# How do linguistic illusions arise? Rational inference and good-enough processing as competing latent processes within individuals

Dario Paape

May 17, 2024

**Abstract**

Non-literal interpretations of implausible sentences such as *The mother gave the candle the daughter* have been taken as evidence for a rational error-correction mechanism that reconstructs the intended utterance from the ill-formed input (*... gave the daughter the candle*). However, the good-enough processing framework offers an alternative explanation: readers sometimes miss problematic aspects of sentences because they are only processing them superficially, which leads to acceptability illusions. As a synthesis of these accounts, I propose that conscious rational inferences about errors on the one hand and good-enough processing on the other are competing latent processes that simultaneously occur within the same comprehender. In support of this view, I present data from a two-dimensional grammaticality/interpretability judgment task with different types of subtly ill-formed sentences. Both conscious rational inference and good-enough processing predict positive interpretability judgments for such sentences, but only good-enough processing also predicts positive grammaticality judgments. By fitting a lognormal race model jointly to judgments and response latencies, I show that conscious rational inference and good-enough processing, as well as purely grammar-driven processing, actively trade off with each other during reading. Furthermore, individual differences measures reveal that participant traits such as linguistic pedantry, interpretational charity, and analytic/intuitive cognitive styles contribute to variability in the processing patterns.

## 1 Introduction

Consider the sentence in (1):

(1) The mother gave the candle the daughter.

Depending on one's point of view, this sentence is either implausible or ungrammatical. If the sentence is interpreted literally, it means that the daughter was given to the candle. If the sentence is not interpreted literally, and we assume that the utterer meant to say that the candle was given to the daughter, they must have either swapped the order of the arguments by mistake, or forgotten the word *to* before *the daughter*. Furthermore, in a spoken conversation, it is also possible that the *comprehender* may have missed or misheard a word. It has been suggested that language users make these kinds of inferences about possible speech errors and other communicative slips all the time, in order to cope with the noisiness inherent in everyday communication (Levy, 2008b; Gibson et al., 2013).[1] This behavior is *rational* in the sense that it is purposive (Chater and Oaksford, 1999): the language user's goal is to reconstruct the intended message by making use of prior knowledge about sensible things that people might say. By combining this knowledge with the (subjective) probability of different kinds of errors, they can recover the most likely interpretation. In the case of (1), the inference would be that it was probably the candle that was given to the daughter, given the plausibility of the scenario and the plausibility of an argument swap as a possible speech error (Poppels and Levy, 2016).

[1] A related approach, the Rational Speech Act (RSA) framework (see Degen, 2023 for a review) also assumes that listeners reason about the communicative intent of the interlocutor. However, the RSA framework assumes that the literal meaning of a given sentence based on the veridical input is the starting point of the reasoning process.

The pure "rationalist" perspective of language processing assumes an idealized comprehender with perfect knowledge of language statistics (Gibson et al., 2013) who is not bounded by cognitive resource constraints. Levy (2008b) assumes that the error-correction process in cases like (1), where mentally editing the input yields an a-priori more plausible analysis, causes additional mental effort compared to cases in which error correction is not needed or not possible (see also Levy et al., 2009; Ryskin et al., 2021; Gibson et al., 2013; Chen et al., 2023).[2] It is thus assumed that comprehenders may go to great lengths to find out what a given sentence is most likely supposed to mean: they may invest more cognitive resources into mental error correction (Levy, 2008b), and/or reread parts of the sentence (Levy et al., 2009). This assumption may not always be realistic.

A longstanding and influential line of work in decision-making research has highlighted the fact that human rationality is bounded by constraints such as time pressure, incomplete information, and fluctuating motivation (e.g., Simon, 1972; Gigerenzer and Selten, 2002; Selten, 1990). Drawing from this literature, the "good enough" processing framework (e.g., Ferreira and Patson, 2007) in psycholinguistics has highlighted the goal of saving cognitive resources by occasionally omitting effortful processing steps. In the case of (1), this could mean that the comprehender only computes a "bag of words" style parse of the sentence and otherwise mostly relies on prior event plausibility (Paape et al., 2020; Kuperberg, 2016), fails to register the absence of the word *to*, and/or fails to actively monitor their comprehension (Glenberg et al., 1982).

Due to the shared prediction that readers sometimes adopt non-literal interpretations and partly rely on event plausibility, it has historically proven difficult to disentangle the predictions of the good-enough processing framework from those of the rational inference framework (Brehm et al., 2021). For this reason, some authors have treated rational inference as a subtype of good-enough processing (e.g., Dempsey et al., 2023; Goldberg and Ferreira, 2022). However, a closer look at the assumptions and predictions of the two frameworks reveals important differences. In contrast with the view that examples like (1) activate a potentially costly error correction mechanism, good-enough processing predicts that such sentences can often be processed with relatively little effort, assuming that the actual compositional tructure is simply ignored or at least heavily downweighted.

Good-enough processing incorporates Simon's (1955; 1956) concept of satisficing: it is assumed that readers do not fully optimize their interpretation processes in the sense that they take all available information into account, but may instead settle for an imperfect representation of a given sentence once their current *aspiration level* is reached (Ferreira et al., 2009; Christianson, 2016). When proper incentives to process utterances deeply and attentively are lacking, readers may save cognitive resources by partly ignoring the syntactic structure of a sentence and adopting a superficially plausible reading (Ferreira, 2003; Christianson et al., 2010), failing to revise initial misinterpretations in the face of disambiguating material (e.g., Christianson et al., 2001), or underspecifying their syntactic analysis of a sentence so that its semantic representation remains vague (Swets et al., 2008; Dwivedi, 2013; von der Malsburg and Vasishth, 2013).

It is clear that rational inference and good-enough processing serve different, potentially conflicting goals: to reconstruct the intended meaning of a sentence by invoking additional processes beyond the "normal" parsing of the literal string, or to save cognitive resources by omitting some parts of "normal" parsing. Given that both goals are plausible drivers of human reading (and listening) behavior, how can they be reconciled? One possibility is to move away from the focus on average behavior that dominates most of psycholinguistics (Yadav et al., 2022), and which may obscure a more complex reality: reconstructive and effort-saving mechanisms may be active to different degrees across different individuals, or within the same individual at different times, or even concurrently (Brehm et al., 2021). Some speakers may set relatively low aspiration levels and often go with their "gut feeling" when interpreting utterances, while others may expend mental effort to try and reconstruct what the other person probably wanted to say.

Importantly, there likely exists a third type of individual or processing mode that is rarely discussed in the literature: people who are pedantic — which I use neutrally here — in the sense that they are completely faithful to the linguistic stimulus, and who take sentences like (1) literally, responding with "That's nonsense!", "I don't know what you're trying to say!", or "What an unusual thing to happen!". Literal interpretations of implausible sentences are robustly attested (e.g., Gibson et al., 2013; Ferreira, 2003) and have been linked to high verbal working memory (Bader and Meng, 2018; Meng and Bader, 2021; Stella and Engelhardt, 2022). A completely input-faithful speaker would arguably be an

[2] Starting with Levy (2011), a different line of work has integrated a noisy context representation with surprisal (Levy, 2008a; Hale, 2001) as the main determinant of processing difficulty. The resulting model makes very different predictions from the original noisy-channel model, which I present in more detail below.

embodiment of pure grammatical competence in the Chomskyan sense (Wray, 1998), which makes literal responses theoretically highly interesting. Any realistic model of sentence comprehension should thus take into account that different people or even one and the same person may, depending on the situation, be "inferencers", "slackers", or "pedants".

How can each of these processing "modes" be identified? One promising avenue is to use error awareness as an indicator. In the "slacker" mode of processing, ill-formed sentences should be very likely to pass unnoticed. Such cases are known as linguistic illusions, where an ungrammatical sentence passes as grammatical and/or an implausible sentence passes as plausible (e.g., Muller (2022); Phillips et al. (2011); Sanford and Sturt (2002)). The rational inference account is underspecified with regard to error awareness: Levy (2008b, p. 237) states that a copy editor needs to "notice and (crucially) correct mistakes on the printed page" but also that "in many cases, these types of correction happen at a level that may be below consciousness — thus we sometimes miss a typo but interpret the sentences as it was intended" (see also Huang and Staub (2021b)).

That rational inference *can* be conscious is implied by studies that have used highly explicit tasks such as retyping of sentences (Ryskin et al., 2018) or judging how likely one sentence is to be changed into a different one due to a speech error (Zhang et al., 2023b). There is also evidence that error correction via rational inference leads to increased P600 amplitudes (Ryskin et al., 2021; Li and Ettinger, 2023), which have been linked to conscious detection of anomalies during reading (Coulson et al., 1998; Rohaut and Naccache, 2017; Sanford et al., 2011). Crucially, while the additional processing steps involved in rational inference may not *always* be conscious, they should be comparatively more likely to rise to consciousness than good-enough processing, which implies the *absence* of one or more processing steps. Finally, the "pedantic" processing mode naturally predicts error awareness, in the sense that the utterance is identified as being ungrammatical or nonsensical.

In addition to differences between people, there are likely to be differences between error types that create variability in how a speaker responds to a sentence (Frazier and Clifton, 2015). Some sentences, such as the "depth charge" sentence *No head injury is too trivial to be ignored* (Wason and Reich, 1979) cause an illusion of acceptability almost invariably across individuals (Paape et al., 2020), while other sentence types may show large amounts of variability both across and within speakers (Goldshtein, 2021; Hannon and Daneman, 2004; Leivada, 2020; Frank et al., 2021; Christianson et al., 2022). The aim of the present study is to quantify this variability between speakers and sentences across six different constructions that are known to cause linguistic illusions, and to identify traits that correlate with speakers' dispositions towards rational inference or good-enough processing. Table 1 lists the six constructions under investigation.

An additional contribution of the present work is the use of a computational modeling approach to investigate how responses to sentences are generated in real time. In the model presented below, conscious rational inference, good-enough processing and outright rejection of a sentence are treated as latent processes that compete and trade off with each other to produce a response. In combination with the empirical breadth of the experimental design, this allows for the comparison of the three processes across different linguistic constructions. Furthermore, the model allows for a systematic investigation of participant-level traits that affect each of the latent processes.

I will now introduce the logic of the experiment, followed by the description of the procedure, and finally the implementation of the computational model.

# 2 Experimental study

The sentence judgment study presented below had three main aims:

1. To quantify the relative contributions of rational inference and good-enough processing to different linguistic illusions.
2. To investigate individual-level trade-offs between rational inference and good-enough processing across different illusions.
3. To investigate the effect of individual-level traits such as linguistic pedantry on rational inference, good-enough processing, and outright rejection of illusion sentences.

In order to achieve the first aim, the study used a novel two-dimensional sentence judgment task in which participants simultaneously judge whether they feel that they understand the sentences ("get it"/"don't get it"), in addition to whether they think that the sentences are formally correct ("correct"/"incorrect"). Under good-enough processing, readers should occasionally miss formal

**Inversion**: Order of direct and indirect object is swapped (e.g. Gibson et al. (2013))

*The mother gave the candle the daughter.*

---

**Agreement attraction**: Verb agrees with intervening noun phrase instead of subject noun phrase (e.g., Bock et al. (2001))

*The waitress who sat the girls unsurprisingly were unhappy about all the noise.*

---

**Depth charge**: Incongruous degree phrase "saved" by negation (e.g., Wason and Reich (1979))

*In Maria's class, no test is too difficult to fail.*

---

**Comparative illusion**: Number of individuals compared to number of events (e.g., Wellwood et al. (2018))

*More engineers relocated to San Francisco than our accountant did.*

---

**Missing VP illusion**: Three clauses with three subjects but only two verbs (e.g., Gibson and Thomas (1999))

*The manuscript that the student who the catalog had confused ____ was missing a page.*

---

**NPI illusion**: Negative polarity item (NPI) *ever* licensed by embedded negation (e.g., Drenhaus et al. (2005))

*The authors that no critics recommended have ever received acknowledgment for a novel.*

Table 1: Constructions used in the experimental study.

grammatical errors, resulting in the impression that illusion sentences are both well-formed and interpretable ("get it, correct"). Under rational inference, by contrast, it is plausible to assume that readers notice the errors — especially when instructed to look out for them — but can nevertheless reconstruct the presumably intended sentence ("get it, but incorrect"). I assume that the explicit task demands will cause all rational inferences to be conscious rather than non-conscious. This linking assumption appears justified given the aforementioned reliance on explicit tasks in the relevant literature on rational inferences, but I will later explore the possibility that some inferences may also be non-conscious. The remaining two judgment options ("don't get it, incorrect"/"don't get it, but correct") are not covered in any depth by either theory, but may nevertheless show differences between illusions: readers may outright reject some ungrammatical constructions more readily than others.

To achieve the second aim, I analyze the correlations between the subject-level random effects in a hierarchical computational model. The model assumes that a given manipulation has an average effect across all participants, and that the individual effects for each participant are normally distributed around this average. Analyzing the correlations between individual effects allows for statements of the form "Participants who are more likely than average to do good-enough processing for illusion X are also more likely than average to do good-enough processing for illusion Y". Furthermore, and perhaps more interestingly, correlations can be analyzed not only within but across response types: "Participants who are more likely than average to do good-enough processing for illusion X are less likely than average to engage in rational inferences for illusion Y". Finding such negative correlations would strengthen the case for shared cognitive mechanisms across different illusions, and crucially also yield insights into whether and how these mechanisms compete within individuals.[3] Finding such trade-offs would lend support to the notion that good-enough processing and rational inference serve conflicting goals, and suggest that both mechanisms draw on a shared pool of resources, as has been suggested in other models of conflict tasks (e.g., Lee and Sewell, 2024).

The third aim is to uncover the underlying factors that contribute to individual differences in the processing of illusion sentences by collecting additional measures outside of the sentence judgment task. The first measure I use is a simple questionnaire that covers linguistic pedantry, interpretational charity, motivation, and attention. The second measure comes from a syllogistic reasoning task with believable and unbelievable syllogisms, which is intended to uncover individual differences in cognitive

[3]See Brown (2021) for a related discussion of how random-effects correlations in linear mixed models can be leveraged to answer specific research questions.

style (Trippas et al., 2015, 2018; Stupple et al., 2011): a more analytic cognitive style "denotes a propensity to set aside highly salient intuitions when engaging in problem solving" (Pennycook et al., 2012, p. 335). Having an analytic cognitive style should increase an individual's tendency to favor analytical grammar rules and logic over "quick and dirty", good-enough processing of illusion sentences.

