Online representations of non-canonical sentences are more than good-enough

Michael G. Cutter[1], Kevin B. Paterson[2], & Ruth Filik[1]

[1] University of Nottingham

[2] University of Leicester

Word Count: 9175

Author Note

Michael G. Cutter, Ruth Filik, School of Psychology, University of Nottingham, Nottingham, United Kingdom; Kevin B. Paterson, Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, United Kingdom.

This research was supported by the Leverhulme Trust Research Project Grant RPG-2019-051.

Correspondence regarding this article should be addressed to Michael G. Cutter, School of Psychology, University of Nottingham, University Park Campus, Nottingham, NG7 2RD. E-mail: michael.cutter@nottingham.ac.uk.

Declarations of interest: none

Abstract

Proponents of good-enough processing suggest that readers often (mis)interpret certain sentences using fast-and-frugal heuristics, such that for non-canonical sentences (e.g. *The dog was bitten by the man*) people confuse the thematic roles of the nouns. We tested this theory by examining the effect of sentence canonicality on the reading of a follow-up sentence. In a self-paced reading study 60 young and 60 older adults read an implausible sentence in either canonical (e.g. *It was the peasant that executed the king*) or non-canonical form (e.g. *It was the king that was executed by the peasant*), followed by a sentence that was implausible given a good-enough misinterpretation of the first sentence (e.g. *Afterwards, the peasant rode back to the countryside*), or a sentence that was implausible given a correct interpretation of the first sentence (e.g. *Afterwards, the king rode back to his castle*). We hypothesised that if non-canonical sentences are systematically misinterpreted then sentence canonicality would differentially affect the reading of the two different follow-up types. Our data suggested that participants derived the same interpretations for canonical and non-canonical sentences, with no modulating effect of age group. Our findings suggest that readers do not derive an incorrect interpretation of non-canonical sentences during initial parsing, consistent with theories of misinterpretation effects that instead attribute these effects to post-interpretative processes.

Keywords: Self-paced reading; good-enough processing; non-canonical sentences; sentence comprehension.

It seems uncontroversial to claim that the main goal of language comprehension is to understand a message that is being communicated, and that to succeed in this task readers must identify each word in a sentence and use a language's syntax to combine these words into a coherent and accurate representation (e.g. see Frazier & Clifton, 1996; MacDonald et al., 1994). However, over the past ~20 years much work has suggested that the mental representations that people form of linguistic input are not necessarily based on a veridical representation of that input. Rather, readers may derive representations that are shallow (see Sanford & Sturt, 2002), based upon fast-and-frugal heuristics (i.e. 'good-enough' processing; see Karimi & Ferreira, 2016, for a review) or noisy-channel inferences (see Gibson et al., 2013). It is these representations – and their effects on downstream processing of subsequent linguistic material – that the current paper focuses upon. Furthermore, we aimed to determine whether cognitive ageing affects the regularity with which readers form such good-enough representations.

The systematic misinterpretation of non-canonical sentences (Ferreira, 2003; see also Christianson et al., 2010) is one key phenomenon that resulted in the theory that readers merely derive 'good-enough' representations while processing language. Ferreira (2003) aurally presented participants with sentences that were plausible or implausible, with the arguments of the verbs within these sentences being presented in canonical or non-canonical order. Canonical here refers to sentences in which the agent of the action (i.e. verb) precedes the patient as is typical in English, as in active sentences (e.g. *The dog bit the man* as a plausible sentence; *The man bit the dog* as an implausible sentence); non-canonical order refers to sentences in which this order is reversed, as in passive sentences (e.g. *The man was bitten by the dog*; see Lim & Christianson, 2013a; Zhou & Christianson, 2016 for similar effects in subject/object relative clauses). After hearing each sentence, listeners were shown a probe asking them to report who the DO-ER or ACTED-ON was in the sentence. Participants

were most accurate for sentences presented in canonical order regardless of plausibility (~96% accuracy across three experiments), less accurate for sentences that were non-canonical and plausible (~86%), and least accurate for implausible non-canonical sentences (~72%). Inaccuracies were systematic, such that participants (mis)interpreted the dog as doing the biting, and the man as being bitten.

Based on these findings, Ferreira proposed that people's mental representations of sentences are strongly influenced by two simple heuristics, rather than formed exclusively via an algorithmic syntactic parse. The first heuristic is to treat the first noun as the subject/agent of an action, and the second noun as the object/patient (the SVO heuristic). In canonical sentences this leads to the correct mapping between a sentence's surface form and its nouns' thematic roles; this is not true for non-canonical sentences, since the patient precedes the agent. This heuristic on its own can cause some misinterpretations, explaining why non-canonical sentences are often misinterpreted regardless of plausibility. The second heuristic involves constructing meaning from simple lexical-semantic relations between words, and pragmatics about what is likely in the real world. For example, disregarding syntactic roles, a sentence including the words *bite*, *dog*, and *man* is more likely to be about a dog biting a man than a man biting a dog, given real-world knowledge about plausible events. While this heuristic is not strong enough on its own to cause misinterpretations, when combined with the SVO heuristic it results in implausible non-canonical sentences being misinterpreted more than plausible non-canonical sentences. More recently, Karimi and Ferreira (2016) further formalised this position, with the online cognitive equilibrium hypothesis. In this approach, it is proposed that to reach a state of 'cognitive equilibrium' between existing mental schemata and novel incoming material, readers process incoming linguistic material in a dual-route of heuristic and algorithmic processes. While both processes begin simultaneously, heuristic processes are faster than algorithmic processes, resulting in heuristic processes having a

disproportionate influence on a sentence's final interpretation (see also Townsend & Bever, 2001). It should be noted that the claim of good-enough processing is not that participants only ever use heuristics to analyse a sentence, with algorithmic processes not being executed at all – rather, it is merely the case that heuristic processes sometimes strongly influence the final interpretation.

More recent work has questioned whether misinterpretation effects occur due to participants' initial interpretations of sentences, or only due to post-interpretative processes driven by the probe after the sentences. Specifically, Bader and Meng (2018; see also Meng & Bader, 2021) argued that the human parsing mechanism initially arrives at a fully correct parse of non-canonical sentences, and that this representation is then accessed in a way that results in apparent misinterpretation in response to thematic role probes. Bader and Meng claimed that responding to such probes requires cue-based retrieval from the sentence representation, with the cues used in response to these specific probes including typical linear position of certain arguments in a sentence and semantic factors, with such cues being unreliable for non-canonical sentences. Here, Bader and Meng (2018) distinguished between a *parsing account* in which the initial processing of a sentence causes misinterpretation, and a *retrieval account* in which the way that people are cued to retrieve information from a fragile but correct representation causes misinterpretation.

In support of their retrieval account, Bader and Meng (2018) elicited speeded plausibility judgements for sentences that were plausible or implausible, presented in canonical or non-canonical form. Participants had to make a binary decision about whether each sentence was plausible. They argued that accuracy should be low for implausible non-canonical sentences if participants parsed them as if they were plausible canonical sentences (i.e. implausible sentences would be classified as plausible), while the remaining sentence types should be parsed and classified accurately within the 'good-enough' framework.

Instead, participants were highly accurate in rating all four sentence types, suggesting that they had been parsed correctly, without reliance on superficial heuristics.

