measures, so much as the average amount by which it increased on the trials where a re-fixation/regression was made. For example, while Post-Regression Re-Reading Time is set to 492 in Row 1, this only results in a 103 ms average increase across trials, since this reading time is only added to gaze durations on.21 of trials. First-pass refixation times were added to first fixation durations to obtain gaze durations, while post-regression re-reading times were added to gaze durations to obtain go-past times.

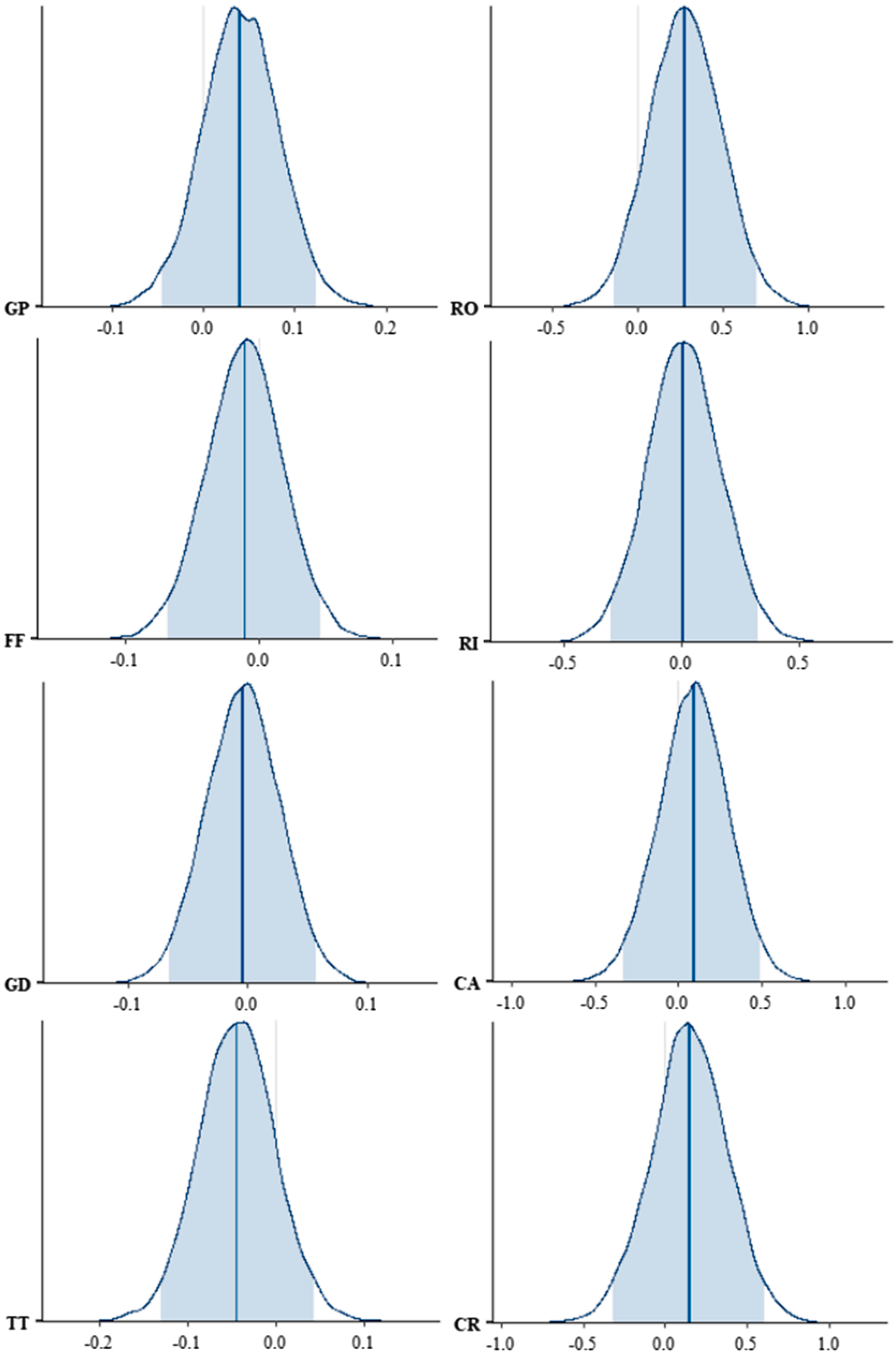

Fig. A1. Posterior distributions for the estimate of the interactive effect in each measure for our main analysis in log-units. The shaded area represents the 95% Highest Density Interval. GP = Go-Past Time; FF = First Fixation Duration; GD = Gaze Duration; TT = Total Time; RO = Regressions Out of Verb; RI = Regressions into Preposition; CA = Comprehension on All Items; CR = Comprehension on Items Probabing Relative Clause Comprehension.