It is important to note that there is broad agreement in the literature that the illusion constructions tested in the current study are ill-formed at both the prescriptive and the descriptive level: there is no variety of English, vernacular or otherwise, that "allows" sentences with missing lexical verbs or sentences in which direct and indirect objects appear in reverse order without an added preposition (e.g., *The mother gave the candle the daughter*). The errors tested here are thus not "native errors" in the sense of Bradac et al. (1980), that is, constructions that speakers might use but that they are taught to avoid in formal education, or on which the majority of naive and expert judgments fundamentally disagree (with the possible exception of depth charge sentences; see general discussion). Consequently, while the judgments of formal correctness elicited in the current study may activate participants' prescriptive attitudes to some degree — which is part of the rationale behind the task — they also plausibly reflect their natural intuitions with regard to what constitutes a well-formed sentence of English. The instructions (reproduced below) were deliberately designed to strike a balance between allowing for superficial, good-enough processing on the one hand and triggering error awareness on the other: no comprehension questions were asked and participants were explicitly instructed not to "overanalyze" the sentences, but they were also told to be on the lookout for errors.

It should be stressed that the goal of the present design is not to establish that the illusions in question *exist*, that is, that the illusion variants of the different constructions are more acceptable/interpretable than their ungrammatical counterparts; against the background of the existing literature, this is taken as a given. By contrast, the present design is focused mainly on the comparison between illusion sentences and their *grammatical* counterparts: good-enough processing predicts "get it, correct" judgments for both illusion sentences and their grammatical controls, whereas rational inference predicts "get it, incorrect" judgments for illusion sentences but not for controls.

## 2.1 Participants

Participants were tested in three groups that were recruited over Prolific (`https://www.prolific.co`; Palan and Schitter, 2018). Group 1 completed only the sentence judgment task, Group 2 additionally completed the questionnaire, and Group 3 additionally completed the questionnaire and the syllogistic reasoning task. Group 1 consisted of 100 self-identified native speakers of English currently living in the US. The first 50 participants were initially paid £2.83 each[4], which was adjusted to £3.55 after review, as the completion time estimate had been too low. The remaining 50 participants were paid £3.55 each. The data of one participant were subsequently removed because response times were consistently too short to be realistic, leaving data from 99 participants. Groups 2 and 3 consisted of 157 participants and 100 participants, respectively, recruited from the same subject pool. Participants in Group 2 were paid £3.55 each while participants in Group 3 were paid £5.65 each due to the longer experiment duration.

## 2.2 Materials

12 inversion sentences were adapted from Cai et al. (2022). Inversion sentences appeared in two conditions, the normal condition and the inverted condition, as shown in (2). In 6 items, both the inverted sentence and the control sentence used the direct object construction, while in the other 6 items both sentences used the prepositional object construction with *to*.

(2) The mother gave { a. the daughter the candle / b. the candle the daughter } before bedtime.

12 agreement attraction sentences were adapted from Parker and An (2018). Agreement attraction sentences appeared in three conditions, as shown in (3). All participants saw the *waitress*/*girls* (attraction) condition. In Group 1, 50 participants saw only the *waitress*/*girl* (ungrammatical) condition as a control, while the other 50 participants saw only the *waitresses*/*girls* (grammatical) condition

[4]Prolific is a UK-based company, so renumeration is calculated in British Pounds.

as a control. Group 2 saw only the grammatical control condition, while Group 3 saw no agreement attraction sentences at all.[5]

(3) The { waitresses / waitress } who sat the { a. girls / b. girl } unsurprisingly were unhappy about all the noise.

12 depth charge sentences were adapted from O'Connor (2015). Depth charge sentences appeared in three conditions, as shown in (4). All participants saw the *no*/*too* (illusion) condition. In Group 1, 50 participants saw only the *no*/*so* condition (sensible) as a control, while the other 50 participants saw only the *some*/*too* condition (not sensible) as a control. The remaining participants saw only the *so* condition as a control.

(4) In Maria's class, ...
   a. ... no test is so difficult that she would fail it.
   b. ... no test is too difficult to fail.
   c. ... some tests are too difficult to fail.

The depth charge illusion is often difficult to spot even for trained linguists (Wason and Reich, 1979). The crucial insight is that the degree phrase *too difficult to fail* is pragmatically incongruous (compare *too easy to fail*). The illusory property of this construction is that having a negation at the beginning of the sentence (*no test*) can completely mask the incongruity. However, in logical terms, asserting that no test has the property of being *too difficult to fail* should not repair the incongruity (Paape et al., 2020).

12 comparative illusion sentences were adapted from O'Connor (2015). Comparative illusion sentences appeared in two conditions, the plural condition (grammatical), and the singular condition (ungrammatical), as shown in (5). All comparative illusion sentences contained a sentence-final subordinate clause (*because ...* ; *due to ...* ) to make them sound more natural.

(5) Last fall, more engineers relocated to San Francisco than { a. accountants / b. our accountant } did, ...

12 missing VP sentences were taken from Langsford et al. (2019). Missing VP sentences appeared in two conditions, the VP2 condition (grammatical) and the no-VP2 condition (ungrammatical), as shown in (6).

(6) The ancient manuscript that the grad student who the new card catalog had confused a great deal { a. was studying in the library / b. —— } was missing a page.

12 NPI sentences were adapted from Parker and Phillips (2016). NPI sentences appeared in three conditions, as shown in (7). All participants saw the *the authors*/*no critics* (embedded negation) condition. 50 participants saw only the *no authors*/*the critics* condition (grammatical) as a control, while the other 50 participants saw only the *the authors*/*the critics* condition (ungrammatical) as a control. The remaining participants saw only the grammatical control condition.

(7) { No authors / The authors } that { a. no critics / b. the critics } recommended have ever received acknowledgment for a best-selling novel.

In addition to the 72 illusion and control sentences, there were also 24 fillers. Of these, 12 were garden-path sentences (e.g., *The farm hand believed that while the fox stalked the geese continued to peck ...* ), 6 contained "malaphors" or "idiom blends" (e.g., *Bill wasn't the sharpest bulb in the box*), and 6 were relatively long but well-formed sentences of different types.

Group 2 completed the same judgment task as Group 1, followed by an additional questionnaire that appeared at the end of the experiment. Participants were asked to what extent they agreed or disagreed with the following four statements:

[5]Group 3 received an additional logic task, which made the experiment longer overall, so the judgment portion was sized down. Agreement attraction sentences were chosen for removal because number agreement has no obvious logical component to it, unlike the other sentence types.

1. Doing the experiment was fun for me.
2. It bothers me a lot when people use incorrect grammar.
3. I usually assume that what people say makes sense.
4. In my everyday life, when I read a text, I always pay close attention to every sentence.

There were five levels on the scale: “completely disagree”, “somewhat disagree”, “undecided”, “somewhat agree”, “completely agree”.

Group 3 completed a syllogistic reasoning task in addition to the judgment task and the questionnaire. The materials consisted of 64 syllogisms in four conditions resulting from crossing the factors validity (valid versus invalid) and believability (believable versus unbelievable). These factors were manipulated between items, that is, each syllogism appeared in only one condition. Examples for each condition are shown in (8). The structure of the syllogisms varied between items. Some materials were novel while others were adapted from previous studies (Goel and Vartanian, 2011; Solcz, 2011; Hayes et al., 2022).

(8) a. Either the sky is blue or it is green. The sky is not green. Therefore, the sky must be blue. (valid, believable)

b. All rabbits are fluffy. All fluffy creatures are tadpoles. Therefore, all rabbits are tadpoles. (valid, not believable)

c. If an animal is a feline, then it purrs. If an animal purrs, then it is a cat. Therefore, if an animal is a cat, then it is a feline. (invalid, believable)

d. Some sodas are beverages. All sodas are carbonated drinks. Therefore, some carbonated drinks are not beverages. (invalid, not believable)

### 2.2.1 Procedure

All subjects gave informed consent to participate in the study,[6] which was run on the PCIbex farm (Schwarz and Zehr, 2021). Given the potential importance of task demands for the results, the experimental instructions are reproduced here in their entirety[7]:

> People often make mistakes when they speak or write. They will say things like “is sufficient enough for”, “pales next to comparison of”, or even “I’m going to get some bed”. Such utterances can be considered incorrect or nonsensical, but it is nevertheless clear what the person in question was trying to say. In this study, you are supposed to judge both the **formal correctness** of the sentences you will read, as well as indicate whether you know what the other person **meant**.
>
> Some of the sentences will be very complex, and you may feel that you don’t understand them, but still get the impression that they are formally correct. Sometimes you may be completely unsure. This is fine; you can just choose the appropriate answer option (“no idea”).
>
> Don’t “overanalyze” the sentences — try to read normally as much as possible. There will be no detailed comprehension test. We are mainly interested in whether you **feel** that you understood the sentence.

In each trial, the stimulus sentence was presented at the center of the screen, with five possible judgment options shown directly below:

1. 😊👍 Get it, correct
2. 😊👎 Get it, but incorrect
3. 😐👍 Don't get it, (probably) correct
4. 😐👎 Don't get it, incorrect
5. ❓❓ No idea

[6] No ethics approval was obtained prior to conducting the study, as non-invasive studies involving healthy participants are exempt from ethics approval according to the relevant laws and research guidelines in Germany.

[7] The speech error examples are taken from Frazier (2015).

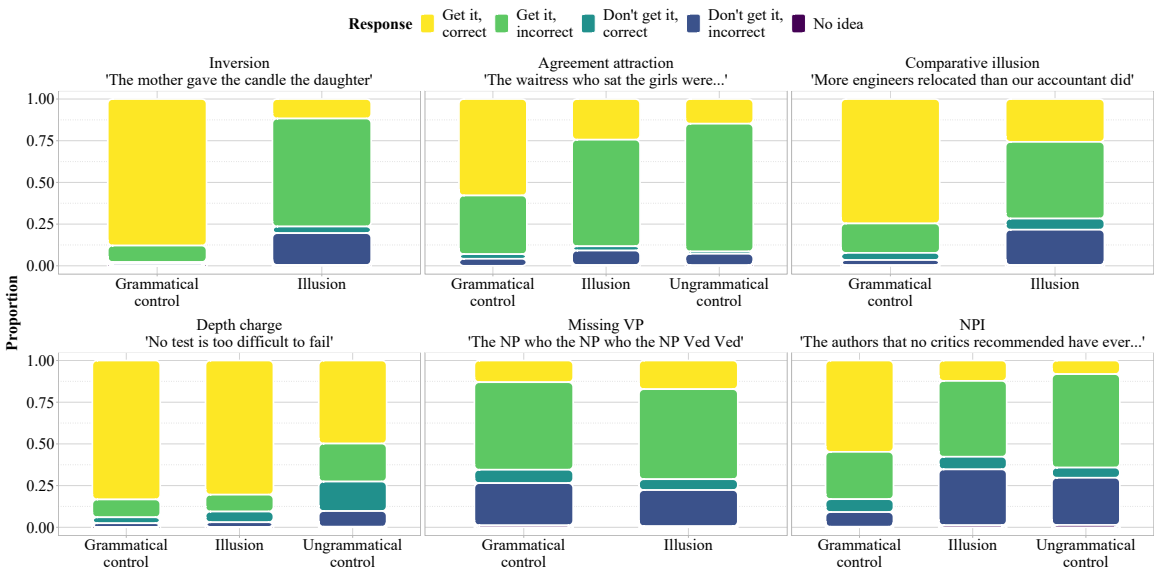


Figure 1: Response proportions across constructions and conditions.

The time elapsed from presentation of the sentence to choosing a response was recorded for each trial.[8] There was no time limit. Responses could be chosen by either clicking on them or by pressing the appropriate number key. Sentences were rotated through the conditions in a Latin squares fashion. There were no practice trials. The median duration of an experimental session in Group 1 was 20 minutes. The median duration of an experimental session for Group 2 was 21 minutes.

In Group 3, the order of the sentence judgment task and the syllogistic reasoning task was counterbalanced across participants, so that 50 participants completed the reasoning task first and the remaining 50 participants completed the judgment task first. In the syllogistic reasoning task, participants were instructed to judge the syllogisms purely based on logic ("logically valid" versus "not logically valid"), and to assume that the premises were true, even if they didn't make sense. As for the judgment task, the entire syllogism was presented on the screen, with the response options shown underneath. At the end of the experiment, the same questionnaire as in Group 2 was administered. The median duration of an experimental session in Group 3 was 33 minutes.

## 2.3 Data preparation

The data were analyzed in R (R Core Team, 2022). The reaction time and response data from all three participant groups were combined into one data set. Trials with response times below 2 seconds or above 60 seconds were dropped, which resulted in a loss of 5% of the data. Additionally, trials with "no idea" responses were dropped, which resulted in a loss of 0.5% of the remaining data.[9] The final data set contained 22,884 observations from 355 participants.

## 2.4 Descriptive results

Figure 1 shows response proportions across constructions and conditions. Figure 2 shows judgment times (reading time + response selection time) across constructions, conditions, and response types.

A number of high-level observations can be made about the descriptive results:

[8] One could have collected two sequential judgments instead: a binary judgment of formal correctness followed by a judgment of meaning recoverability, or vice versa. However, this would not have aided interpretability: the judgment latencies presumably do not reflect the entire decision process, as readers likely prepare their judgment already while reading the sentence. Furthermore, sequential judgments may have biased readers towards a certain processing style ("meaning-first" or "correctness-first"), which would complicate the interpretation of the results.

[9] This step resulted in the complete removal of data from one subject, who always responded with "no idea".