Further evidence suggesting that misinterpretation effects may be paradigm-dependent was reported by Gibson et al. (2013), who presented participants with a range of implausible sentences – including actives (e.g. *The ball kicked the girl*) and passives (e.g. *The girl was kicked by the ball*) – with these being followed by a comprehension question to assess interpretation (e.g. *Did the girl kick something/someone?*). The good-enough framework would predict more erroneous 'yes' answers based on misinterpretations for passive versus active sentences. While such an effect was found, it was very small, with passives being misinterpreted on an extra 1.8%, 4.1%, and 2.8% of trials than actives across each of three experiments (see Gibson et al., 2017 for auditory presentation). Thus, while evidence for greater misinterpretation of non-canonical sentences is present in comprehension questions that are designed to probe a reader's understanding of the sentence, these effects are far smaller than for more specific thematic role probes.

The misinterpretation of non-canonical sentences has far-reaching consequences for theories of the human parsing mechanism, and the extent to which it always arrives at a fully algorithmic parse of a sentence. As discussed above, investigations of this phenomenon have typically relied on eliciting conscious judgements from participants, with the effect being large for thematic role probes (Ferreira, 2003), small for comprehension questions (Gibson et al., 2013; 2017), and absent for plausibility judgements (Bader & Meng, 2018; Meng & Bader, 2021). In the current paper, we take an alternative, more naturalistic approach to assessing the semantic propositions derived from implausible canonical and non-canonical sentences, by examining downstream processing consequences on a follow-up sentence. This approach has been highly informative in assessing the interpretation of garden-path sentences (see Slattery et al., 2013; Sturt, 2007), but has not been applied to non-canonical sentences.

The appeal of this approach relative to prior studies is that it allows us to probe a reader's interpretation of the sentence without asking them to think about it any more than they would during normal reading. Crucially, this allows us to obtain evidence to further distinguish between a parsing account (Ferreira, 2003) and a retrieval account (Bader & Meng, 2018); if good-enough representations are formed in the initial parsing of non-canonical sentences we should detect evidence of this in the reading of follow-up text, whereas if misinterpretations only occur in response to certain cues then there should be no evidence of good-enough processing in the reading of later text.

Specifically, within a self-paced reading paradigm, we presented participants with an implausible sentence in canonical (1a,c) or non-canonical form (1b,d), followed by a sentence that was either plausible only with a correct algorithmic reading of the first sentence (1a,b; henceforth referred to as Algorithmically Consistent) or an incorrect good-enough reading of the first sentence (1c,d; henceforth Good-Enough Consistent).[1]

1.a It was the peasant| that executed| the king.| Afterwards,| the peasant| rode back to| the countryside.

1.b It was the king| that was executed by| the peasant.| Afterwards,| the peasant| rode back to| the countryside.

1.c It was the peasant| that executed| the king.| Afterwards,| the king| rode back to| his castle.

1.d It was the king| that was executed by| the peasant.| Afterwards,| the king| rode back to| his castle.

A correct, algorithmically derived representation of the first sentence in this item would state that the king is dead, while the peasant remains alive. Conversely, the representation

[1] The | symbols present in all example sentences represent how these sentences were segmented in our self-paced reading experiment.

that would be derived using fast-and-frugal heuristics would state that the peasant is dead, while the king remains alive. In (1a) and (1c) this sentence appears in canonical form, and thus participants should almost always proceed with the algorithmically derived representation. In (1b) and (1d), however, this sentence appears in non-canonical form. As such, if readers do form good-enough representations during the initial processing of non-canonical sentences then on some portion of trials they should carry across a heuristically derived representation of this sentence into the processing of the follow-up sentence.

Which of these representations participants have in mind as they read the follow-up sentence will affect how plausible they find each follow-up sentence. In (1a,b) the follow-up sentence is Algorithmically Consistent, in that it is plausible if readers believe the peasant to have not been executed, whereas if readers believe the peasant to have been executed then this sentence becomes implausible at the word *rode*. This is because a dead person cannot typically go riding. As such, readers may have more difficulty processing this sentence when it appears after a non-canonical sentence (1b) rather than canonical sentence (1a), due to sometimes carrying a heuristically derived representation of the non-canonical sentence forwards. In contrast, in (1c,d) the follow-up sentence is Good-Enough Consistent, in that it is plausible given the heuristically derived representation of the first sentence in which the king has not been executed, but is implausible given the algorithmically derived representation of the first sentence in which the king was executed. Thus, if readers truly parse non-canonical sentences using fast-and-frugal heuristics (Ferreira, 2003) then we should find that the Algorithmically Consistent follow-up will on average be read more quickly after a canonical than non-canonical sentence, while the opposite should be true for Good-Enough Consistent follow-ups. These effects should appear as an interaction between follow-up type and first sentence canonicality.

If, on the other hand, the misinterpretation effects observed in prior studies occur only due to post-interpretative processes– as suggested by Bader and Meng (2018) –then first sentence canonicality should not affect the reading of the follow-up sentence. Rather, there should merely be a main effect of follow-up type, whereby participants read the Algorithmically Consistent follow-up more quickly than the Good-Enough Consistent follow-up, since the former sentence type will always be more consistent with the representation that readers carry forward from the first sentence. It should be noted that it is not necessarily the case that participants are not performing retrieval from a fragile memory representation of the sentence in this account. Rather, it could simply be the case that the cues to retrieval in an artificial task such as that used by Ferreira (2003) are likely to result in the retrieval of the wrong information, while the cues used in natural sentence processing leads to more accurate retrieval.

Language processing and ageing

As well as examining whether good-enough representations affect the processing of later text, we tested whether age differences exist for this process. As people age there are many alterations in their cognitive processing capacities that may affect language processing, and potentially the extent to which good-enough processing occurs (Christianson et al., 2006; Malyutina & den Ouden, 2016; Stine-Morrow et al., 2006). Specifically, as people age they generally experience declines in working memory capacity (Anders et al., 1972; Waters & Caplan, 2001) and inhibitory control (Hamm & Hasher, 1992; Hasher et al., 1999), but an increase in linguistic knowledge (Hartshorne & Germine, 2015). These factors may interact to alter the balance between algorithmic and heuristic processing. An increase in linguistic knowledge and experience may result in stronger biases towards processing sentences as Agent-Verb-Patient, and what constitutes a plausible agent or patient of a verb, increasing the speed and strength of heuristic processing. Decreases in working memory capacity and the

ability to manipulate information in working memory may make algorithmic processing more effortful than in young adults, decreasing the speed and strength of such processes. Finally, decreases in inhibitory control may affect the ability to suppress or erase irrelevant inferences from memory. Consequently, even if a reader completes the algorithmic processing of a sentence having formed a heuristic representation, full deletion of the erroneous heuristic parse is less likely.

If the above is true, older adults may be more likely to derive incorrect representations of the non-canonical sentences in (1b,d), with this resulting in larger canonicality effects on the processing of subsequent text. To test this, we collected data from 60 young adults (aged 18-25) and 60 older adults (aged 65+). We hypothesised that any two-way interaction between first-sentence canonicality and follow-up sentence type would be larger in older than younger adults, if it was present in the population at all.