paper would not technically represent a failure to replicate Levy et al. (2009), while evidence for these effects would represent a failure to replicate the exact pattern observed by Levy et al. Nonetheless, we will proceed on the assumption that any effect in these measures is actually evidence for the theoretical position advanced in Levy et al. (2009), though an absence of an effect would in no way be problematic for this theoretical position. In first fixation durations (Intercept $b = 5.55$) we found equivocal evidence for an effect of verb ambiguity ($b = -0.03$, CrI $[-0.05, 0.\ 0.00]$, P($\widehat{b} < 0$) = 0.060; $\text{BF}_{10} = 0.519$), evidence against an effect of preposition form ($b = 0.02$, CrI$[-0.00, 0.04]$, P($\widehat{b} > 0$) = 0.94; $\text{BF}_{10} = 0.154$) and evidence against an interaction between these two factors ($b = -0.01$, CrI$[-0.07, 0.05]$, P($\widehat{b} > 0$) = 0.35; $\text{BF}_{10} = 0.064$), with any trend towards an interaction actually being in the opposite direction to hypothesised (i.e. a larger ambiguity effect following *toward* than *at*). In gaze duration (Intercept $b = 5.71$) we found evidence for an effect of verb ambiguity ($b = -0.06$, CrI$[-0.10, -0.03]$, P($\widehat{b} < 0$) = 1; $\text{BF}_{10} > 1000$), equivocal evidence for an effect of preposition form ($b = 0.03$, CrI$[0.00, 0.06]$, P($\widehat{b} < 0$) = 0.98; $\text{BF}_{10} = 0.540$), and evidence against an interaction between these two variables ($b = -0.00$, CrI $[-0.06, 0.06]$, P($\widehat{b} < 0$) = 0.45; $\text{BF}_{10} = 0.056$). Finally, in total reading times (Intercept $b = 6.52$) we found evidence for an effect of verb ambiguity ($b = -0.11$, CrI$[-0.16, -0.05]$, P($\widehat{b} < 0$) = 1; $\text{BF}_{10} > 1000$), and evidence against an effect of preposition form ($b = -0.01$, CrI$[-0.05, 0.03]$, P($\widehat{b} > 0$) = 0.30; $\text{BF}_{10} = 0.046$) and evidence against an interaction ($b = -0.04$, CrI$[-0.13, 0.04]$, P($\widehat{b} > 0$) = 0.15; $\text{BF}_{10} = 0.103$).

Analysis of Levy et al.'s stimuli only

It could be argued that the analysis we have presented thus far does not count as a direct replication of Levy et al. (2009) due to the fact that in addition to the 24 items used in the original study we also presented 32 items that we designed ourselves. While we took care to make these items as similar as possible to those used by Levy et al. (2009), it is possible that subtle differences emerged between the items. This could lead to our items suppressing an effect which was present in the original 24 items. In order to test this possibility we analysed just the data for the 24 items which were identical to or adapted from Levy et al.'s items. Separate conditional means for the original Levy et al. items and our own original items are shown in Table 4. We also re-ran our statistical models for just the stimuli used by Levy et al. For brevity's sake we will report only the estimate for the interactive effect from our brms models, as well as the Bayes factor for this effect. These analyses showed that within just the 24 items from Levy et al. there was evidence against the presence of an interaction in go-past time ($b = -0.00$, CrI$[-0.14, 0.13]$, P($\widehat{b} > 0$) = 0.48, $\text{BF}_{10} = 0.085$), first fixation durations ($b = 0.03$, CrI $[-0.06, 0.11]$, P($\widehat{b} > 0$) = 0.73, $\text{BF}_{10} = 0.113$), gaze durations ($b = -0.01$, CrI$[-0.10, 0.08]$, P($\widehat{b} > 0$) = 0.41, $\text{BF}_{10} = 0.088$), total time ($b = -0.08$, CrI$[-0.23, 0.07]$, P($\widehat{b} > 0$) = 0.14, $\text{BF}_{10} = 0.197$), the probability of making a regression out of the critical verb ($b = 0.00$, CrI$[-0.66, 0.68]$, P($\widehat{b} > 0$) = 0.49), the probability of making a regression into the preposition ($b = -0.04$, CrI$[-0.51, 0.44]$, P($\widehat{b} > 0$) = 0.44) or comprehension performance for all items ($b = -0.09$, CrI$[-0.73, 0.58]$, P($\widehat{b} < 0$) = 0.61) and just those probing relative clause understanding ($b = 0.06$, CrI$[-0.67, 0.79]$, P($\widehat{b} < 0$) = 0.56).