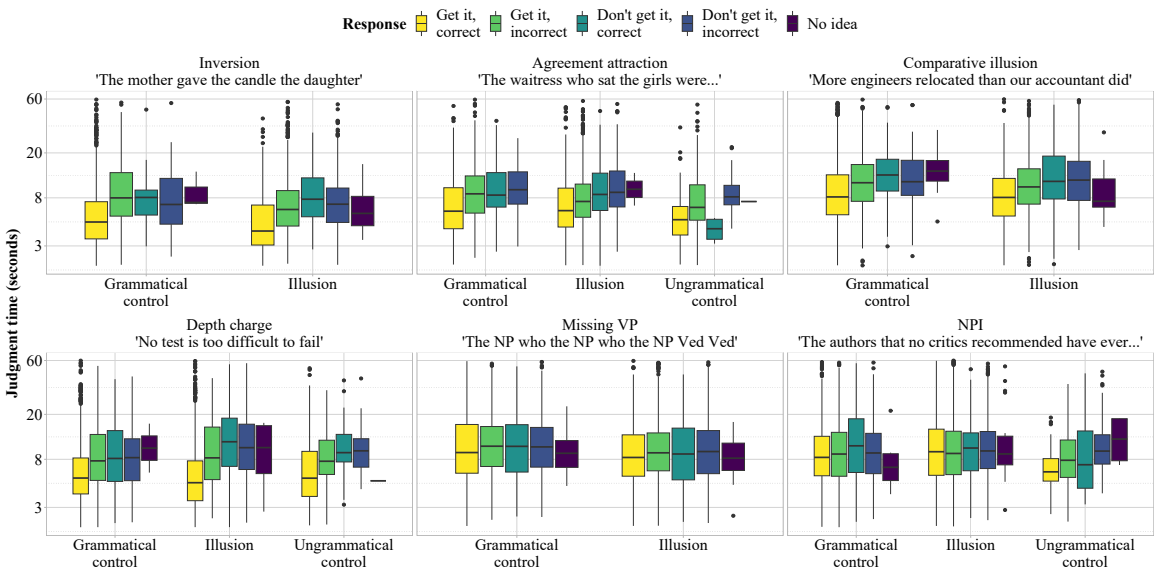


Figure 2: Judgment times (reading time + response selection time) across constructions, conditions, and response types. Note that the y-axis is log-scaled.

- Across constructions, "get it, incorrect" judgments are more common in the illusion conditions than "get it, correct" judgments, suggesting that rational inference is more common under the current instructions than good-enough processing. The only exception is the depth charge construction, which shows more than 75% "get it, correct" judgments.
- Across constructions, even the fully grammatical control sentences show non-zero proportions of "incorrect" judgments. This tendency is especially pronounced for the agreement attraction sentences of Parker and An (2018)[10] and the NPI sentences of Parker and Phillips (2016).
- It is informative to compare the illusion conditions to both grammatical and ungrammatical control conditions. For agreement attraction sentences, the judgment pattern in the illusion condition is close to midway between the grammatical and ungrammatical conditions in terms of "correct" judgments, while for NPI sentences, the illusion condition is much closer to the ungrammatical condition. This is in line with the qualitative pattern observed by Xiang et al. (2013) and Langsford et al. (2019), who also compared the two constructions.
- The missing VP illusion shows comparably high amounts of "get it, incorrect" and "don't get it, incorrect" judgments both in the grammatical *and* in the illusion condition. This is consistent with the assumption that multiple center embeddings overload the parser's capacity (e.g., Gibson and Thomas, 1999).
- For the depth charge illusion, the grammatical condition with *so* is almost indistinguishable from the illusion condition with *too*, which essentially patterns like a fully grammatical sentence. This result is unexpected if the depth charge illusion is the result of a rational inference mechanism, as recently argued by Zhang et al. (2023b).
- Across constructions, latencies tend to be shorter for "get it, correct" judgments, as would be expected if the sentence is either grammatical or an ungrammatical sentence is processed superficially. However, this pattern does not seem to hold for missing VP and NPI sentences, which casts doubt on the assumption that positive grammaticality judgments for illusion sentences are always the result of effort-saving mechanisms.

[10]Speculatively, the high rejection rates may be due to the placement of the adverb between the attractor and the verb (*The waitress who sat the girls unsurprisingly were* . . . ), which is anecdotally regarded as incorrect by some speakers.

# 3 Computational modeling

## 3.1 Preprocessing of individual differences predictors

Model-based preprocessing was applied to the individual differences measures from Groups 2 and 3 prior entering them into the computational model. The questionnaire responses were subjected to a principal components analysis with polychoric correlations using the psych package (Revelle, 2023). The number of factors was set to 2. The resulting factor loadings are shown in Table 2. As the table shows, Factor 1 loaded positively on having fun during the experiment, being bothered by grammatical errors, and paying attention during reading, but loaded negatively on assuming that utterances typically make sense. Factor 2, by contrast, loaded positively on making this default assumption of sensibleness, but less positively on attention, and negatively on being bothered by grammatical errors.

| | Factor 1 (PEDANTRY) | Factor 2 (CHARITY) |
|---|---|---|
| had fun | 0.62 | 0.58 |
| bothered by incorrect grammar | 0.66 | −0.46 |
| assume that people make sense | −0.16 | 0.82 |
| attentive reading | 0.84 | 0.09 |

Table 2: Factor loadings from the principal components analysis of the questionnaire data.

Given these loadings, I will assume that Factor 1 captures motivation to rigorously apply grammatical rules and pay attention to linguistic detail, which I will call PEDANTRY, while Factor 2 captures the default assumption that sentences are sensible without much regard for correct grammar, which I will call CHARITY. These labels are purely descriptive and should be treated with some caution. As Table 2 shows, the two factors are not complementary: both load positively on enjoying the experiment, and both also load positively on attention, though the association is stronger for PEDANTRY. If the proposed labeling is on the right track, individuals who score high on PEDANTRY should show less good-enough processing for illusion sentences, meaning fewer "correct" responses, while individuals who score high on CHARITY should give more "correct" responses and fewer "don't get it" responses.

For both factors, the factor scores for each participant were entered as scaled, centered predictors into the main analysis. The data from the syllogistic reasoning task completed by Group 3 were analyzed using a hierarchical logistic regression model in brms (Bürkner, 2017). The factors validity and believability were sum-coded for this analysis. Response time was also entered as a predictor to account for speed-accuracy trade-off. The subject-level random effects for validity, believability and their interaction were extracted from this model and entered into the main analysis as centered, scaled predictors. I will call these predictors LOGIC, BELIEF, and CONFLICT.

## 3.2 Modeling rationale and implementation

In psycholinguistics, so-called "online" measures such as reading times are usually analyzed separately from "offline" measures such as acceptability ratings. In cognitive psychology, by contrast, the standard is to look at the measured latencies and observed responses in a task as reflecting the same mental process, which is often conceptualized in terms of evidence accumulation or sequential sampling (Evans and Wagenmakers, 2019; Ratcliff et al., 2016). The core assumption of evidence accumulation models is that the time spent processing a given stimulus — modulo the time required to, say, press a keyboard key — reflects the time needed to extract enough information from the stimulus to be able to respond to it. In the context of a sentence judgment study, one can assume that while the participant reads a given sentence, they are extracting information that is relevant to the task of making the judgment: identifying words, accessing their meaning, (possibly) computing the compositional structure of the sentence, consulting their mental grammar, and potentially making inferences beyond the literal input.

One way of modeling the accumulation of evidence is to assume a race between accumulators that compete to produce a response. Race models have been applied with great success to many decision tasks in cognitive psychology (Heathcote and Matzke, 2022). A race model that is relatively straightforward to implement and interpret is the lognormal race model of Rouder et al. (2015). In the lognormal race model, the different response options accrue evidence in parallel, with the fastest option

winning and determining the observed response in a given trial. Under this model assumption, each trial also yields information about the unobserved responses, given that they must have been slower than the observed response. It is an open question whether conceiving of good-enough processing, rational inference and rejection as parallel processes is warranted; they could, in principle, occur sequentially, or there could be "switching" between different modes of processing within a trial (Ferreira and Huettig, 2023). Nevertheless, the parallelism assumption is a useful simplification, and does not preclude the possibility that the latent processes have considerable overlap in terms of the information they make use of (see below).

The notion that different mechanisms are available to derive sentence meaning, and that these mechanisms operate in parallel, is well-established in the literature, especially in the context of "heuristic" versus "algorithmic" interpretations (e.g., Ferreira, 2003; Ferreira and Huettig, 2023; Dempsey et al., 2023; Kuperberg, 2007). For instance, most variations of the good-enough processing approach do not assume that the sentence processor *only* applies superficial heuristics, but that heuristics and fully compositional processing are carried out in parallel. Heuristic processing is often, but not necessarily always faster than fully compositional processing (Dempsey et al., 2023), which may outperform heuristic processing in cases where the input cannot be easily coerced to fit into a"template" such as agent-verb-object (Ferreira, 2003). The novel aspect of my model is that a third parallel route to meaning, namely rational inference, is assumed.

What is crucial to the model is the assumption that the responses in the experimental task can be reliably mapped to the proposed processes. Given the recent proposal that rational inference can be pre-perceptual (Huang and Staub, 2021b,a), there is a possibility that "get it, correct" responses may be produced by mental "repairs" that the reader takes no conscious notice of. However, the conscious nature of the experimental task should make this type of non-conscious rational inference unlikely. Nevertheless, I compare the race model against a serial model that incorporates non-conscious rational inference using cross-validation, and find that the race model has better predictive fit to the data.

I implement the lognormal race model in Stan (Stan Development Team, 2023) and fit it to the judgments and their associated reading/judgment times, that is, the time taken to read the entire sentence and produce a response. This measure is, of course, very coarse; the "atomic" events of reading such as fixations on single words, and possibly rereading of (parts of) the sentence, are not captured. However, the model is not intended as a model of these lower-level processes, but rather treats them as means to an end; the goal is to accumulate evidence to reach a decision about a judgment: does the sentence make sense, and is it grammatically well-formed?

As a further simplification, the two types of "don't get it" responses ("correct"/"incorrect") are coded as a single response category, so that there are three accumulators in the model: REJECT, INFER, and GOOD. Their finishing times FT in a given trial $i$ are each sampled from a lognormal distribution with mean $\mu$ and standard deviation $\sigma$:

$$\begin{aligned} FT\_\text{REJECT}_i &\sim Lognormal(\mu_1, \sigma_1) && \text{(Response "Don't get it")} \\ FT\_\text{INFER}_i &\sim Lognormal(\mu_2, \sigma_2) && \text{(Response "Get it, but incorrect")} \\ FT\_\text{GOOD}_i &\sim Lognormal(\mu_3, \sigma_3) && \text{(Response "Get it, correct")} \end{aligned} \tag{9}$$

The REJECT accumulator is taken to represent complete failure to understand the sentence, the INFER accumulator is taken to represent conscious rational inference processes, and the GOOD accumulator is taken to represent good-enough processing. As the stimulus sentence is the same for each assumed latent process, some percentage of the finishing time across the three accumulators for each trial is going to be identical. For instance, each of the response options presumably requires the words in the sentence to be identified by fixating them and accessing the mental lexicon. However, things may already start to diverge at this point: under good-enough processing, less information may be retrieved from the lexicon compared to the other racing processes, which should make the accumulator faster on average. By contrast, rational inference by assumption involves additional reasoning steps beyond the veridical encoding of the input, meaning that the INFER accumulator must have some percentage of its finishing time devoted to these steps, presumably slowing it down relative to the GOOD accumulator. Figure 3 schematically shows the contributions of the component processes to the total accumulation time of each accumulator. The different $\mu$ parameters encode these base differences in speed between accumulators, which are additionally assumed to be affected by the experimental manipulations and individual differences between participants.

The three accumulators are assumed to be active across grammatical, ungrammatical, and illusion sentences. A somewhat puzzling observation based on the descriptive statistics is that some partic-

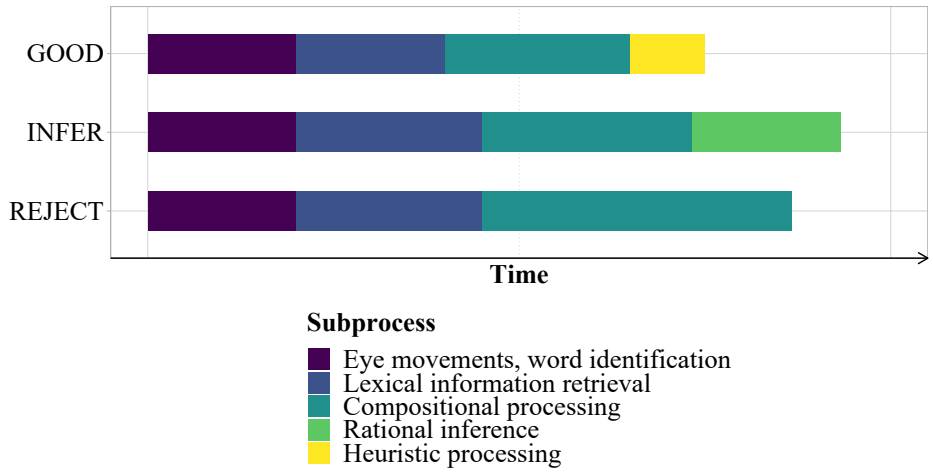


Figure 3: Contributions of component processes to the finishing times of the three accumulators in a hypothetical experimental trial.

ipants give "get it, incorrect" or even "don't get it" judgments for grammatical sentences, which begs the question why the INFER and REJECT accumulators should ever win if the sentence is fully well-formed. One possibility is that participants find some of the grammatical sentences stylistically odd, and that the experimental design may have created some amount of hypervigilance. Especially for missing VP sentences, another possibility is that participants are unable to parse the well-formed structure, and therefore reject it or try to infer the meaning rather than compostionally deriving it. Ultimately, the question of why rejection rates are rarely zero for sentences that are designed to be fully grammatical applies to may (psycho)linguistic studies, and should be given more attention in the literature, especially if the aim is to design naturalistic stimuli. Nevertheless, the presence of a grammatical baseline is important for gauging "how far away" illusion sentences are from it in terms of acceptability and interpretability, even more so if the baseline does not show perfect acceptability either.