Summary

In sum, we test whether readers form good-enough representations of non-canonical sentences in a manner consistent with a parsing account of misinterpretation effects. To do so, we examine self-paced reading times on sentences following an implausible sentence presented in canonical or non-canonical order, with follow-up sentences being either Algorithmically Consistent or Good-Enough Consistent. We will assess interpretation of the first sentence by examining follow-up sentence reading, with the assumption that an erroneously formed good-enough representation of a non-canonical sentence would differentially affect the plausibility of the two different follow-up sentence types. The processing of Algorithmically Consistent follow-up sentences should become more difficult with a good-enough representation of the first sentence, while the processing of a Good-Enough Consistent follow-up sentence should become easier. Thus, if people form good-

enough representations of non-canonical sentences, there should be an interaction between first-sentence canonicality and follow-up sentence type, with effects emerging in the penultimate and final sentence regions. Furthermore, we predicted that any evidence of good-enough processing may be larger in older compared to young adults. For added control, we also present participants with plausible sentences in canonical or non-canonical order to allow us to examine spill-over effects driven by syntactic processing difficulty, and sentences including an implausibility which was not dependent upon earlier good-enough processing in order to ensure that our participants do measurably react to violations of plausibility. These stimuli are detailed below, in the Materials and Design section.

Method

Participants

Sixty older adults (mean age = 69.5; age range = 65-88; 34 male; mean years of education = 14.54; mean hours reading per week = 20.35) and 60 younger adults (mean age = 22.4; age range = 21-25; 13 male; mean years of education = 14.85; mean hours reading per week = 12.57) participated. All older and 52 young adults were recruited via Prolific academic, completing the study for payment. Eight young adults were recruited from the University of Nottingham for course credit. All participants were native English speakers.

Materials and Design

For the current study we developed sentences similar to those used by Ferreira (2003), which could be presented canonically or non-canonically, and in which the arguments of the verb were plausible in one order, but not in the reverse order. Our stimuli were modelled after what Ferreira (2003) called *biased reversible* sentences (e.g. *the dog bit the man*) rather than *irreversible sentences* (e.g. *the chef cleaned the pan*). To confirm our intuitions about sentence plausibility we performed a plausibility rating study, in which participants rated

sentences on a scale of 1 (entirely implausible) to 7 (entirely plausible).[2] We created 59 biased reversible sentences, which were presented alongside the 24 biased reversible items, 24 irreversible items, and 24 reversible items from Ferreira (2003). Each rater only saw one version of each item, rating ~50% of items in one order and the other ~50% in reverse order. All items were presented as actives, as in prior studies (Bader & Meng, 2018; Ferreira, 2003). We obtained ratings from 16 older adults and 32 young adults. We used the ratings from the young and older adults separately to calculate a median score for each item in each condition.

Table 1

Mean Rating Score for the Items Used in our Experiment, with the Rating of the Weakest Item in Brackets.

	Younger adults	Older adults
Plausible first sentence	6.82 (4.5)	6.75 (4.5)
Implausible first sentence	1.74 (3.5)	1.69 (3.5)
First-sentence difference	5.08 (3)	5.06 (3.5)
Algorithmically Consistent difference	5.44 (3.5)	5.68 (3.5)
Good-Enough Consistent difference	5.70 (4)	5.83 (3.5)

Note. A score of 7 represents a perfectly plausible Sentence, while 1 represents a completely implausible sentence. Algorithmically Consistent difference was calculated by subtracting the rating of the Algorithmically Consistent follow-up sentence (e.g. *Afterwards, the peasant rode back to…*) when it followed the plausible version of the first sentence (e.g. *The king executed the peasant*) vs. the implausible version of the first sentence (e.g. *The peasant executed the king*). Good-Enough Consistent difference was calculated by subtracting the rating of the Good-Enough Consistent follow-up sentence (e.g. *Afterwards, the king rode*

[2] The full instructions for all norming studies can be found at https://osf.io/wc2g4/, alongside ratings for individual items.

back to…) when it followed the implausible version of the first sentence vs. the plausible version of the first sentence.

In a second norming study we took the most promising items from the first stage, and created a pair of follow-up sentences for each item. One follow-up sentence was designed to be plausible given a correct reading of the first sentence (i.e. Algorithmically Consistent) but implausible given a good-enough reading of this sentence. The other follow-up sentence was designed to be Good-Enough Consistent. To ensure the follow-up sentences were as (im)plausible as intended, we presented participants with the follow-up sentence preceded by the sentence from the first stage of norming in their plausible (e.g. *the king executed the peasant*) or implausible form (e.g. *the peasant executed the king*). Participants were told to rate the follow-up sentence's plausibility under the assumption that the proposition in the first sentence was true, even if somewhat implausible. Forty younger adults and 25 older adults participated. We calculated median scores for each version of each item separately for each age group. This norming process resulted in the selection of 44 items for the experiment. Mean ratings for the stimuli for each age group are in Table 1.

In addition to stimuli designed to assess our main theoretical question, we included stimuli for two 'control' experiments in our study, to rule-out alternative explanations of effects observed in our main stimuli. The first set of control stimuli consisted of 22 sentences in which the arguments of a verb were plausible in either direction, with these sentences presented in canonical (2a) or non-canonical form (2b). Ferreira (2003) referred to such sentences as reversible. These sentences were followed by a sentence, which, while similar in structure to the follow-up sentences in our main experimental items, did not contain a plausibility manipulation.

(2a) It was the sister| that hugged| the brother.| She knew that| she would| miss him when| he went to university.

(2b) It was the brother| that was hugged by| the sister.| She knew that| she would| miss him when| he went to university.

These *canonicality control* stimuli were included to allow us to assess whether syntactic processing difficulty caused by the uncommon non-canonical sentence spilled-over onto the processing of a follow-up sentence, and whether any difficulty persisted until the equivalent of the critical region from our main experiment (i.e. *miss him when*). While such an effect is not necessarily problematic for addressing our theoretical question, knowledge of such effects will aid interpretation of data obtained for our main experimental stimuli.

The second set of control items were targeted more towards assessing whether– regardless of good-enough processing– older adults might show differential responses to plausibility violations within a sentence. Specifically, we presented participants with twenty sentences like (3a) and (3b) below, taken from Rayner et al. (2004). In (3a) a knife is a sensible thing to chop carrots with, while in (3b) an axe is not a sensible thing to chop carrots with. Prior work has shown that, due to this, readers experience processing difficulty upon encountering *carrots* in (3b) compared to (3a).

(3a) John used a knife| to chop the large| carrots for dinner| last night.

(3b) John used an axe| to chop the large| carrots for dinner| last night.

There is controversy regarding whether older adults make more, equal, or less use of context in reading compared to young adults (see Payne & Silcox, 2019), with effects varying depending on experimental paradigm and the response variable examined. As such, older adults may experience reduced or increased reading difficulty relative to young adults upon encountering implausible material in self-paced reading. This would be problematic for our

main experiment, since age differences in response to our manipulation could be due to older adults engaging in more good-enough processing of non-canonical sentences, or due to them reacting differently to plausibility violations. By testing for an interaction between age and plausibility in our *plausibility control* experiment, we can rule this explanation out.