Skipping contingent analysis

It could be argued that if participants do maintain uncertainty about the identity of a word, then this uncertainty should be greater when they have skipped over this word. In our Stage 1 Report, we committed to testing this possibility in an additional analysis. To begin with, it is worth noting that skipping of the preposition was unsurprisingly considerably higher when *at* was used ($m = 0.71$) as opposed to *toward* ($m = 0.12$). An ideal analysis based on skipping would test a) whether ambiguity effects were greater when *at* was skipped rather than fixated and b) whether ambiguity effects were equivalent regardless of whether or not *toward* was skipped. Unfortunately, the low skipping of *toward* somewhat limits our ability to conduct a full and proper analysis here. With only 12% of the trials in which the preposition *toward* was used involving skipping of the preposition, any estimates of the effect of skipping in these sentences will be based on very little data and highly noisy. As such, for the following analysis we focused exclusively on the sentences featuring *at*. We acknowledge this as a post-hoc analysis decision, rather than one made as part of our original plans. It is worth noting there was still an imbalance for the sentences in which *at* was used (i.e. 29% of trials in which the preposition was fixated vs. 71% in which it was skipped).

In Table 5 we present mean reading times per condition depending upon whether *at* was fixated or skipped. We constructed Bayesian mixed-models in which verb ambiguity and whether or not the preposition was skipped – as well as the interaction between these two factors – were treated as predictor variables. For brevity's sake, we report only the estimate and Bayes factors for the interactive effect. There was no interaction between verb ambiguity and preposition skipping in any measure. Specifically, in go-past time (Intercept $b = 5.90$) there was evidence against an interaction between ambiguity and preposition skipping ($b = -0.02$, CrI$[-0.16, 0.11]$, P($\widehat{b} > 0$) = 0.36, $\text{BF}_{10} = 0.092$). In the probability of participants making a regression out of the target verb (Intecept $b = -1.62$) there was no evidence of an interaction ($b = -0.07$, CrI$[-0.70, 0.55]$, P($\widehat{b} > 0$) = 0.41). In the probability of participants regressing back into *at* (Intercept $b = -1.05$) there was no evidence for an interaction between ambiguity and preposition skipping ($b = -0.02$, CrI$[-0.52, 0.48]$, P($\widehat{b} > 0$) = 0.47). In overall comprehension rates (Intercept $b = 1.92$) there was no evidence for an interaction ($b = 0.00$, CrI$[-0.69, 0.66]$, P($\widehat{b} > 0$) = 0.51). In comprehension for just items probing reduced relative clause understanding (Intercept $b = 1.67$) there was no evidence for an interaction ($b = 0.16$, CrI$[-0.51, 0.84]$, P($\widehat{b} > 0$) = 0.71). In first fixation durations (Intercept $b = 5.55$) there was evidence against an interaction ($b = -0.00$, CrI$[-0.08, 0.07]$, P($\widehat{b} > 0$) = 0.45, $\text{BF}_{10} = 0.089$). In gaze durations (Intercept $b = 5.70$) there was evidence against an interaction between ambiguity and preposition skipping ($b = 0.01$, CrI$[-0.09, 0.10]$, P($\widehat{b} > 0$) = 0.56, $\text{BF}_{10} = 0.088$). In total time (Intercept $b = 6.54$) there was evidence against an interaction between ambiguity and preposition skipping ($b = -0.03$, CrI$[-0.16, 0.10]$, P($\widehat{b} > 0$) = 0.33, $\text{BF}_{10} = 0.095$).