The mathematical structure of $\mu_x$ for each accumulator $x$ in trial $i$ is that of a linear mixed-effects model with an intercept $\alpha$, slopes $\beta$ for the predictors $1...n$, and normally distributed by-subject and by-item adjustments $u$ and $w$ for both intercepts and slopes. Predictors include the factor-coded conditions within each construction type, the individual differences measures, and their interactions. During fitting, the observed response in a given trial determines which accumulator's $\alpha$ and $\beta$ parameters are assumed to have produced the observed reaction time. As non-observed responses from the remaining accumulators must have been slower by assumption, some information is also gained regarding their corresponding $\alpha$ and $\beta$ parameters.[11]

$$\mu_{x,i} = \alpha_x + u_i + w_i + \beta_{n,x,i} \cdot \text{predictors} \tag{10}$$

The observed reaction time in a given trial $i$ is the finishing time of the winning accumulator, plus a shift parameter estimated from the data that represents non-decision time, which can vary between subjects.

$$RT_i = min(FT\_\text{REJECT}_i, FT\_\text{INFER}_i, FT\_\text{GOOD}_i) + \text{SHIFT} \tag{11}$$

IF...ELSE constructions in Stan are used to exclude the individual differences predictors where no data is available, that is, for the syllogistic reasoning predictors in Groups 1 and 2, and for the

[11] In practice, this is achieved by using the complementary cumulative distribution function ("all values greater than x") rather than the probability density for the accumulator in question.

questionnaire-based predictors in Group 1. To account for differences in sentence length between constructions, sentence length in characters is added as a predictor to each $\mu$.

Given the assumptions of the lognormal race model, if one accumulator becomes faster in a given condition, there will be more responses of the associated type, and fewer responses in the other response categories. This is true even if the finishing times of the other accumulators are unaffected by the manipulation, because, like in a real-life race, it is the *relative* speed of the competitors that determines the winner. Thus, if a given manipulation speeds up the INFER accumulator, it does not follow that the GOOD and REJECT accumulators are slowed down by the same amount, or even slowed down at all. Unlike in competition-based models, where strengthening of one response option automatically implies weakening of the other options through inhibition, race models in principle allow response options to accrue evidence completely independently from one another (Teodorescu and Usher, 2013). However, whether the racing processes are truly independent is a question of implementation (Heathcote and Matzke, 2022), and ultimately a question of whether assuming independence is plausible and empirically warranted. For instance, in the present implementation, the accumulators share a common global intercept, as well as shared hierarchical adjustments to the finishing times, which are meant to capture global differences in information extraction speed between participants.

Despite some shared structure across accumulators, the focus of the present work is on disentangling rational inference and good-enough processing, and on identifying possible trade-offs between the different mechanisms. Due to the structure of the judgment task, giving one response naturally precludes giving another response in the same trial. However, through the inclusion of reaction times in the model, it can be inferred if the other responses are being "actively" or "passively" suppressed. *Active trade-offs* would entail that the data are more likely under a model where the $\mu$ parameters of multiple accumulators are affected by a manipulation, whereas *passive trade-offs* would entail that only one accumulator's $\mu$ parameter is affected. In such cases, there will be more responses of the given type and fewer responses of the other types, but no indication in the data that a speedup in one accumulator is accompanied by a slowdown in the others. Individual-level trade-offs are modeled via random-effects correlations in the model. The empirical question is whether, for instance, a subject with a strong REJECT tendency, as reflected by a hierarchical adjustment that encodes faster finishing times for that accumulator, will show *slower* finishing times for the GOOD and/or INFER accumulators.

The fundamental ability of the model to correctly identify active trade-offs was established by means of a simulation study. Response and latency data were generated from a race process with three accumulators. Experimental manipulations were simulated that either affected multiple accumulators in opposite directions or just one of the accumulators, and the model was fitted to the generated data. Despite similar descriptive patterns in the summary statistics, the model was able to correctly recover the data-generating parameters in all cases.

## 3.3 Modeling results, cross-validation, and discussion

### 3.3.1 Population-level effects

Figure 4 shows the predicted finishing times of the three accumulators in the lognormal race model across constructions and conditions. The figure allows for several types of visual comparisons: each panel shows one construction, while the outline colors show the different conditions (illusion versus control). For each accumulator (GOOD, INFER, REJECT), the conditions can be compared within a given construction, with *shorter* finishing times meaning *more* responses of that type. In addition, finishing times can be compared *across* constructions by visually comparing the respective panels. For instance, finishing times for the GOOD accumulator in the grammatical control condition (green) are faster for inversion sentences than for comparative illusion sentences, so that more and faster "get it, correct" judgments are predicted for that condition, in line with the descriptive results.

Figure 5 shows highest density intervals for the slope parameters (differences between conditions) across constructions.[12] Across all constructions, the estimated mean finishing times of the accumulators are slower than the empirically observed judgment times, which follows from the model assumptions: in trials where a given response is *not* observed, it is assumed that the associated accumulator was slower than the observed response time.

[12]The slope estimates for ungrammatical agreement attraction sentences, ungrammatical depth charge sentences, and ungrammatical NPI sentences are notably wider than for the other constructions, given that these sentence types were only used for one half of Group 1.

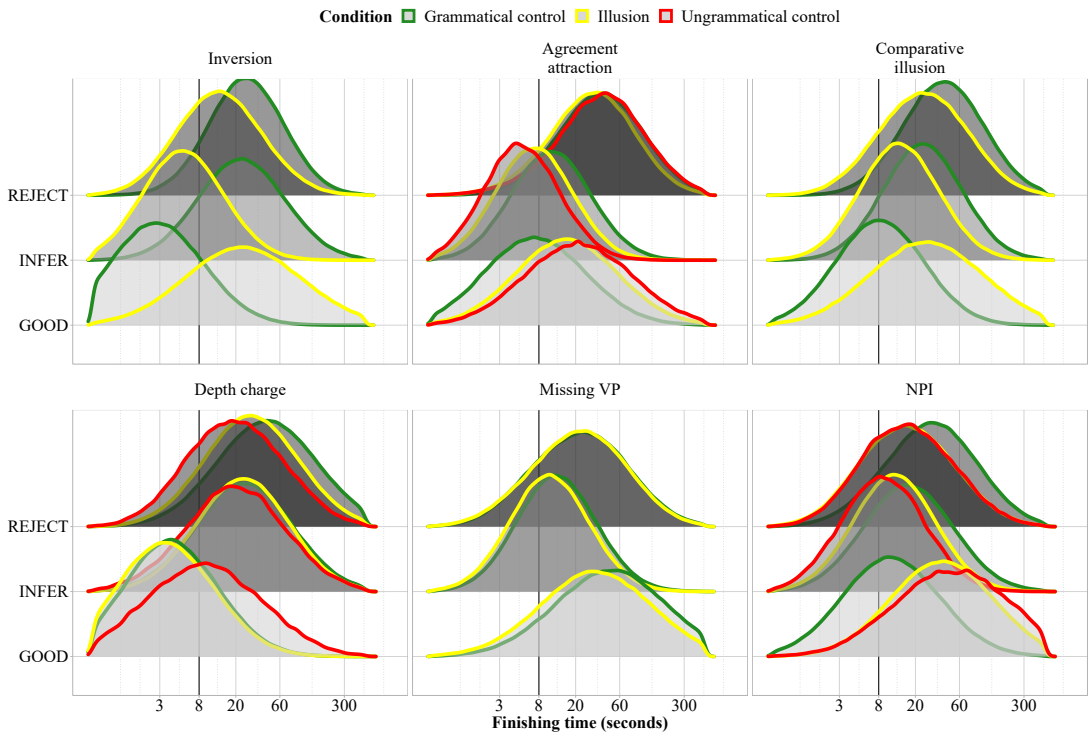


Figure 4: Posterior predictive distributions of finishing times (250 samples) of the three accumulators across constructions and conditions. Faster finishing times correspond to a higher expected number of responses of the respective type. Finishing times more than ± 2.5 SD away from the log mean have been removed. Reference line added at 8 seconds. Note that the x-axis is log-scaled.

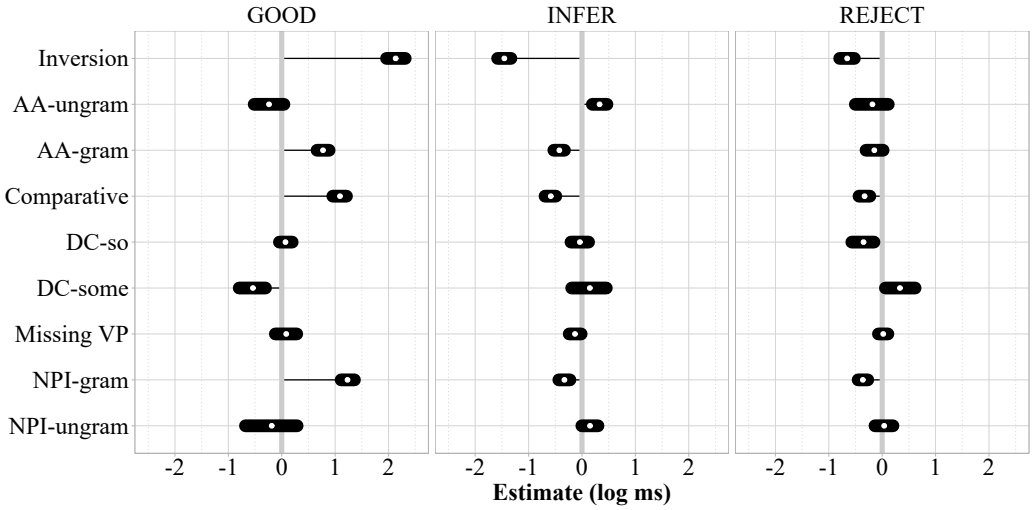


Figure 5: 95% highest density intervals of slope parameters across all constructions on the log ms scale. Across all constructions, the slope is the difference between the illusion condition and the indicated control condition(s). Positive estimates correspond to a slowdown of the accumulator, that is, fewer responses of the associated category in the illusion condition. Reference line added at 0 log ms.

In general, the parameter estimates and predicted finishing times align well with the descriptive results. Figure 6 shows the distribution of accumulator finishing times for the illusion conditions only

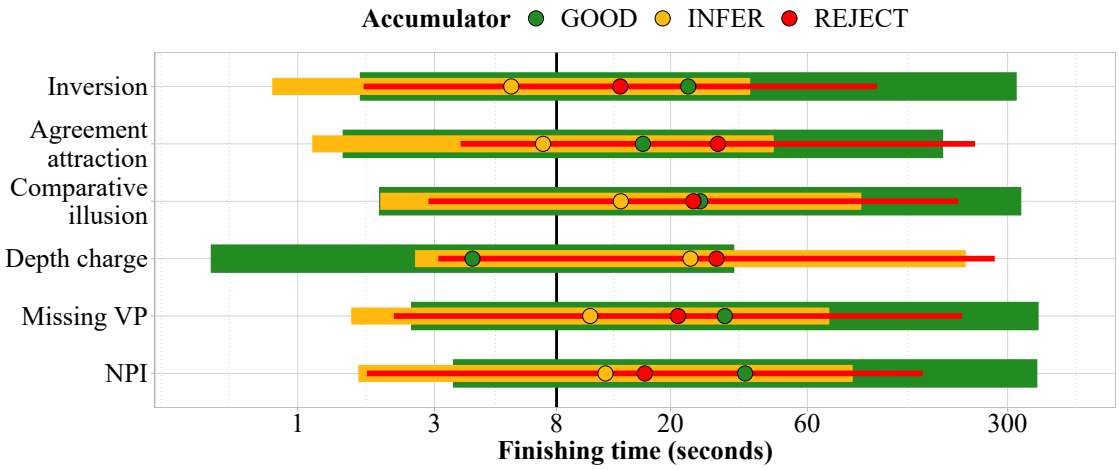


Figure 6: Mean ± 2 SD of predicted accumulator finishing times (250 samples) for the illusion condition across constructions. Faster finishing times correspond to a higher expected number of responses of the respective type. Finishing times more than ± 2.5 SD away from the log mean have been removed. Reference line added at 8 seconds. Note that the x-axis is log-scaled.

across constructions. The pattern confirms that the depth charge illusion behaves differently from the other constructions, as it shows a considerable speed advantage for the GOOD accumulator over the other accumulators, leading to many fast "get it, correct" judgments. As seen in Figure 5, the inversion construction shows the largest difference in judgments between the illusion and control conditions, followed by the comparative illusion and the NPI illusion. The agreement attraction effect is somewhat smaller, while for depth charge sentences and missing VP sentences the illusion and grammatical control conditions are almost indistinguishable, though the REJECT accumulator does show some graded sensitivity to the depth charge manipulation.

Most constructions show active trade-offs between the accumulators at the population level: the inversion manipulation slows down GOOD while simultaneously speeding up INFER and REJECT, as do the comparative illusion manipulation, the agreement attraction manipulation, and the NPI illusion manipulation. The presence of these active trade-offs is informed by the latency data: based only on the responses, it would be unclear if one process dominates the response pattern because it is strengthened by a manipulation, or because the competing processes are weakened.

REJECT is the slowest accumulator on average, being [11 s, 26 s] (95% highest density interval) slower than GOOD across all constructions. Across illusion and control sentences, INFER accumulates evidence more slowly than GOOD by [0.2 s, 9 s] on average. In terms of variability, the opposite picture emerges: REJECT shows the lowest amount of variability in finishing times at [11 s, 12 s], followed by INFER [14 s, 15 s], and finally GOOD [20 s, 23 s] with the highest amount of variability. The GOOD accumulator being fastest but less consistent in its speed of evidence accumulation than the other accumulators is in line with the basic assumptions of the good-enough processing framework, which assumes that readers create imperfect, "quick and dirty" sentence representations in some proportion of trials.

Increasing sentence length in characters by one standard deviation slows down REJECT by [0.2 s, 4.5 s], and GOOD by [3.8 s, 10.1 s], while the effect on INFER crosses zero: [−0.9 s, 2.4 s]. This pattern is somewhat unexpected, given that longer sentences should, in principle, offer more opportunities for errors and "repairs", but recall that there is a passive trade-off between the accumulators: REJECT and GOOD being slowed down in longer sentences results in more trials in which INFER wins, thus predicting more "get it, incorrect" judgments compared to shorter sentences.