Twelve younger and twelve older adults rated the items we used from Rayner et al. for plausibility on the same scale of 1-7 used for our main experimental stimuli. This was important for establishing that these items were no more implausible than our main experimental items. This norming study showed that the plausible items were rated as more plausible by both age groups (YA mean = 6.23; OA mean = 6.14) than the implausible items (YA mean = 3.47; OA mean = 2.17), with the difference between the two item types being less extreme than our main items.

Procedure

Our procedure was approved by the University of Nottingham's School of Psychology Ethics Committee [F1258]. We presented our sentences using Gorilla.sc, a web-based client for conducting behavioural research (see Anwyl-Irvine et al., 2020) with a level of precision appropriate for detecting standard reading time effects (see Bridges et al., 2020, for a recent timing study).

Participants first provided informed consent by ticking a series of separate checkboxes for each aspect of consent, and to confirm that they were aware of their rights. After, participants provided demographic information (age, weekly hours of reading, years of education, highest educational achievement) and performed a bot-check. Next, participants performed the self-paced reading task, implemented in Gorilla's Reading Zone feature. Self-paced reading was phrase-by-phrase and non-cumulative, such that participants first saw the initial words of a sentence, while later regions were masked. When the participant hit the

space bar, the next region was revealed and the prior region masked. Our main stimuli and the canonicality control stimuli were presented in seven regions, illustrated in examples (1a-d) and (2-a-b); Region 1 consisted of the start of the first sentence up to the first noun (e.g. *It was the* [*peasant/king*]). Region 2 included the verb and surrounding function words (e.g. *that* [*was*] *executed* [*by*]) and Region 3 the remainder of this sentence (e.g. *the* [*king*/*peasant*].). Region 4 consisted of the initial words of the following sentence (e.g. *Afterwards*,) and Region 5 the material preceding our target region (e.g. *the* [*peasant*/*king*]). It was at the first word of Region 6 we expected plausibility effects dependent upon first-sentence interpretation to first affect readers, with this region consisting of this word and two spill-over words (e.g. *rode back to*). Region 7 consisted of the rest of the sentence, in which plausibility effects may have persisted (e.g. *the countryside*/*his castle*).

For the plausibility control stimuli, we used four presentation regions. Region 1 included the sentence beginning (e.g. *John used* [*a knife*/*an axe*]), Region 2 the next few words (e.g. *to chop the large*), Region 3 the point at which the implausibility appeared plus two spill-over words (e.g. *carrots for dinner*) and Region 4 the sentence's end (e.g. *last night*). It is in Regions 3 and potentially 4 that we expected to observe plausibility effects. Yes-no comprehension questions appeared after 40% of all items, with questions not appearing after any implausible canonical/non-canonical sentences.

After the self-paced reading task, participants performed a reading span and Stroop task, measuring verbal working memory and inhibitory control, respectively. These data were collected so that, if older adults did engage in more good-enough processing than young adults, we could assess the extent this was due to reduced working memory capacity and/or reduced inhibitory control.

Results

R scripts used to analyse our data, as well as the data itself, are available at https://osf.io/wc2g4/. Before analysis, we removed extreme data values below 250ms or above 5000ms. We used Bayesian statistical methods to analyse log-transformed reading time, using two complementary methods. First, to estimate the size of any effects, we used the brms package (Bürkner, 2017) in R (R Core Team, 2020) to construct Bayesian linear mixed models. Each model included main effects for all predictor variables, all interactions between variables, and random intercepts and slopes for items and participants. For each variable one level was coded as .5 while the other level was coded as -.5 in the contrast matrix. We used weakly informative priors of *Normal*(0, 1) with a regularization of 2 on the covariance matrix of random effects. Our models were run with four chains of 6000 iterations, with 1000 iterations used as warmup. Model family was set to lognormal. From these models, we report median effects estimates (*b*), 95% credible intervals for the effect size (CrI), and the probability of an effect being larger than 0 (p(*b*>0)) for each predictor variable.

In addition, we calculated Bayes factors (see Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012) for each effect to determine whether our data a) represented evidence for an effect's existence or b) represented evidence against the effect's existence. Essentially, a Bayes factor provides a ratio of a dataset's marginal likelihood under two competing hypotheses instantiated within statistical models, allowing us to infer which model/hypothesis more likely represents the processes that generated the data. The value of a Bayes factor comparing two models represents the ratio of evidence for one model versus the other. For a Bayes factor comparing Model A to Model B (BF_{AB}) values above 1 represent evidence for Model A while values below 1 represent evidence for Model B. The precise magnitude of the ratio represents the strength of evidence for one model versus the other. In general, ratios between 3/1 and 1/3 are not considered large enough to be treated as evidence in favour of

either model, while ratios of 3/1 to 10/1 (or 1/3 to 1/10), 10/1 to 30/1, 30/1 to 100/1, and greater than 100/1 are treated as moderate, strong, very strong, and extreme evidence for one hypothesis over the other (see Jeffreys, 1961; Lee & Wagenmakers, 2013). To calculate the Bayes factors for three-way interactions we compared a model including the three-way interaction with a model only including all two-way interactions. To calculate the Bayes factors for each two-way interaction we compared models in which one two-way interaction was removed at a time with a model in which all two-way interactions were included. To calculate Bayes factors for main effects we compared a model including a main effect of the variable in question and the remaining two-way interaction with a model not including the main effect. In all models testing interactions we included random slopes for each main effect. In models testing a main effect we removed a random slope for that particular effect.

Main Experiment

We begin with the analysis of our main experiment; that is to say, the stimuli in which an implausible sentence in canonical versus non-canonical form was followed by an Algorithmically Consistent or Good-Enough Consistent sentence. Mean reading times for analysed regions, collapsed across age groups, are in Table 2, and results from the Critical and Post-Critical region are presented graphically in Figure 1. Separate mean reading times for each age group are available in the Appendix. We report inferential statistics for Regions 5, 6, and 7.[3] Our models included main effects of Age Group, Canonicality, and Follow-Up Type, and two- and three-way interactions between these variables. In addition, we also included a main effect of the number of characters in a region. This variable was included to account for the fact that Algorithmically Consistent and Good-Enough Consistent follow-ups

[3] A reviewer of the prior version of our manuscript was curious as to whether there was any hint of an interaction between canonicality and follow-up type in Regions 3 and 4. There was very little evidence of any such interaction, with any effects being inconsistent in direction and numerically small. The means for these regions can be seen in Table A1 of the Appendix.

were not always equal in length, and as such apparent main effects of sentence type could have been driven by length effects rather than plausibility.

Region 5 was the pre-critical region, and directly preceded the point at which the processing of the follow-up sentence should be affected by first sentence interpretation; analysis of this region was conducted to rule out continued syntactic processing difficulty from the non-canonical sentences. The model intercept was at 6.38 (CrI[6.32,6.44]). Analyses revealed that older adults read more slowly than younger adults (b = 0.39, CrI[0.28,0.49], p(b>0) = 1; BF_{10} > 1000), and that there was evidence against persisting effects of canonicality as a main effect (b = 0.02, CrI[-0.00,0.03], p(b>0) = 0.971; BF_{10} = 0.278) and as part of an interaction with Follow-Up (b = -0.00, CrI[-0.03,0.03], p(b>0) = 0.498; BF_{10} = 0.042), Age Group (b = -0.02, CrI[-0.05,0.01], p(b>0) = 0.144; BF_{10} = 0.082), or both Age Group and Follow-Up (b = -0.04, CrI[-0.10,0.03], p(b>0) = 0.120; BF_{10} = 0.158). There was evidence against a main effect of Follow-Up (b = -0.01, CrI[-0.03,0.01], p(b>0) = 0.134; BF_{10} = 0.064) and interaction between Age Group and Follow-Up (b = 0.02, CrI[-0.01,0.05], p(b>0) = 0.890; BF_{10} = 0.097). Longer regions took longer to read (b = 0.06, CrI[0.05,0.07], p(b>0) = 1, BF_{10} > 1000). In summary, the only effect in this region was the main effect of age.