Filler analysis

As outlined in the methods section, among our stimuli we included 40 filler items in which a target word was manipulated to be either high- or low-frequency. These items were included so that we could ensure that our participants exhibited a psycholinguistic effect that is undoubtedly replicable. We cleaned the data for these items in the same way as for the main analysis presented above, and examined first fixation and gaze durations on the target word using the same statistical methods as above. This analysis showed that our participants did indeed show evidence of frequency effects in first-fixation durations ($b = 0.11$, CrI$[0.08, 0.14]$, P($\widehat{b} > 0$) = 1, $\text{BF}_{10} > 1000$; HF mean = 235, LF mean = 264) and gaze durations ($b = 0.19$, CrI$[0.14, 0.23]$, P($\widehat{b} > 0$) = 1, $\text{BF}_{10} > 1000$; HF mean = 257, LF mean = 312). This effect, alongside the main effect of ambiguity observed in several measures in our main experiment allows us to be confident that our participants were indeed exhibiting the type of measurable reading behaviour we would expect in an eye-tracking study.

Discussion

In the current study we attempted to replicate the findings of Levy et al. (2009) with a greater number of both participants and experimental stimuli. To recapitulate, Levy et al. found that readers experienced more processing difficulty at an ambiguous reduced relative verb (e.g. *tossed*) compared to an unambiguous reduced relative verb (e.g. *thrown*) when these verbs were preceded by a perceptually ambiguous preposition (i.e. *at*) as opposed to a perceptually unambiguous preposition (i.e. *toward*). This interactive effect primarily manifested in go-past times at the verb, regressions out of the verb, and comprehension performance. It was thus these three measures that we view as the ones in which we should observe the same effects as Levy et al. in order to consider the current paper to have successfully replicated their work. Overall, our data suggests that the findings of Levy et al. are not replicable, with our statistical analysis offering evidence against the existence of an interaction between verb ambiguity and preposition form.

In go-past time there was a slight trend toward the interaction observed by Levy et al. in the conditional means. However, our statistical analysis suggested our data were actually more in line with the conclusion of there being no interaction between the two independent variables, with a Bayes factor offering strong evidence for the null hypothesis. In comprehension rates there was no evidence for an interaction, with any trend in the means towards such an effect actually being in the opposite direction to that observed by Levy et al. In regressions out of the target region there was again a slight trend in the means that was consistent with Levy et al.'s findings, but statistical analyses offered little evidence that this effect was reliable. There was also no evidence for the key interaction in any other measures of eye-movement behaviour during reading.

A series of more specific analyses also failed to offer any support to the replicability of Levy et al.'s original finding, and the more general idea that readers maintain word-level uncertainty during reading. In one sub-analysis we examined reading behaviour exclusively on the items used by Levy et al. (2009). Here, we found that in both go-past time and regressions out of the target verb any numerical trends were in fact in the opposite direction to those observed in the original study. We also examined whether readers might be more likely to maintain uncertainty about the word *at* when they skipped over it, as opposed to directly fixating it. Again, statistical evidence was against the existence of any such effect.

Based on the present data set we can say with a high degree of confidence that the effects first observed by Levy et al. (2009) do not replicate, or at the very least that if an effect does exist in the wider population then it is much smaller than the effect first observed by Levy et al. (2009), and much smaller than the reduced effect on which we based our Bayes factor design analysis and sampling plans. Our study is not the only recent investigation of whether readers target regressions to sources of uncertainty earlier in the sentence. Paape et al. (2022) had participants read German sentences similar to our own, in which later material might be expected to cause participants to make targeted regressions in order to confirm earlier material. Specifically, in their study the later material may have led participants to 1) doubt the presence of the article *dass* earlier in the sentence (i.e. the German equivalent of *that*); 2) consider that they may have missed a colon, and; 3) consider that an earlier noun may have been capitalised. The pattern of regressions that readers made in this study only fulfilled one of three predictions made for the noisy-channel account, with this prediction also being consistent with other accounts of sentence processing. The findings from this study are largely consistent with our own, in that the effects offered relatively little support to the idea that people act on uncertainty about prior input. However, our own study provides considerably stronger evidence against this possibility, since within Levy's (2008) noisy-channel framework the manipulation used in the current study would predict a considerably larger effect than that used by Paape et al. Furthermore, the manipulation used here and by Levy et al. (2009) remains the only one to have examined uncertainty about the identity of a specific lexical item, rather than the presence/absence of words earlier in the sentence. It is also worth re-iterating that we were previously unable to replicate Levy et al.'s findings, albeit using self-paced reading methodology as opposed to eye-tracking during reading (Cutter et al., 2022).