In Group 3, receiving the logic task prior to the sentence judgment task as opposed to the other way around also affects the speed of the accumulators: doing the logic task first speeds up REJECT by [−24.2 s, −3.2 s] and GOOD by [−7.2 s, −3.9 s]. The order effect on INFER crosses zero: [−6.5 s, 2.8 s]. Due to trade-off between the accumulators, this means that there were fewer "get it, incorrect"

answers but more "don't get it" and "get it, correct" answers when the logic task was completed first. The observed pattern is in line with participants being mentally exhausted after completing the logic task, and preferring to either completely reject sentences or to uncritically accept them rather than engaging in effortful rational inference processes.

### 3.3.2 Graphical posterior predictive checks

Figure 7 shows predicted versus observed response proportions across constructions and conditions. Figure 8 shows predicted versus observed judgment times across constructions, conditions, and responses.

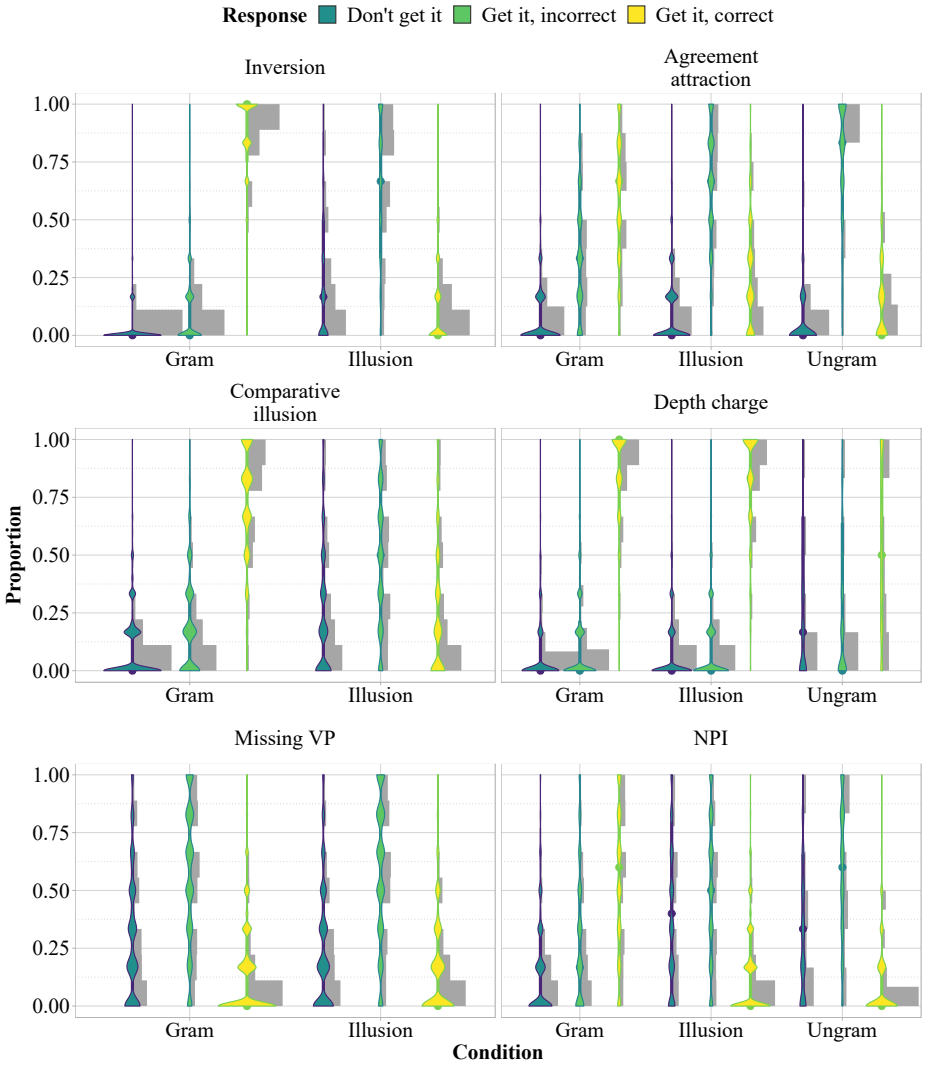


Figure 7: Predicted (colored) versus observed (gray) response proportions across constructions and conditions. Proportions were computed by participant. The visible stratification is due to between-participant variability, which is accurately captured by the model.

The response proportions predicted by the model are in good qualitative and quantitative alignment with the observed data, even at the level of individual participants, as shown by the stratification in Figure 7. The predicted judgment times also reproduce the qualitative patterns. The quantitative fit is also adequate, though predicted judgment times for "get it, correct" judgments tend to be somewhat longer than those seen in the data for some of the ungrammatical and illusion conditions.

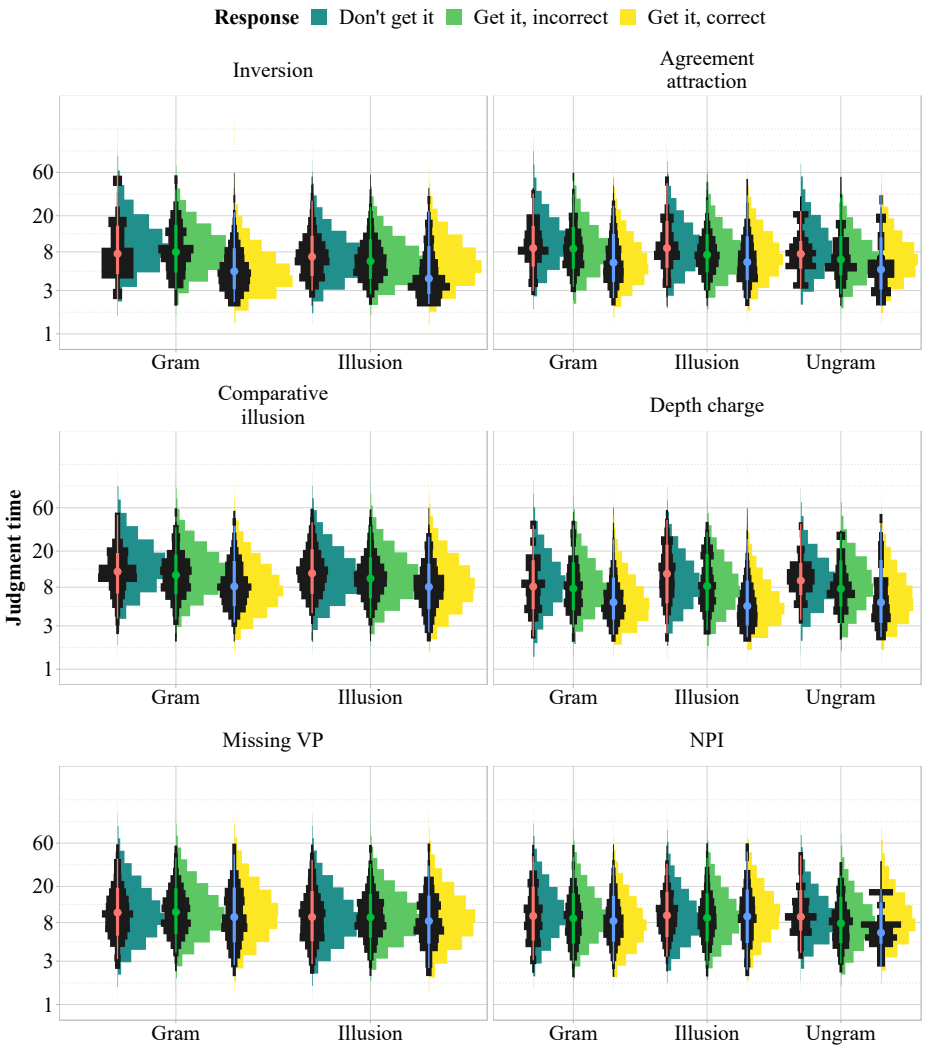


Figure 8: Predicted (histograms) versus observed (eye plots) judgment time distributions across constructions, conditions, and responses.

### 3.3.3 Individual-level trade-offs

Figure 9 shows the distribution of correlation estimates between subject-level random effects across all constructions by accumulator pair. I limit the discussion to slope parameters, that is, to differences between conditions. Active trade-offs predict negative correlations between the slopes for each pair of accumulators. By contrast, correlations between slopes within the *same* accumulator (SELF) should be positive, assuming that one and the same subject will react similarly to the different illusions. Figure 10 shows correlations for which the probability of direction is above 0.95, that is, for which the correlation estimate is mostly positive or negative. Note that this criterion does not correspond to a frequentist test of significance, but merely singles out effects for which the parameter estimates mostly point in a given direction. Due to the large number of parameters (9 slopes $\times$ 3 accumulators), the results should be interpreted with some caution.

The correlations of the subject-level random slopes show some active trade-offs within constructions, but crucially also across constructions. As Figure 9 shows, well-evidenced correlations tend to go in the expected direction for GOOD-INFER and SELF correlations, but overall the data are mostly inconclusive. As can be seen from Figure 10, participants who show larger-than-average effects of the inversion illusion on the GOOD accumulator tend to show smaller-than-average effects on the

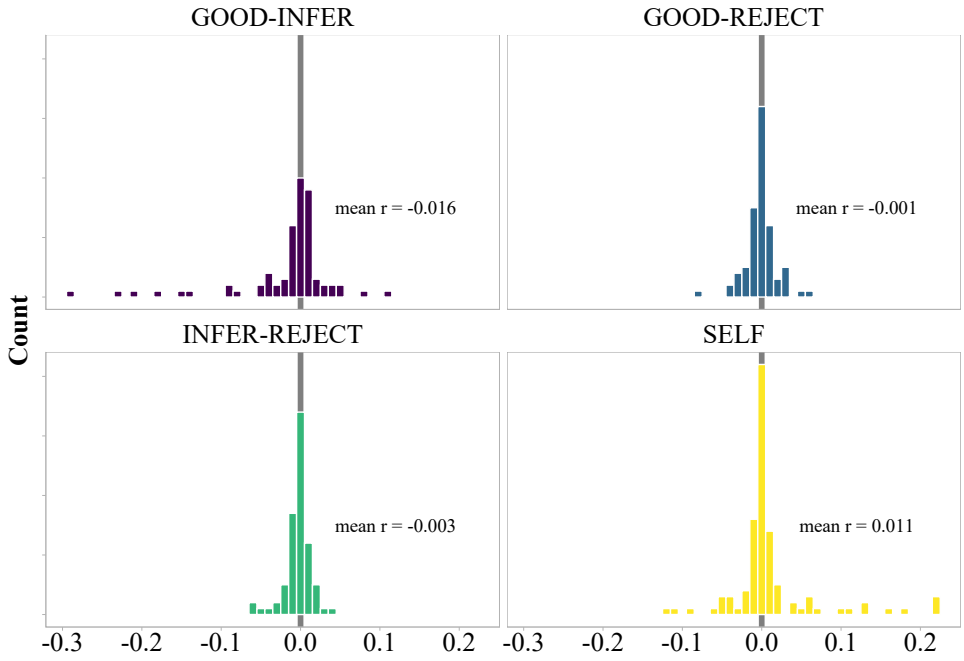


Figure 9: Histogram of subject-level random-effects correlations across all constructions by accumulator pair, weighted by probability of direction.

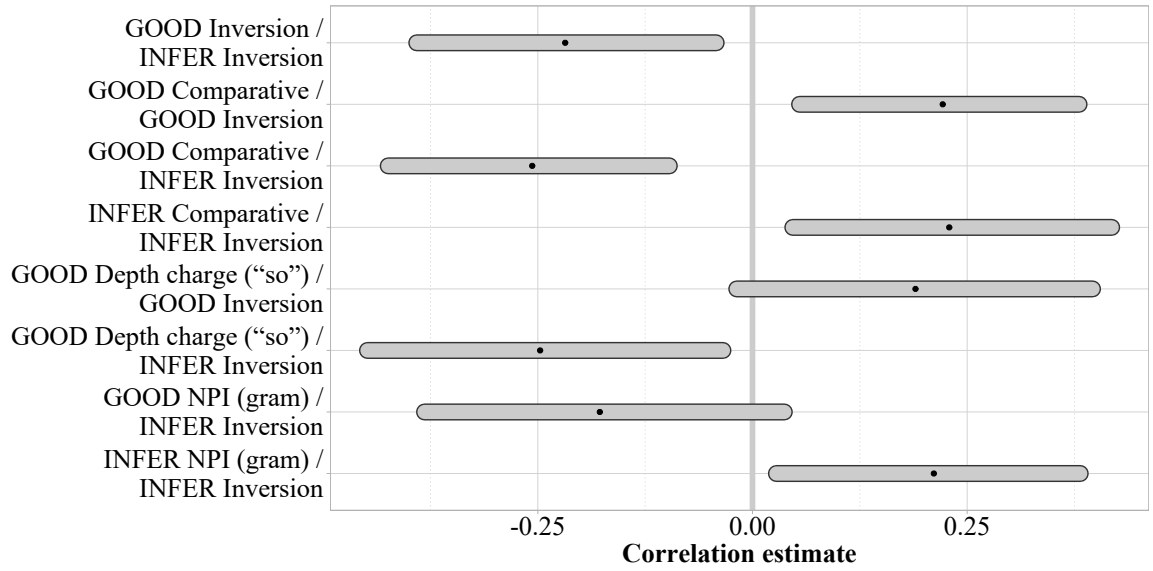


Figure 10: Correlation estimates of subject-level random effects and associated 95% highest density intervals. Only correlations with probability of direction $> 0.95$ for slope parameters (differences between conditions) are shown.