Region 6 was the critical region, and the point at which interpretation-based processing difficulties should have emerged. In this region the model intercept was at 6.57 (CrI[6.51,6.64]), and we observed longer reading times in older adults (b = 0.38, CrI[0.26,0.51], p(b>0) = 1; BF_{10} > 1000), and evidence against this interacting with Canonicality (b = 0.03, CrI[-0.01,0.07], p(b>0) = 0.900; BF_{10} = 0.111) and Follow-Up type (b = 0.01, CrI[-0.03,0.05], p(b>0) = 0.692; BF_{10} = 0.061), and evidence for an effect of region length (b = 0.07, CrI[0.05,0.09], p(b<0) = 1; BF_{10} > 1000). We also found evidence against main effects of Follow-Up Type (b = 0.01, CrI[-0.01,0.04], p(b>0) = 0.904; BF_{10} = 0.12) and

Canonicality (*b* = -0.01, CrI[-0.03,0.01], p(*b*>0) = 0.172; BF_{10} = 0.055). Crucial to our argument in the current paper, there was evidence against an interaction between Canonicality and Follow-Up, both as a two-way interaction (*b* = 0.00, CrI[-0.04,0.04], p(*b*>0) = 0.550; BF_{10} = 0.043) and part of a three-way interaction (*b* = -0.02, CrI[-0.10,0.06], p(*b*>0) = 0.304; BF_{10} = 0.075). Furthermore, any trend towards an interaction in our data was in the opposite direction to that predicted, such that Algorithmically Consistent follow-ups were read faster following a non-canonical sentence, and Good-Enough consistent follow-ups were read faster after a canonical sentence. In summary, there was evidence for a main effect of ageing and region length in this region, with evidence against any effects relating to our canonicality and follow-up type manipulations.

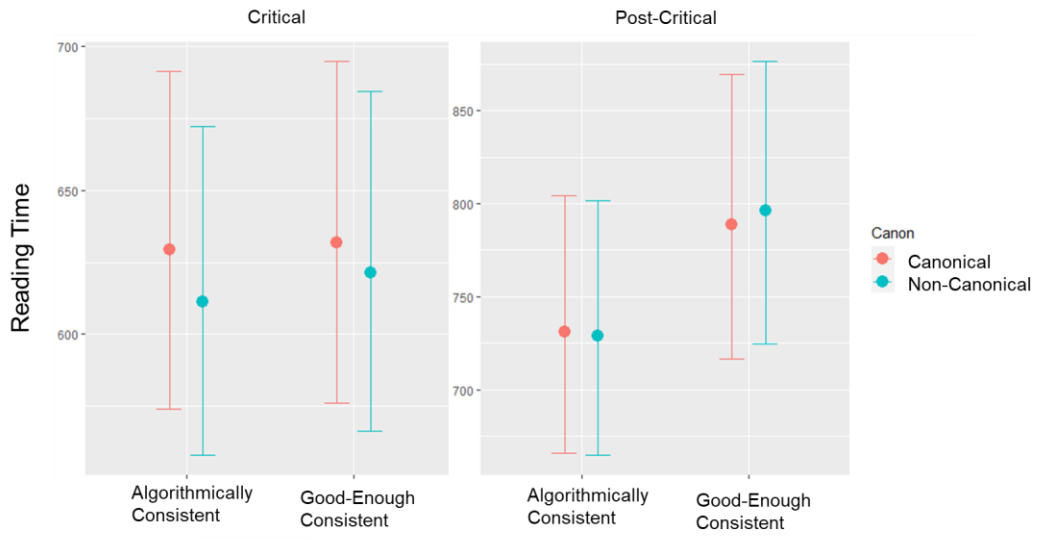

Figure 1. Estimated effects of Canonicality and Follow-Up from our Bayesian linear mixed-models for the critical and post-critical regions, with 95% credible intervals.

Region 7 was the post-critical region. In this region the model intercept was at 6.82 (CrI[6.75,6.88]), and there were main effects of age group (*b* = 0.51, CrI[0.38,0.64], p(*b*>0) = 1; BF_{10} > 1000), Follow-Up type (*b* = 0.08, CrI[0.04,0.12], p(*b*>0) = 0.999; BF_{10} > 1000), and

region length (b = 0.17, CrI[0.14,0.20], p(b>0) = 1, BF_{10} > 1000) but evidence against an effect of Canonicality (b = -0.00, CrI[-0.03,0.02], p(b>0) = 0.346; BF_{10} = 0.034). There was evidence against all interactions (Canonicality * Follow-Up b = 0.00, CrI[-0.04,0.05], p(b>0) = 0.579; BF_{10} = 0.045; Canonicality * Age b = -0.02, CrI[-0.06,0.03], p(b>0) = 0.239; BF_{10} = 0.071; Follow-Up * Age b = 0.00, CrI[-0.04,0.05], p(b>0) = 0.556; BF_{10} = 0.062; Canonicality * Follow-Up * Age b = -0.01, CrI[-0.10,0.07], p(b>0) = 0.374; BF_{10} = 0.065). In summary, there was evidence for main effects of age, follow-up type, and region length, with evidence against all other effects.

Table 2. Mean (Standard Error) Reading Times per Condition in Each Region of our Main Experimental Sentences, Collapsed by Age Group.

	Algorithmically Consistent		Good-Enough Consistent	
	Canonical	Non-Canonical	Canonical	Non-Canonical
Pre-Critical	665 (12)	675 (12)	650 (11)	665 (12)
Critical	817 (15)	799 (14)	843 (15)	834 (15)
Post-Critical	1031 (19)	1011 (18)	1141 (22)	1139 (22)

Note. See Appendix for separate reading times for each age group.

Canonicality control sentences

In the analysis of our canonicality control sentences, we were interested in testing whether readers took longer to read a follow-up sentence preceded by a non-canonical as opposed to canonical sentence. The purpose of this analysis was to determine what influence canonicality may have had on follow-up processing independently of misinterpretation effects. As such, we tested effects of canonicality in the critical and post-critical region, in addition to effects of age group and the interaction between these variables. Reading times are presented in Table 3. In the critical region (Intercept = 6.54, CrI[6.46,6.62]) there was an age group effect whereby older adults took longer to read (b = 0.39, CrI[0.27,0.52], p(b>0) = 1; BF_{10} > 1000), but no canonicality effect (b = 0.01, CrI[-0.01,0.04], p(b>0) = .837; BF_{10} = 0.080) or interaction between these factors (b = 0.04, CrI[-0.01,0.09], p(b>0) = 927; BF_{10} =

0.199). In the post-critical region (Intercept = 6.68, CrI[6.59,6.76]) there was an age group effect (b = 0.55, CrI[0.42,0.68], p(b>0) = 1; BF_{10} > 1000), but no canonicality effect (b = 0.02, CrI[-0.00,0.05], p(b>0) = 0.958; BF_{10} = 0.253) or interaction (b = 0.02, CrI[-0.04,0.08], p(b>0) = 0.744; BF_{10} = 0.095).