It is briefly worth considering if there was anything specific about the Levy et al. (2009) study which may have resulted in the erroneous detection of an effect, or if this effect was simply a false positive driven by a low sample size. While we believe the latter to be the case, we will touch upon some factors that could have resulted in a difference. One factor worth focusing on is the possibility of some fundamental difference between the participants tested in the current investigation compared to those tested by Levy et al. Recall that our original interest in the phenomena under study was in terms of whether older adults might be more likely to maintain uncertainty about word identity than younger adults (Cutter et al., 2022). As such, it is worth considering if there could have been some differences between participant groups across studies. There are actually some hints in the data that this could be the case. For example, comprehension amongst our participants was higher (78.5%) than amongst Levy et al.'s participants (68.5%). In addition, there was an unusually low skipping rate in Levy et al.'s participants (e.g. 0.5% on *smiled*; 2% on *toward* vs. 12% in the current study; 58.5% on *at* vs. 71% in the current study; 4.5% on *tossed/thrown*). The low comprehension combined with very low skipping rates might suggest that the participants tested by Levy et al. were in some way fairly poor readers, and it could be the case that these poorer readers might maintain uncertainty about what they have read in a way that more skilful readers do not, due perhaps to a factor such as low lexical quality (see Perfetti, 2007). While we are of the opinion that it is simply the case that readers do not generally maintain uncertainty about word identity during reading, further work may consider investigating whether reading skill may play a role. However, the clear indication from the present findings is that maintaining uncertainty about what has been read is not associatated with skilled reading. Consequently, while it may prove interesting to obtain pattern of results consistent with those reported by Levy et al. in a sample assessed to be poorer readers, this would not provide evidence for a general theory of reading. It is also worth reiterating that we also observed evidence against the interaction effect when we restricted our analysis to just the items used by Levy et al.; as such, we can be confident that there was nothing about these stimuli which drove the effect first observed by Levy et al. but not in the current study.

At a more general level, it is important to be clear what theoretical positions our results are and are not problematic for. The present replication was aimed specifically toward testing the key evidence for the possibility that readers maintain word-level uncertainty during reading, which is one aspect of the larger noisy-channel framework of language processing (Gibson et al., 2013; Levy, 2008). While we consider our results to be problematic for the idea of readers maintaining word-level uncertainty, they are not necessarily problematic for other aspects of the noisy-channel processing framework. There remains a large amount of independent evidence that readers will often infer an alternative meaning or form of an implausible sentence when the addition/deletion of words allows for a more sensible interpretation (Gibson et al., 2013; 2017; Ryskin et al. 2018; Warren et al., 2017). However, it is worth noting that prior work on such noisy-channel inferences has generally been based purely upon participants' final interpretations of these sentences, with little attention being given to the online processing of such sentences. It may be that tracking readers' eye movements as they process these sentences would grant us further insights into the underlying processes determining whether or not such sentences are interpreted correctly, and the extent to which levels of perceptual input extracted from the sentence governs the interpretative process. Such work could also reveal the time course of any revisions that are made to potentially noisy input, and whether such revisions are made rapidly during initial processing, or only later as a post-perceptual process (see Huang & Staub, 2021). Furthermore, such work would demonstrate whether readers are more likely to act upon uncertainty over prior input when the veridical representation of the sentence is implausible as is the case in the sentences used by Gibson et al., as opposed to the veridical representation simply being syntactically unlikely as was the case in the sentences used in the current investigation.