INFER accumulator for the same manipulation. This negative correlation suggests that participants who tend to do good-enough processing in the inversion construction ("slackers") are less likely to engage in rational inference. Such a trade-off is expected under the assumption that good-enough processing and rational inference serve different, conflicting goals: to conserve mental energy or to expend additional energy to infer meaning. The pattern of trade-offs holds across all correlations shown in Figure 10: larger participant-level effects on the GOOD accumulator correlate with smaller participant-level effects on the INFER accumulator. Importantly, all correlation estimates that reached

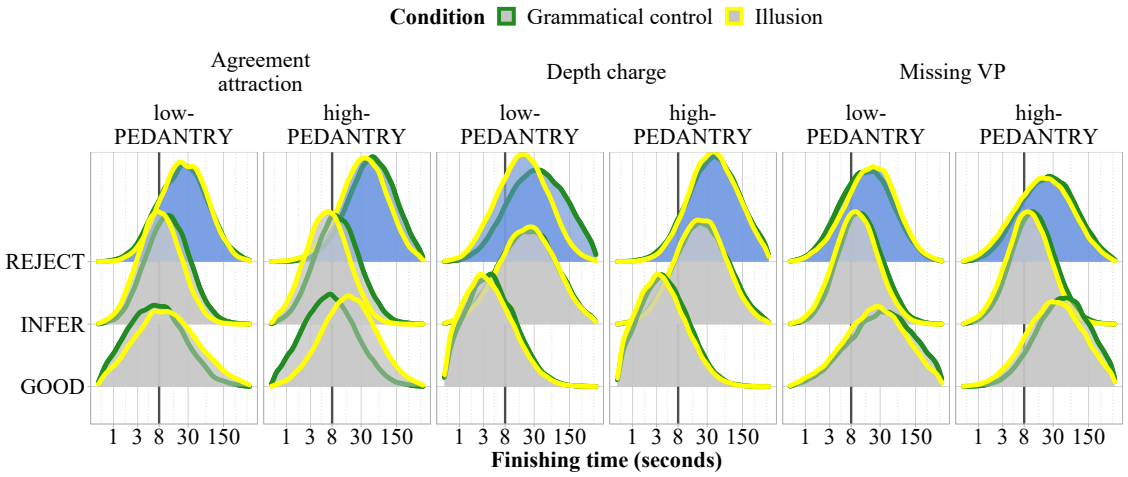


Figure 11: Posterior predictive distributions of finishing times (250 samples) of the three accumulators across constructions and conditions, by PEDANTRY score. Interactions with probability of direction > 0.95 are highlighted in blue.

the 0.95 probability-of-direction cutoff include the inversion construction, likely because the average effects in this construction are the largest. The inversion construction is thus a good candidate for a stable indicator of individual differences in illusion processing: an individual's processing preferences for inversion sentences can be used to predict the same individual's processing preferences for comparative illusion sentences, depth charge sentences, and NPI illusion sentences. However, due to the large number of correlations tested, future confirmatory studies should aim to test these correlations more rigorously.

### 3.3.4 Individual differences measures

In terms of interactions with the individual differences measures, the main question of interest is whether the difference between the illusion and control conditions for a particular constructions varies with a participant's self-reported pedantry and interpretational charity, and/or according to their logical reasoning ability. Due to the large number of possible interactions across constructions and accumulators, I will report only the results for constructions in which at least one interaction parameter reached probability of direction $> 0.95$. Across all predictors, I plot the finishing time distributions of the three evidence accumulators for the 20 highest- and lowest-scoring participants against each other.

**Pedantry** Figure 11 shows interactions with the PEDANTRY predictor, which is thought to reflect a participant's motivation to rigorously apply grammatical rules. PEDANTRY effects are seen in the agreement attraction, depth charge, and missing VP constructions. Contrary to my prediction, the interactions mainly affect the REJECT accumulator rather than the GOOD accumulator. For agreement attraction and missing VP sentences, high-PEDANTRY participants show faster finishing times for the REJECT accumulator compared to the control condition, that is, more "don't get it" responses. This pattern is still broadly in line with the assumption that PEDANTRY captures the strict application of grammatical rules in ungrammatical illusion sentences: pedantic individuals tend to give "don't get it" responses if the grammar does not license an interpretation, as opposed to ignoring errors or drawing inferences beyond the literal input. For depth charge sentences, however, the pattern is reversed: low-PEDANTRY participants tend to respond "I don't get it" more often in the illusion condition. I discuss this surprising finding below.

**Charity** Figure 12 shows interactions with the CHARITY predictor, which is thought to reflect a participant's default assumption that sentences are formally correct and sensible. Interactions with CHARITY are seen for inversion, agreement attraction, and missing VP sentences. As a general

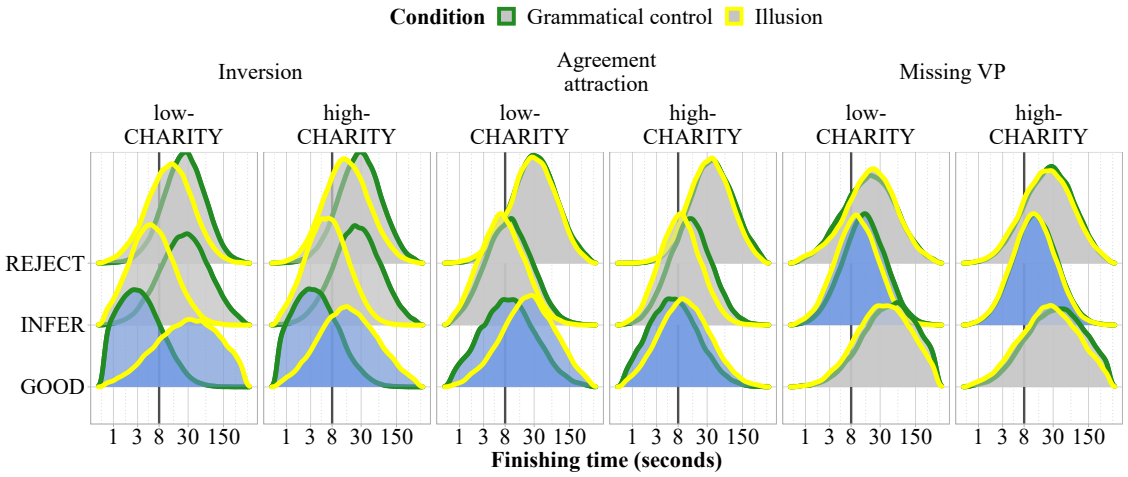


Figure 12: Posterior predictive distributions of finishing times (250 samples) of the three accumulators across constructions and conditions, by CHARITY score. Interactions with probability of direction > 0.95 are highlighted in blue

pattern, high-CHARITY participants tend to distinguish less between illusion and control sentences for these constructions, which is plausible given the interpretation of the predictor. However, there are differences between the constructions with regard to which accumulator is affected: for inversion illusion and agreement attraction sentences, high CHARITY leads to more "get it, correct" judgments (good-enough processing), while for missing VP sentences, *low* CHARITY leads to more "get it, incorrect" judgments (rational inference) for illusion compared to control sentences; this suggests that conscious rational inference in missing VP sentences may require *less* charitable assumptions about sentences usually being correct and sensible.

Moving on to the predictors derived from the syllogistic reasoning task, recall that there are three individual-differences measures: the main effect of logical validity (LOGIC), the main effect of believability (BELIEF), and the validity $\times$ believability interaction, which is commonly interpreted to signal a conflict between rule-based reasoning and intuition (CONFLICT). No effects reached probability of direction $> 0.95$ for BELIEF, so I will focus on the other two predictors.

**Logic** Figure 13 shows effects of the LOGIC predictor for inversion sentences, depth charge sentences, missing VP sentences, and NPI sentences. Similarly to the PEDANTRY predictor, the effect of LOGIC is such that high-LOGIC participants distinguish more strongly between grammatical and illusion sentences for these constructions, which mainly affects the GOOD accumulator. Depth charge sentences are unique in that LOGIC mainly affects "don't get it" responses: high-LOGIC participants tend to reject illusion sentences of this type more often than control sentences. Overall, the results are in line with the assumption that individuals with strong logical abilities are less likely to process sentences superficially, and are thus more likely to spot grammatical errors, and/or, in the case of depth charge sentences, fail to understand the sentence when the compositional meaning is nonsensical.

**Logic-belief conflict** The CONFLICT predictor affects comparative illusion, depth charge, and NPI sentences, as shown in Figure 14. For comparative illusion and depth charge sentences, low-CONFLICT participants distinguish more strongly between illusion and control sentences; this difference is visible in rational inferences ("get it, incorrect") for comparative illusion sentences, but in rejections ("don't get it") for depth charge sentences. This pattern is broadly in line with the assumption that CONFLICT measures the amount to which intuition interferes with rule-based processing. The pattern in NPI sentences, however, is unexpected: high-CONFLICT participants show a slightly larger difference between conditions on the INFER accumulator than low-CONFLICT participants.

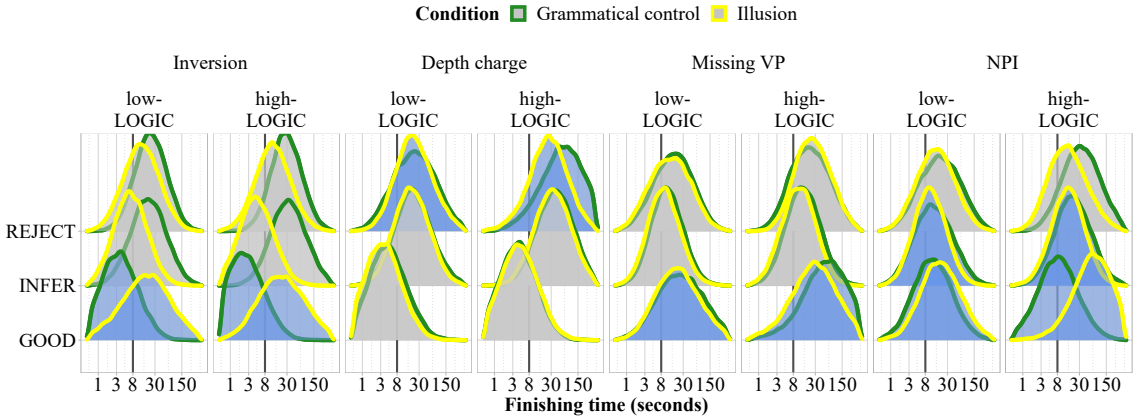


Figure 13: Posterior predictive distributions of finishing times (250 samples) of the three accumulators across constructions and conditions, by LOGIC score. Interactions with probability of direction > 0.95 are highlighted in blue.

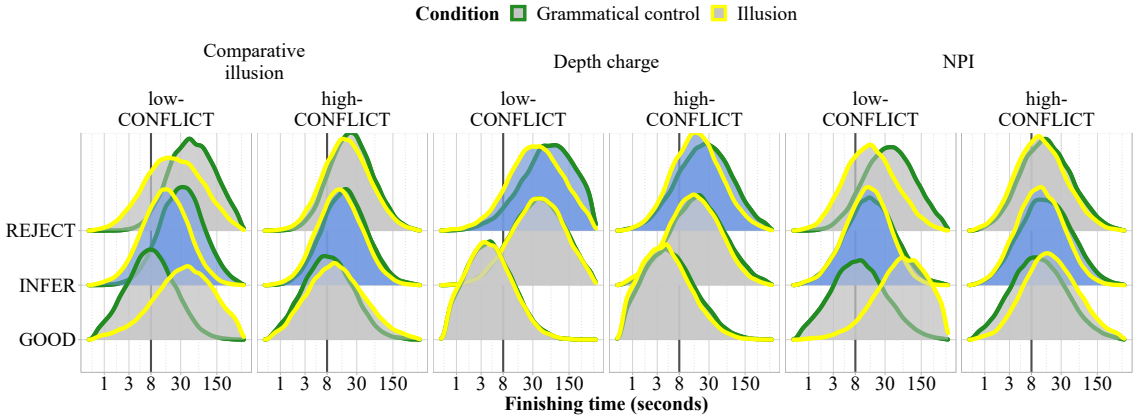


Figure 14: Posterior predictive distributions of finishing times (250 samples) of the three accumulators across constructions and conditions, by CONFLICT score. Interactions with probability of direction > 0.95 are highlighted in blue.

**Summary** To summarize, pedantic individuals and individuals who consistently apply logical rules in the face of distracting believability information tend to experience fewer acceptability illusions for some constructions, presumably because they tend to stick to closely to prescriptive grammar when evaluating meaning. By contrast, individuals who make charitable assumptions about sentence interpretability and formal correctness appear to tend more towards good-enough processing, at least for some illusions.

### 3.3.5 Model comparisons via cross-validation

Despite the plausible parameter estimates and promising results of graphical posterior predictive checking, there is a question of whether the proposed race model generalizes well to unseen data, or whether it is overfitted to the current data set. One well-known way of approaching this question is through cross-validation: the model is fitted repeatedly to only a subset of the data, and the remaining data points are used as novel data that the model's predictive fit is evaluated against (see Vehtari and Ojanen, 2012; Yarkoni and Westfall, 2017 for introductions). Since no model can realistically be expected to give a perfect fit (Box, 1979), it is usually the *relative* predictive performance of different models that is of interest. Here, I focus on three questions:

1. What is the relative contribution of each accumulator to the race model's predictive fit?
2. What is the contribution of the individual differences measures to the race model's predictive fit?
3. How does the predictive fit of the race model compare to alternative response models?

Question 1 can be answered by completely removing all slope parameters related to the experimental manipulations from one of the accumulators. This amounts to the assumption that the respective accumulator is "passive" in the sense that it is unaffected by the properties of the sentence, and that the associated response is only produced differentially across conditions as a consequence of the *other* accumulators being affected by the manipulations. The response associated with the unaffected accumulator can thus be seen as a default response in the simplified model (Paape and Zimmermann, 2020; Nicenboim and Vasishth, 2018). Question 2 can be approached in a similar way, namely by removing all slope parameters related to the individual differences measures from the model. Question 3 can be answered by implementing models that do *not* assume a race mechanism as the process generating the observed responses and latencies. Model comparisons are carried out using approximate leave-one-out cross-validation via Pareto-smoothed importance sampling, as implemented in the loo R package (Vehtari et al., 2015, 2019). The measure of interest is the expected log pointwise predictive density ($\widehat{elpd}$) of each model, which quantifies how likely the unseen data points are under the model.

In principle, the model space for the problem at hand is $\mathscr{M}$-open (Piironen and Vehtari, 2017): there are infinitely many possible models that could account for the data. Considering different race models first, in order to keep the number of models to be compared manageable, I consider only the three models with one "passive" accumulator each (GOOD, INFER, or REJECT), as well as three models in which either both sets of individual-differences slopes (questionnaire-based and logic-task-based) or only one of them are removed from *all three* accumulators. In terms of models that do not assume a race between accumulators as the data-generating process, I will limit myself to two additional models: A multinomial processing tree (MPT) model (see Erdfelder et al., 2009; Batchelder and Riefer, 1999 for reviews) that assumes "noisy channel" mechanisms as the sole source of the different response types, and a "theory-free" multivariate model that can be fitted straightforwardly via the brms package (Bürkner, 2017).