Table 3. Mean (Standard Error) Reading Times per Condition in our Canonicality Control Sentences.

	Younger		Older	
	Canonical	Non-Canonical	Canonical	Non-Canonical
Critical	631 (12)	624 (12)	913 (19)	956 (22)
Post-Critical	668 (15)	685 (16)	1150 (26)	1194 (28)

Simple plausibility effects

Finally, we examined the effect of our simple plausibility manipulation, in both Region 3 (critical) and Region 4 (post-critical) of these sentences (see Table 4 for mean reading times). In Region 3 (Intercept = 6.83, CrI[6.73,6.93]) there were main effects of plausibility (b = -0.06, CrI[-0.10,-0.03], p(b>0) = 0.00025; BF_{10} = 5723), and age group (b = 0.35, CrI[0.23,0.48], p(b>0) = 1; BF_{10} > 1000), but evidence against these factors interacting (b = 0.03, CrI[-0.03,0.09], p(b > 0) = 0.849; BF_{10} = 0.126). In Region 4 (Intercept = 6.76, CrI[6.68,6.84]) there was an age group effect (b = 0.56, CrI[0.44,0.68], p(b > 0) = 1; BF_{10} > 1000), but no plausibility effect (b = 0.01, CrI[-0.03,0.04], p(b>0) = 0.664; BF_{10} = 0.052), or interaction (b = -0.03, CrI[-0.10,0.04], p(b>0) = 0.191; BF_{10} = 0.119).

Table 4. Mean (Standard Error) Reading Times per Condition for our Plausibility Control Items.

	Younger		Older	
	Plausible	Implausible	Plausible	Implausible
Critical	835 (18)	919 (21)	1202 (29)	1257 (29)
Post-Critical	748 (20)	711 (15)	1274 (32)	1278 (29)

Discussion

We examined whether the misinterpretation of non-canonical sentences, first observed by Ferreira (2003) using thematic role probes, is evident in reading behaviour on a follow-up sentence. The presence of such an effect would support the idea that initial interpretations of such sentences are influenced by fast-and-frugal heuristics rather than derived purely from a detailed syntactic analysis. In contrast, the absence of an effect would support accounts in which the initial parse is algorithmic, and misinterpretations only occur due to information being retrieved from this representation in response to specific cues (Bader & Meng, 2018; Meng & Bader, 2021). Participants read implausible sentences presented in canonical or non-canonical form, followed by an Algorithmically Consistent or Good-Enough Consistent sentence. If readers formed good-enough representations of non-canonical (but not canonical) first sentences, the reading of the two follow-up sentence types should have been affected differently by first sentence canonicality. We observed evidence against such effects in our data, suggesting readers did not form good-enough representations of non-canonical sentences, but rather formed a fully specified representation. As such, the current study is more consistent with Bader and Meng's (2018; Meng & Bader, 2021) retrieval account of misinterpretation errors than a parsing account in which the initial interpretation is derived using fast-and-frugal heuristics (Ferreira, 2003).

As well as our main experiment, we presented participants with two 'control' experiments, to rule out alternative explanations for the findings of our main experiment. One alternative explanation for the lack of interaction between canonicality and follow-up type is that our participants simply did not react in a measurable way to plausibility violations, even if processing difficulty was experienced. This is ruled out by data from both our main experiment and plausibility control experiment. In our main experiment there was evidence that readers took longer to read the post-critical region of Good-Enough Consistent follow-

ups than Algorithmically Consistent follow-ups regardless of first-sentence canonicality. This suggests that, when the follow-up sentence was inconsistent with the situation described in the first, reading took measurably longer. Second, in our plausibility control items there were plausibility effects at the plausibility violation, despite the violations in these sentences being rated as less severe. Clearly, our methods and sample were appropriate for detecting any plausibility-based difficulty that did occur.

A second concern was that effects in our main experiment may be obscured by spill-over effects from processing syntactically unusual non-canonical sentences. To rule this out, we presented participants with plausible canonical and non-canonical sentences, with fully plausible follow-up sentences, and examined reading times at the regions in which we expected to observe evidence of good-enough processing in our main experiment. This data revealed evidence against a canonicality effect in the critical and post-critical region of these items. Thus, it is unlikely that spill-over effects due to non-canonicality obscured effects in our main study, and we are confident in treating our data as clear evidence against an interaction between canonicality and follow-up sentence type.

Having established that our data represent evidence against the hypothesis formulated above, it is important to consider whether alternative ways of characterising good-enough processing may better explain our data. There are two possibilities here, neither of which we find compelling. The first is that, rather than the processing of non-canonical sentences being ‘good-enough’ in terms of resulting in incorrect thematic role assignment, processing was good-enough and shallow in terms of either thematic roles not being assigned at all and remaining ambiguous (see Swets et al., 2008), readers not attempting to integrate a shallow representation of the first sentence with the second sentence (see Dwivedi, 2013), or readers having the option of integrating the follow-up sentence with either an algorithmic or good-enough representation of the first sentence (see Lim & Christianson, 2013a, 2013b). Applied

to our study, it could be argued that, for canonical sentences, readers always had an accurate representation which was integrated with the follow-up sentence. In contrast, for non-canonicals the representation would either merely suggest that there was *a peasant, king,* and *execution* with little commitment as to the agent/patient of the action, or participants would simply not attempt to integrate this representation with the second sentence. This account would predict no canonicality effect on the reading of the Algorithmically Consistent follow-up, as we found, since the follow-up sentence would be no less plausible with an undefined representation than a correct reading of the first sentence. However, this explanation is ruled out by the null canonicality effect on reading the Good-Enough Consistent follow-up. Here, readers should have experienced difficulty when reading the follow-up sentence with a fully specified representation of the first sentence suggesting *the king is dead,* while difficulty should not have occurred with underspecified representations. Thus, reading should have been faster after a non-canonical vs. canonical sentence for this follow-up in this formulation of good-enough processing. No such effect was observed, suggesting that the processing of our sentences was neither good-enough in terms of representations containing wrong thematic roles, nor underspecified thematic roles, nor a lack of integration between sentences.

A second argument against our data representing evidence against good-enough processing is that what is 'good-enough' can vary depending upon the task at hand (e.g. see Karimi & Ferreira, 2016; Swets et al., 2008). By this argument, it could be that our non-canonical sentences were processed more deeply than those in Ferreira's (2003) study by virtue of accurate representations being required to process a subsequent sentence, while Ferreira presented sentences in isolation. Thus, the reason we observed no downstream processing consequences could be due to anticipated downstream processing causing participants to process the initial sentence more deeply. We find this unlikely. In Ferreira's study, participants had to actively state who was the agent or patient of an action on 33% of

trials, and were instructed how to make these decisions before the experiment began. To us, it seems that if anything is likely to increase the care with which people assign thematic roles during language processing, it is asking them to state those roles, rather than having them read a follow-up sentence. Due to this, it seems unlikely that processing would have been deeper as a result of participants reading multi-sentence texts. Furthermore, such an account would reinforce the point that these effects only occur in fairly artificial experimental conditions, rather than as a part of normal everyday language comprehension.