We opened this article with a consideration of some of the ways in which Levy's (2008) proposal of word-level uncertainty may be problematic for various theories focused upon foundational aspects of the reading process, including models of visual word identification and syntactic parsing. As outlined in depth above, word-level uncertainty seems incompatible with models of word identification built upon an interactive-activation framework (McClelland & Rumelhart, 1981). As such, the current work is particularly important for the continued feasibility of the vast array of models that assume the definitive identification of words during reading (e.g. Coltheart et al., 2001; Davis, 2010; Grainger & Jacobs, 1996; Perry et al., 2007). These models typically were developed to account for the identification of words in isolation, outside of the context of sentence processing. However, recently attempts have been made to better integrate such word identification models with other aspects of the reading process, most completely in the Über-Reader model (Reichle, 2021). As work increasingly focuses upon the integration of models for different levels of the reading process it will become ever more important to consider how – or, in the case of the current study, whether – higher-level parsing operations may affect lower-level processes, such as the recognition of individual words.

In closing, we set out to replicate the findings of Levy et al. (2009). Our replication study employed a considerably larger participant sample and set of stimuli than was used in this earlier study. Using both Bayesian and frequentist analytical techniques we observed no evidence which suggests the effect first reported by Levy et al. is in fact replicable, bringing into doubt the highly influential idea that, as a consequence of noisy linguistic input, readers often maintain uncertainty about word identity during reading.

CRediT authorship contribution statement

Michael G. Cutter: Conceptualization, Methodology, Software, Formal analysis, Resources, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Ruth Filik:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Kevin B. Paterson:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

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.

Acknowledgements

This research is supported by a Leverhulme Trust Research Project Grant [RPG-2019-051].

Appendix

See Table A1.
See Fig. A1.

References

BNC Consortium. (2007). British national corpus. *Oxford Text Archive Core Collection.*

Bürkner, P. C. (2020). Brms. Version 2.14.4. https://github.com/paul-buerkner/brms.

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology, 70*, 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Cutter, M. G., Paterson, K. B., & Filik, R. (2022). No evidence of word-level uncertainty in younger and older adults in self-paced reading. *Quarterly Journal of Experimental Psychology, 75*, 1085–1093. https://doi.org/10.1177/17470218211045987

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713–758. https://doi.org/10.1037/a0019738

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*, 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science, 28*, 703–712. https://doi.org/10.1177/0956797617690277

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518–565. https://doi.org/10.1037/0033-295X.103.3.518

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Huang, A., & Staub, A. (2021). Why do readers fail to notice word transpositions, omissions, and repetitions? A review of recent evidence and theory. *Language and Linguistics Compass, 15*, Article e12434. https://doi.org/10.1111/lnc3.12434

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modelling: A practical course.* Cambridge University Press.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceeding of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234-243).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086–21090. https://doi.org/10.1073/pnas.0907664106

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*, 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Morey, R. D., & Rouder J. N. (2018). BayesFactor: Computations of Bayes factors for common designs. Version 0.9.12-4.2. <https://richarddmorey.github.io/BayesFactor/>.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357. https://doi.org/10.1037/0033-295X.113.2.327

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review, 119*, 517–545. https://doi.org/10.1037/a0028450

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. Doi: 10.1126/science.aac4716.

Paape, D., Vasishth, V., & Engbert, R. (2022). Does local coherence lead to targeted regressions and illusions of grammaticality? *Open Mind, 5*, 42–58. https://doi.org/10.1162/opmi_a_00041

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357–383. https://doi.org/10.1080/10888430701530730

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273–315. https://doi.org/10.1037/0033-295x.108.1.204

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201. https://doi.org/10.3758/BF03197692

Reichle, E. D. (2021). Computational models of reading: A handbook. *Oxford University Press*. https://doi.org/10.1093/oso/9780195370669.001.0001

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movements control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*, 445–526. https://doi.org/10.1017/s0140525x03000104

Reichle, E. D., & Schotter, E. R. (2020). A computation analysis of the constraints on parallel word identification. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.* 164-170.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*, 1–21. https://doi.org/10.3758/PBR.16.1.1

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition, 181*, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128–142. https://doi.org/10.3758/s13423-017-1230-y

Van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language, 45*, 225–258. https://doi.org/10.1006/jmla.2001.2773

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language, 94*, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Warren, T., Dickey, M. W., & Liburd, T. J. (2017). A rational inference approach to group and individual-level sentence comprehension performance in aphasia. *Cortex, 92*, 19–31. https://doi.org/10.1016/j.cortex.2017.02.015

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 205–223.