The MPT model assumes two types of possible mental edits to the sentence representation that occur serially and stochastically: edits that occur *before* a grammatical violation arrives and edits that occur *in response* to such violations. The concept of surprisal (Levy, 2008a; Hale, 2001) posits that a word's processing difficulty is proportional to its predictability in a given context. Levy (2011) presented a model in which surprisal can be affected by mental edits to the sentence representation before the critical word arrives. For instance, in an agreement attraction sentence like *The key to the cabinets were rusty*, if the preamble is mentally edited into *The keys to the cabinets ...*, the ungrammatical word *were* is now much less surprising than in a setting where the preamble is always represented veridically (Yadav et al., 2023). Similarly, in a missing VP sentence like *The apartment that the maid who the cleaning service sent over was well-decorated*, if the preamble is edited to contain only two as opposed to three subject nouns, *not* encountering a third verb would not be surprising to the reader (Futrell et al., 2020; Hahn et al., 2022). If such context edits happen pre-perceptually or at least before perceptual information is integrated at higher levels (Huang and Staub, 2021b,a), this would explain why illusion sentences can be perceived as being formally correct. The surprisal-plus-mental edits model thus offers an alternative explanation for the "get it, correct" judgments seen in the present study.

The MPT model assumes that the process DISTORTION occurs first with probability $p_1$, potentially followed by the process REPAIR with probability $p_2$. The process probabilities are estimated from the data. Additionally, the following assumptions are encoded in the model:

1. Distortions are equally likely to occur in grammatical, ungrammatical and illusion sentences, but can differ between constructions. Distortions lead to slowdowns and rejections in grammatical sentences but to speedups and acceptance ("get it, correct") in ungrammatical and illusion sentences. The amount of slowdown/speedup D differs between constructions and is estimated from the data.
2. Repairs can also occur across all conditions and incur a cost R that is estimated from the data. Both repair probability and repair cost differ between constructions. Illusion sentences are more likely to be repaired than ungrammatical sentences. Grammatical sentences can only be "repaired" if they have first been distorted, in which case the slowdown will be D+R with probability $p_1 \cdot p_2$.

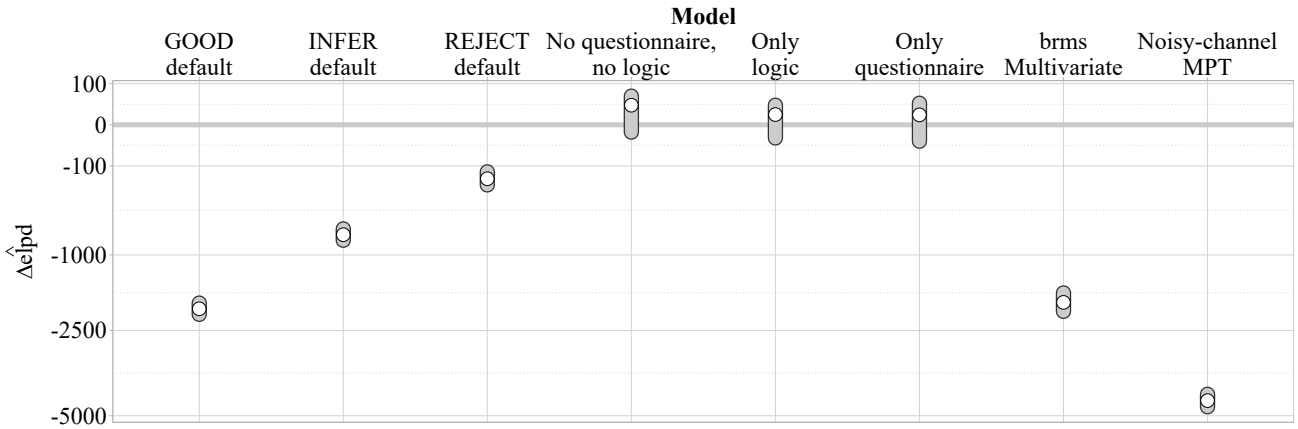


Figure 15: Estimated differences in log pointwise predictive density between the full model and alternative models. A positive difference means that the alternative model has a better predictive fit than the full model. Error bars show 95% confidence intervals. Note that the y-axis is square-root-scaled.

3. Sentence length and individual differences can affect both distortion and repair rates. Distortion and repair costs differ between individuals. Judgment times are lognormally distributed, with an added SHIFT (non-decision time) parameter.

The current implementation of the noisy-channel proposal as an MPT model is considerably less detailed than the ones provided by Hahn et al. (2022) and Yadav et al. (2023), as it neither takes into account the size of the required edits, nor the corpus frequencies of the different constructions. On the other hand, it is, to my knowledge, the first implementation that takes into account both distortions and repairs across grammatical, ungrammatical, and illusion sentences.

The "theory-free" model was implemented using the brms syntax for multivariate models (Bürkner, 2024). The brms model is not a cognitive process model but purely a statistical one that treats the experimental responses as having a categorical distribution with a $\theta$ parameter for each response ("get it, correct", "get it, incorrect", "don't get it") that is affected by the experimental manipulations and individual differences. Response times are assumed to have a shifted lognormal distribution with mean $\mu$ and standard deviation $\sigma$, where $\mu$ is affected by the experimental manipulations and individual differences. Crucially, the model does not assume any connection between responses and their associated latencies, apart from possible correlations between random effects for a given subject and item.

Figure 15 displays the difference in expected log pointwise predictive density for each alternative model against the full race model. Differences in $\widehat{elpd}$ have an associated uncertainty, which can be quantified by computing their standard errors and the resulting confidence intervals (Vehtari et al., 2017). As the figure shows, turning any of the accumulators in the race model into a "passive" default response sharply reduces predictive fit. The reduction in predictive fit is strongest for the GOOD accumulator, suggesting that the effects of the manipulations on this accumulator are the most important, while removing the slopes on the REJECT accumulator yields the smallest predictive loss. Nevertheless, the predictive fit of the full model is still higher than for any of the three simpler "default" models.

The two non-race models both have decisively lower predictive fit than the full race model. The multivariate brms model is on par with the worst-fitting race model, but the noisy-channel MPT model does not appear to predict the data well. Of course, this only applies to this specific implementation of the noisy channel approach. One should keep in mind that the MPT model has much fewer parameters than the race model, as well as more constraints: repairs have a strictly positive cost, while distortions are assumed to result in speedups or slowdowns depending on the condition, which are constrained to be of equal magnitude. Hard constraints are important for theory building and testing, especially if the empirical fit of the constrained model turns out to be suboptimal (Roberts and Pashler, 2000). Nevertheless, a different implementation of the noisy channel proposal may outperform the race

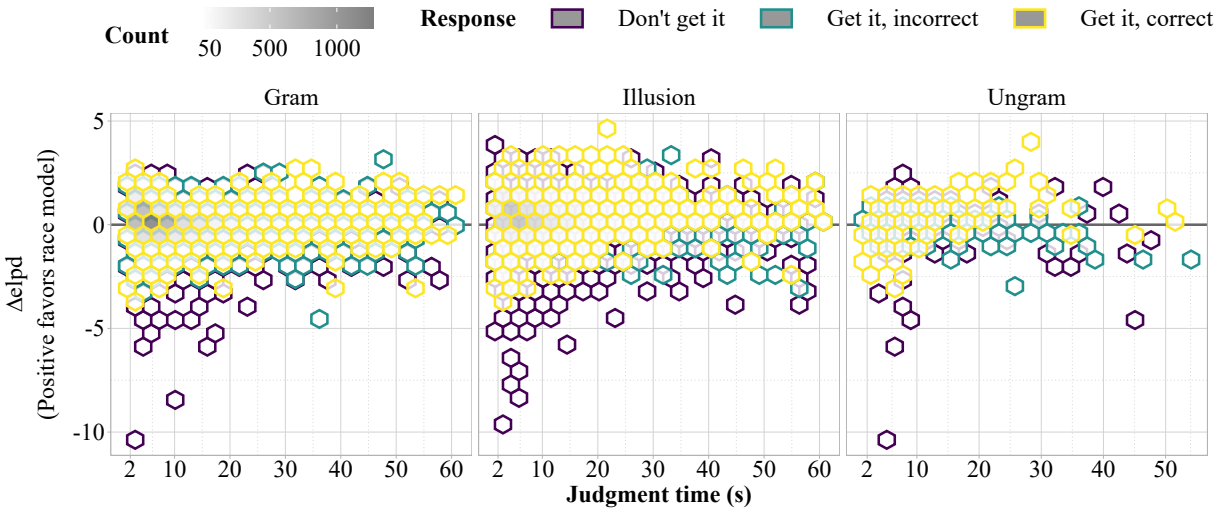


Figure 16: Difference in log pointwise predictive density between the full race model and the noisy-channel MPT model by condition, response, and judgment time.

model, but a comparison of different noisy channel models is beyond the scope of the current paper. That being said, it is potentially informative to compare the predictive performance of the race and MPT models for individual observations, as the model with overall worse predictive performance may nevertheless show an advantage for specific aspects of the data (Nicenboim and Vasishth, 2018). Figure 16 shows differences in $\widehat{elpd}$ between the two models by condition, observed response, and judgment time.

As the figure shows, the noisy channel MPT model actually tends to have better predictive capability than the race model for fast rejections ("don't get it") in the grammatical and illusion conditions, presumably because rejections don't incur a repair cost. In the race model, fast rejections are unlikely due to REJECT being the slowest accumulator on average. The noisy channel model also tends to have an advantage for slow "get it, incorrect" responses, which is expected given the assumption of a repair cost. By contrast, the race model appears to get most of its predictive advantage from the "get it, correct" responses, especially in the illusion conditions, and especially from the many responses of this type with relatively fast judgment times. Overall, the results are sensible given each model's assumptions, and show that the noisy channel model does well at predicting certain aspects of the data.

The removal of the individual differences slopes does *not* result in a significant drop in predictive fit compared to the full race model. While the confidence intervals of the differences in expected log pointwise predictive density against the full model all cross zero, the pattern suggests that the individual differences parameters do not improve the generalizability of the model, but may potentially reduce it instead. However, the model comparison is inconclusive, and the results rest on the assumption that the effects of the individual differences predictors are linear on the log scale.

Exploring potential non-linear effects of the predictors on the accumulators in detail is also beyond the scope of this paper. Nevertheless, there is reason to believe that the combination of linear predictors and additional by-participant adjustments to the accumulator finishing times can capture non-linear patterns. Figure 17 shows LOESS plots (e.g. Jacoby, 2000) of the observed and model-predicted relationships between LOGIC score and response proportions across constructions, and reveals that the model captures captures the empirical patterns reasonably well.

### 3.3.6 Discussion

Joint modeling of the reaction time and response data with a lognormal race model and comparing the predictive fit of the model against alternative models yielded a number of key results:

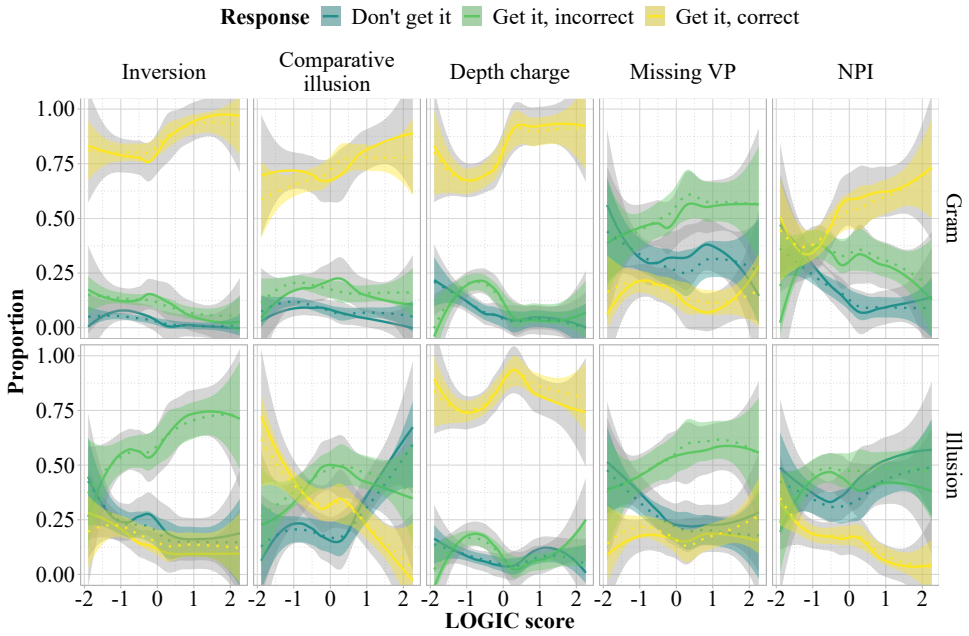


Figure 17: LOESS curves of response proportions by LOGIC score, construction, and condition. Solid lines are based on the experimental data; dotted lines are based on model predictions.

- At the population level, rational inference, good-enough processing, and outright rejection actively trade off against each other: effects on the GOOD accumulator tend to be accompanied by effects of the opposite sign on the INFER and REJECT accumulators across constructions.
- At the individual level, there is also some indication of active trade-offs between good-enough processing and rational inference across constructions, as well as positive correlations across constructions for the *same* accumulator within a given participant, suggesting some stable tendencies towards processing in the "inferencer", "slacker" or "pedant" modes, though the results are not conclusive.
- At the individual level, participants' interpretational charity, linguistic pedantry, and logical ability affect the accumulators' finishing times in (mostly) theoretically plausible ways across constructions. However, given the cross-validation results, it is not clear how well these effects generalize.
- In terms of predictive ability, the lognormal race model outperforms both a "theory-free" multivariate model and an implementation of the noisy channel proposal that includes both pre- and postperceptual context edits.