Given the lack of alternative explanations for our data, we consider the current study to represent evidence against the idea that people often form inaccurate or underspecified mental representations of non-canonical sentences during parsing. As such, our work supports arguments made in Bader and Meng (2018) and Meng and Bader (2021), in which misinterpretations of non-canonical sentences occur due to post-interpretative retrieval processes driven by thematic role probes. The current work may suggest that participants are more able to accurately retrieve thematic role information from non-canonical sentences when faced with the range of cues available in natural reading as opposed to the unusual probes used in prior studies, or, alternatively, that in natural reading they retrieve such information from a more semantically rich situation model as opposed to fragile syntactic representations of the first sentence. This may explain why different paradigms tend to show different levels of good-enough processing of these type of stimuli, such that the accuracy of retrieval of information from non-canonical sentences depends upon the specific cues that trigger retrieval. Assuming this account of misinterpretation effects is correct, it is interesting to consider how our manipulation– and the processing of subsequent material in general– may affect readers' susceptibility to such memory-based effects. For example, Christianson et al. (2010) probed representations of non-canonical sentences using cues along the lines of "EXECUTIONER = PEASANT?", finding lower accuracy for implausible non-canonical

than implausible canonical sentences. In future work it may be interesting to test how the presence of a follow-up sentence affects responses to such probes. It could be that the presence of an Algorithmically Consistent follow-up sentence increases accuracy while a Good-Enough Consistent follow-up would decrease accuracy. Such an effect would suggest that integration between the two sentences provides readers with an extra set of cues to use when attempting to retrieve thematic role information.

Our study is not the first to clarify what sort of online representations are formed for sentences interpreted in a good-enough manner. Much work has shown that after reading a garden-path sentence (e.g. *While Mary bathed the baby giggled happily*) people often answer "yes" to questions such as "Did Mary bathe the baby" despite a full syntactic analysis of the sentence ruling this interpretation out (Christianson et al., 2001; Christianson et al., 2006; Ferreira et al., 2001). This finding is treated as evidence of good-enough processing, with it being proposed that readers do not form fully syntactically licit representations of garden-path sentences. However, in more recent work Slattery et al. (2013) tracked eye movements across such sentences, showing that readers responded to a later manipulation in a way consistent with having fully reanalysed the sentence. Slattery et al. argued that evidence of good-enough processing observed in comprehension may be due to a lingering semantic representation of an initial misanalysis not being fully purged from memory after reanalysis, rather than a failure to properly parse the sentence. Our own study and converging evidence from other work (e.g. Meng & Bader, 2021) similarly suggests that it is not the parsing of a non-canonical sentence that is good-enough, so much as the memory processes involved in retrieving information from the representation of that sentence.

Bader and Meng's (2018) retrieval account of misinterpretation effects is not the only theoretical framework that can explain our findings. An alternative approach explains misinterpretation effects through noisy-channel inference (e.g. Gibson et al., 2013; Levy,

2008). The basic assumption here is that perception is noisy, and so if a reader perceives an implausible utterance there is a chance this was driven by a fault in perceptual encoding. When a more plausible interpretation can be derived by relatively minor changes to the perceptual input, readers assume the more probable meaning. For implausible non-canonical sentences a more sensible meaning can be derived by assuming that the inclusion of *was* and *by* in the sentence was erroneous (e.g. *The king ~~was~~ executed ~~by~~ the peasant*) while for implausible canonical sentences the same words can be inserted to reach a plausible interpretation. Crucially, nothing in this approach dictates that readers cannot derive a veridical representation of a sentence before making post-perceptual inferences about the intended meaning, in response to the veridical representation making little sense. Thus, a lack of effect of sentence type on follow-up sentences is not necessarily problematic for noisy-channel accounts of misinterpretations, since these reinterpretations may only be made further down the line. More generally, sentences such as those used in the current study, which require multiple word changes to reach a plausible meaning, are unlikely to lead to regular misinterpretations in any case. Indeed, in the paradigm used by Gibson et al. to assess noisy channel inferences, people only make errors for non-canonical sentences ~3% of the time. Similar work examining the effect of potential misinterpretations on follow-up reading using initial sentences that are more likely to be misinterpreted in a noisy-channel account (e.g. *The mother handed the candle the daughter*; ~45% misinterpretations in Gibson et al., 2013) may be more appropriate for assessing when readers make noisy-channel inferences.

Ageing effects

As well as looking at good-enough processing in and of itself, we examined whether older adults were more likely than young adults to form good-enough representations, as prior work suggests (e.g. see Christianson et al., 2006; Malyutina & den Ouden, 2016). While we observed evidence against the three-way interaction that would have suggested that older

adults engage in more good-enough processing, the current study should not be treated as evidence against this possibility; it is entirely possible that older adults may be more likely to *retrieve* information from a representation of a non-canonical sentence in a way that leads to good-enough interpretations in response to thematic role probes. The current study merely shows that there is no evidence of either young or older adults deriving good-enough representations of non-canonical sentences in the context of discourse processing.

It is also worth discussing the lack of age differences in the effect of plausibility violations in our control items. As mentioned, much controversy exists over whether context use in language processing is age invariant (see Payne & Silcox, 2019). The data from our control items may be informative here, in that they show that, in self-paced reading, older adults experience similar disruption to young adults upon encountering implausible utterances. This mirrors results from electrophysiological work (e.g. see Lee & Federmeier, 2012) showing that older adults exhibit an enhanced N400 in response to plausibility violations in a similar way to young adults. Thus, a reader's ability to detect a clear implausibility is not affected by cognitive ageing. This lack of interaction occurred alongside a main effect of age, with older adults taking longer to read than younger adults. This finding does, however, come with several caveats. In our study the implausible word was the first word in a three-word region. It is possible that the speed with which older and younger participants detected our manipulation varied within this three-word region, even though the absolute effect size was equivalent. Furthermore, in the offline norming of these stimuli older adults did rate the implausible items as more implausible than the young adults, and so a study with more carefully controlled stimuli may still find a difference between age groups.[4]

[4] A reviewer also pointed out that the older adults in our sample spent a lot more time reading per week than the younger adults, and that it could be the case that this extra experience compensated for declines in other areas, thus resulting in our null effect. However, an analysis in which we excluded the eight older adults with most hours read per week and eight youngest adults with least hours read per week showed little difference compared to our main analysis, such that the size of the plausibility effect in each group remained near

Conclusion

In summary, we tested whether readers form 'good-enough' interpretations of implausible non-canonical sentences during their initial parse of such sentences, by examining the effect of canonicality on the reading of a follow-up sentence. Our data suggested that this was not the case, with reading behaviour being more consistent with the follow-up sentence being integrated with a veridical representation of the first sentence. This finding supports the idea that good-enough processing is more likely the product of post-interpretative retrieval processes, as opposed to being driven by the representations that people form during the initial parsing of a sentence.

identical. Thus, it seems unlikely that the greater hours read per week in older adults somehow masked a more general decline.

Declaration of Conflicting Interests

The authors declare that there are no conflicting interests.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Leverhulme Trust [grant number RPG-2019-051].

References

Anders, T. R., Fozard, J. L., & Lillyquist, T. D. (1972). Effects of age upon retrieval from short-term memory. *Developmental Psychology, 6*, 214–217. https://doi.org/10.1037/h0032103

Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. & Evershed, J. K. (2020). Gorilla in our midst: An online behavioural experiment builder. *Behavior Research Methods, 52*, 388-407. https://doi.org/10.3758/s13428-019-01237-x

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1286-1311. https://doi.org/10.1037/xlm0000519

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ, 8*:e9414. https://doi.org/10.7717/peerj.9414

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

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368-407. https://doi.org/10.1006/cogp.2001.0752

Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 538-544. https://doi.org/10.1037/a0018027

Christianson, K., Williams, C. C., Zacks, R. T., Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, *42*, 205-238. https://doi.org/10.1207/s15326950dp4202_6

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristics first, algorithmic second processing. *PloS one*, *8*, e81461. https://doi.org/10.1371/journal.pone.0081461

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164-203. https://doi.org/10.1016/S0010-0285(03)00005-7

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*, 3-20. https://doi.org/10.1023/A:1005290706460

Frazier, L., & Clifton, C., Jr. (1996). *Construal*. MIT Press.

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

Hamm, V. P., & Hasher, L. (1992). Age and the availability of inferences. *Psychology and Aging*, *7*, 56–64. https://doi.org/10.1037/0882-7974.7.1.56

Hartshorne, J. K.., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the lifespan. *Psychological Science*, *26*, 433-443. https://doi.org/ 10.1177/0956797614567339

Hasher, L., Zacks, R. T., & May, C. P. (1999). *Inhibitory control, circadian arousal, and age.* In D. Gopher & A. Koriat (Eds.), *Attention and performance. Attention and*

performance XVII: Cognitive regulation of performance: Interaction of theory and application (p. 653–675). The MIT Press.

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

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representation and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*, 1013-1040. https://doi.org/10.1080/17470218.2015.1053951

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

Klaus, J., & Schriefers, H. (2016). *Measuring verbal working memory capacity: A reading span task for laboratory and web-based use*. PsyArXiv. https://doi.org/10.17605/OSF.IO/NJ48X

Lee, C. L., & Federmeier, K. D. (2011). Differential age effects on lexical ambiguity resolution mechanisms. *Psychophysiology*, *48*, 960-972. https://doi.org/10.1111/j.1469-8986.2010.01158.x

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).

Lim, J. H., & Christianson, K. (2013a). Second language sentence processing in reading for comprehension and translation. *Bilingualism: Language and Cognition*, *16*, 518-537. https://doi.org/10.1017/S1366728912000351

Lim, J. H., & Christianson, K. (2013b). Integrating meaning and structure in L1-L2 and L2-LI translations. *Second Language Research, 29*, 233-256. https://doi.org/10.1177/0267658312462019.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703. https://doi.org/10.1037/0033-295X.101.4.676

Malyutina, S., den Ouden, D. B. (2016). What is it that lingers? Garden-path (mis)interpretations in younger and older adults. *Quarterly Journal of Experimental Psychology, 69*, 880-906. https://doi.org/10.1080/17470218.2015.1045530

Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology, 74*, 1-28. https://doi.org/10.1177/1747021820947940

Payne, B. R., & Silcox, J. W. (2019). Aging, context processing, and comprehension. *Psychology of Learning and Motivation*, *71*, 215-264. https://doi.org/10.1016/bs.plm.2019.07.001

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject.org/.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The Effect of Plausibility on Eye Movements in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1290–1301. https://doi.org/10.1037/0278-7393.30.6.1290

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.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehensions: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*, 382-386. https://doi.org/10.1016/S1364-6613(02)01958-7

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*, 104-120. https://doi.org/10.1016/j.jml.2013.04.001

Stine-Morrow, E. A. L., Miller, L. M. S., & Hertzog, C. (2006). Aging and self-regulated language processing. *Psychological Bulletin, 132*, 582–606. https://doi.org/10.1037/0033-2909.132.4.582

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*, 477-488. https://doi.org/10.1016/j.cognition.2006.10.009

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*, 201-216. https://doi.org/10.3758/MC.36.1.201

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension. The integration of habits and rules.* MIT Press.

Waters, G. S., & Caplan, D. (2001). Age, working memory, and on-line syntactic processing in sentence comprehension. *Psychology and Aging, 16*, 128–144. https://doi.org/10.1037/0882-7974.16.1.128

Zhou, P., & Christianson, K. (2016). I “hear” what you’re “saying”: Auditory perceptual simulation, reading speed, and reading comprehension. *Quarterly Journal of Experimental Psychology*, *69*, 972-995. https://doi.org/10.1080/17470218.2015.1018282

Appendix

Table A1. Mean (Standard Error) Reading Times per Condition in Multiple Regions of our Main Experimental Sentences, Separated by Age Group.

	Young				Older			
	Algorithmically Consistent		Good-Enough Consistent		Algorithmically Consistent		Good-Enough Consistent	
	Canonical	Non-Canonical	Canonical	Non-Canonical	Canonical	Non-Canonical	Canonical	Non-Canonical
Region 3	768 (19)	855 (24)	773 (22)	841 (24)	1227 (26)	1310 (28)	1247 (27)	1324 (30)
Region 4	652 (14)	665 (15)	646 (15)	664 (15)	911 (20)	934 (18)	887 (18)	952 (22)
Pre-Critical	525 (9)	532 (9)	504 (8)	528 (11)	799 (20)	816 (21)	788 (18)	798 (21)
Critical	664 (15)	637 (13)	694 (17)	676 (16)	969 (26)	960 (23)	991 (23)	994 (24)
Post-Critical	784 (21)	786 (21)	874 (24)	875 (24)	1279 (30)	1239 (27)	1415 (33)	1406 (34)

Stroop and Reading Span Test

As mentioned in the main manuscript, we included a Stroop task and Reading Span Test to measure inhibitory control and working memory, respectively. Performance on these tasks, split by age group, are shown in Table A2.

In our implementation of the Stroop task readers were shown the word *red*, *green*, or *blue* presented in red, green, or blue and had to hit one of three keys as quickly as possible to indicate the colour of the word. The Stroop score presented below was obtained by calculating a participant's mean correct response time for trials in which the printed word and the colour of the word were congruent (e.g. *red* shown in red) and subtracting it from that participant's mean correct response time for trials in which the printed word and the colour of the word were incongruent (e.g. *red* shown in blue).

The reading span task we used was adapted from an online open access version presented and tested by Klaus and Schriefers (2016), with us implementing their procedure and stimuli on Gorilla.sc. Please see the original paper for procedural details and scoring method.

Table A2. Mean Scores on Reading Span and Stroop Task in Two Age Groups.

	Young	Older
Reading Span Plausibility Judgement Accuracy	0.92 (0.26)	0.94 (0.24)
Reading Span Plausibility Judgement Speed (ms)	3977 (1603)	4581 (1658)
Reading Span Memory Performance	0.74 (0.11)	0.76 (0.14)
Stroop Accuracy	0.97 (0.18)	0.95 (0.22)
Stroop Score	107 (73)	199 (131)