The observation of active trade-offs between the accumulators makes sense under the assumption that each accumulator serves a different goal: the GOOD mechanism aims to avoid cognitive effort by using superficial, heuristic processing, the INFER mechanism makes use of linguistic experience to "repair" errors and infer intended meanings, and the REJECT accumulator rigorously applies grammatical rules. Some parts of sentence processing, such as word identification, serve all of these goals simultaneously, while others are exclusive to a single goal. The implication of the observed trade-offs is that good-enough processing and rational inference compete for cognitive resources, which goes against recent proposals that they are essentially variations of the same theoretical approach to non-literal interpretations (e.g., Dempsey et al., 2023; Goldberg and Ferreira, 2022).

Despite the generally plausible pattern of results with regard to the individual differences measures, there were also two unexpected findings: low-PEDANTRY participants rejected depth charge illusion sentences (*No test is too difficult to fail*) more often than high-PEDANTRY participants, and high-CONFLICT participants made more rational inferences for NPI illusion sentences than low-CONFLICT participants.

In this context, recall that PEDANTRY not only includes disposition but also motivation: pedantic individuals invest energy into finding errors. It has been suggested that depth charge sentences are so complex that it may be impossible to parse them correctly even for highly motivated participants (Wason and Reich, 1979). Speculatively, low-PEDANTRY participants, who are less motivated, may reject depth charge sentences more often in the illusion condition because they are semantically more complex than the sentences in the control condition (*No test is so difficult that she would fail it*). Regarding the unexpected interaction between CONFLICT and rational inference in NPI sentences, it may be related to the fact that both NPI licensing and syllogistic reasoning have been linked to extralinguistic pragmatic processing (Xiang et al., 2013; Tessler et al., 2022). Speculatively, high-CONFLICT individuals may be more sensitive to the pragmatic licensing conditions on NPIs, and thus better able to draw inferences beyond the literal input. Alternatively, due to the large number of parameters in the model, the unexpected results may simply be false positives, especially in view of the inconclusive cross-validation results. This possibility can be addressed in future replication studies.

# 4 General discussion

The main aim of the current study was to investigate trade-offs between conscious rational inference mechanisms and good-enough processing. To study these trade-offs, I collected data from a large participant sample on different sentence types that are known to cause linguistic illusions: sentences with argument inversions (*The mother gave the candle the daughter*), agreement attraction sentences, depth charge sentences (*No test is too difficult to fail*), comparative illusion sentences, missing VP sentences, and NPI illusion sentences. The current study is, to my knowledge, the only one to date that tested all of these constructions within the same sample of participants, that used combined judgments of grammaticality and interpretability, and that additionally collected individual differences measures. The study of Langsford et al. (2019) was comparable in terms of sample size and empirical breadth, but didn't cover individual differences. In addition, the data from the current study are complemented by a computational cognitive model that jointly accounts for judgments and judgment latencies, and that can dissociate the underlying latent processes.

There is a wealth of work that has explored specific illusions in depth by using targeted manipulations of sentence structure and lexical content, as well as grammatical differences between languages (e.g., O'Connor, 2015; Wellwood et al., 2018; Bhatia and Dillon, 2022; Orth et al., 2021; Frank and Ernst, 2019). This kind of work is highly valuable, but introduces the risk of overfitting theories to particular constructions instead of aiming to develop a "theory of everything". Studies that compare different illusions and try to identify possible shared mechanisms, on the other hand, are comparatively rare (e.g., Langsford et al., 2019; Brehm et al., 2021; Parker and Phillips, 2016). The current study aimed to broaden the empirical picture by presenting six illusion types to the same participants, and by testing the predictions of two accounts that are general enough to cover a wide range of illusions.

As discussed in the introduction, the rational inference approach and the good-enough processing approach are different both in spirit and in their predictions. The rational inference approach in the version originally proposed by Levy (2008b) posits that readers can mentally correct errors in sentences they read, which can result in non-literal interpretations that are influenced by prior expectations about plausible meanings (e.g., Gibson et al., 2013; Cai et al., 2022). By contrast, good-enough processing posits that readers sometimes process sentences superficially, which can result in input information such as grammatical mismatches being ignored (e.g., Ferreira (2003)). Crucially, under both rational inference and good-enough processing, readers should be under the impression that they *understood* the meaning of the sentence. The predictions of the two accounts only start to diverge when the precise nature of the rational error-correction mechanism is taken into account: given an appropriate task, such as judging the formal correctness of sentences, it is likely that error corrections will rise to consciousness (Levy, 2008b; Ryskin et al., 2018). By contrast, an obvious corollary of the good-enough processing assumption is that sentence anomalies are sometimes missed when processing is shallow (Karimi and Ferreira, 2016; Sanford and Sturt, 2002; Christianson, 2008; Frazier and Clifton, 2015), even when readers/listeners are specifically instructed to monitor for errors or give acceptability ratings (Erickson and Mattson, 1981; Sanford et al., 2011; Paape et al., 2020).

## 4.1 Two-dimensional judgments and competing latent processes

In the current study, readers were asked to simultaneously judge the formal correctness and interpretability of linguistic illusion sentences and control sentences. The relevant linking assumption is that if processing is shallow, readers should not even notice that an "inversion" sentence like *The mother gave the candle the daughter* is implausible, and thus give "get it, correct" judgments. By contrast, if readers engage in rational error correction, and if error correction is conscious given the explicit task demands, they should give "get it, incorrect" judgments. Finally, if readers are pedantic about grammar, and assume that grammar is the only mediator between input and meaning, they may notice the error but refrain from trying to infer the (presumably) intended interpretation, responding with "I don't get it".

Mapping the available response types to mental mechanisms in a one-to-one fashion is, of course, a simplification. All human decision processes are subject to noise, and sentence judgments are no exception. Thus, a "get it, incorrect" judgment on a given trial cannot be unequivocally interpreted as evidence for conscious rational inference having taken place on that particular trial. What *can* be interpreted, however, are the differences between conditions within the illusion sentences, as well as the observed differences between illusions in terms of which mechanism "wins" most often. I believe that the empirical results vindicate the linking assumptions: For instance, inversion illusions largely showed "get it, incorrect" judgments, which is in line with the claim of Gibson et al. (2013) that these sentences are processed via rational inference. On the other hand, the depth charge illusion in sentences such as *No test is too difficult to fail*, which the majority of authors has attributed to good-enough processing (see Paape et al., 2020 for a review), showed a large preference for "get it, correct" judgments.

Across the six illusions tested, results showed a mixture of the three response types, with "get it, incorrect" judgments dominating in the majority of the illusion conditions, except in the depth charge construction. The overall picture is consistent with an advantage for rational inference under task demands that favor conscious error detection (and potentially correction), though it is notable that large amounts of good-enough processing still occur even when participants are tasked with finding errors. Given that the good-enough processing framework explicitly aims to offer a more "naturalistic" vision of language processing by taking into account the uncertainty and error-proneness of everyday communication (Ferreira et al., 2002), it is plausible to assume that embedding the illusion sentences in longer passages to make them less conspicuous and choosing a less formal task would shift the balance away from rational inference and towards good-enough processing, though this would require an explicit investigation.

The pattern seen in depth charge sentences suggests that this particular illusion may not be due to rational error correction, as recently proposed by Zhang et al. (2023b), but is either caused by superficial processing (e.g., Wason and Reich, 1979; Paape et al., 2020) or even by (partial) grammaticalization of the anomalous structure (Fortuin, 2014; Cook and Stevenson, 2010). Another interesting case is the missing VP illusion, where both the grammatical and ungrammatical conditions received high amounts of "incorrect" judgments. This pattern is consistent with the assumption that multiply center-embedded sentences overload the parser's memory (e.g., Gibson and Thomas, 1999), and that such structures, while theoretically well-formed, may not be part of the average reader's grammatical competence (e.g., Schlesinger, 1975).

## 4.2 The lognormal race model and the role of individual differences

In order to further dissociate the latent processes involved in the processing of linguistic illusions, I fitted the lognormal race model of Rouder et al. (2015) to the judgment and latency data.[13] The lognormal race model assumes that while the stimulus — in this case, the sentence — is processed, the different response options accumulate evidence in parallel. The first option that accumulates enough evidence to pass a threshold determines the observed response. Under this model, rational inference, good-enough processing, and outright rejection of the sentence are seen as parallel, competing processes within the same individual in each trial. Given this perspective, it is also important to ask to what extent underlying individual differences between participants contribute to different response patterns.

[13] For previous applications of race models to sentence processing, see e.g. Nicenboim and Vasishth (2018), Paape and Zimmermann (2020), Lissón et al. (2021), Logačev and Vasishth (2016).

The modeling results showed that there are active trade-offs between rational inference and good-enough processing within individuals, both within and across illusion constructions: participants who compute fast rational inferences tended to show less good-enough processing, especially for inversion sentences, and vice versa. There was no indication that rejection ("don't get it") traded off with the other response options at the individual level, but this null result may be due to insufficient data. There was, however, some indication that individuals who scored high on linguistic pedantry – based on the questionnaire responses – were more likely to reject agreement attraction and missing VP sentences, and that participants who scored high on interpretational charity were more likely to do good-enough processing of inversion and agreement attraction sentences. By contrast, individuals who scored high on logic and individuals who scored low on logic-belief conflict tended to distinguish *more* strongly between illusion and control conditions for inversion sentences, depth charge sentences, and NPI illusion sentences.

As highlighted in the introduction, a salient point of distinction between the rational inference and good-enough processing frameworks is the assumed goal of the comprehender: to reconstruct an intended meaning as optimally as possible, or to save cognitive effort by "cutting corners" during interpretation. In the context of a race model, the question is how and why "cutting corners" can occasionally result in *slower* evidence accumulation than rational inference, that is, how rational inference can ever "win" the race in such a model. As previously mentioned, it is an empirical fact that participants sometimes prefer to "stick to the rules" when presented with ill-formed sentences (e.g., Gibson et al., 2013; Ferreira, 2003; Bader and Meng, 2018; Meng and Bader, 2021; Stella and Engelhardt, 2022), which requires explanation under any model of non-literal processing. Besides individual disposition and motivation (Christianson et al., 2022) an obvious factor that may determine such behavior is how *accessible* a non-literal interpretation is given the specific input. Rational inference may be faster in cases where the error is more frequent and thus easier to reconstruct (Frazier and Clifton, 2015; Poppels and Levy, 2016). Conversely, the speed of heuristic processing may depend on the accessibility and reliability of the heuristic in question, as well as on the fit between the heuristic template and the stimulus (Bellur and Sundar, 2014; Chen and Chaiken, 1999). For instance, Ferreira (2003) has argued that a sentence like *The dog was bitten by the man* can be heuristically matched to a stored agent-verb-patient template learned from linguistic experience (dog-bite-man). By contrast, a complex missing VP sentence such as *The manuscript that the student who the catalog had confused was missing a page* presumably does not trigger immediate activation of such a template. In such cases, it's more likely that the good-enough stream "cobbles together" an analysis on the fly, perhaps using multiple partial parses (Kamide and Kukona, 2018), in hopes of arriving at an approximate meaning. This process may take longer than rational inference, which is informed by error probabilities, and longer than fully algorithmic processing, assuming that the reader has the appropriate linguistic capacity to parse the veridical structure. In short, while instances of supposedly "fast algorithmic" and "slow heuristic" processing require explanation via specific stimulus properties, their mere existence does not invalidate the race-based approach.

## 4.3 Integrating breadth- and depth-focused approaches

Overall, the current work has highlighted the value of adopting a breadth-focused approach when investigating linguistic illusions. However, it is clear that the more widely used "depth"-focused approach, in which a single sentence type is manipulated in different ways, continues to yield crucial insights into the cognitive mechanisms involved in acceptability illusions, especially with regard to the role of fine-grained grammatical constraints. The current work was not intended to uncover the precise processing steps that give rise to illusory acceptability, but to highlight the fact that broad frameworks exists into which more specific mechanisms can be integrated. For instance, it has been suggested that agreement attraction, missing VP, and NPI illusions can be explained by a cue-based retrieval mechanism that sometimes accesses incorrect elements in working memory (Wagers et al., 2009; Vasishth et al., 2008; Häussler and Bader, 2015). Cue-based retrieval can potentially be integrated with good-enough processing, either in the sense that memory representations can be faulty because not all aspects of the stimulus were processed or retained (Yadav et al., 2023), or in the sense that there is an additional error monitoring step after retrieval that may be omitted when processing is superficial.

In any case, a theory is needed that can explain the interplay between inferring likely sentence meanings and consciously noticing formal errors, including effects of task demands and individual differences between speakers. Such a theory will likely include aspects of both the rational inference

framework and the good-enough processing framework. For instance, both frameworks assume that the prior plausibility of a given interpretation plays a role, and it has been suggested that there is a correlation between the "naturalness" of an error ("Could have happened to me!") and its noticeability and/or subjective probability (Frazier and Clifton, 2015; Zhang et al., 2023b; Poppels and Levy, 2016). Furthermore, the rational inference account assumes that readers and listeners rationally adapt to the type and frequency of errors in their environment, and change their interpretation strategies accordingly (Ryskin et al., 2018). What is unclear, however, is how detailed the reader's mental model of the interlocutor needs to be: does the reader or listener engage in a full simulation of the speaker that reverse-engineers the mental processes behind an utterance (Pickering and Garrod, 2013), including likely errors (Frazier and Clifton, 2015; Poppels and Levy, 2016)? A full simulation would possibly be too effortful in realistic scenarios, where cognitive resources are finite (Pöppel, 2023), and requires a comprehender who is able to accurately simulate the errors of the interlocutor without introducing errors of their own in the process.

The data collected in the current study can serve as a benchmark for the predictions of yet-to-be-developed models that can make detailed predictions about error noticeability and rational inference, ideally based on fine-grained properties of sentences. In this context, it will likely be fruitful to take into account how state-of-the-art language and dialogue models like GPT and ChatGPT process illusion sentences, as they are purely data-driven: language models neither have an explicit error correction mechanism, nor do they have a reservoir of motivation and/or attention that can be depleted, unlike humans. There is already some data showing how language models differ (or not) from humans in the domain of linguistic illusions (e.g., Dentella et al., 2023; Cai et al., 2023; Shin et al., 2023; Paape, 2023; Zhang et al., 2023a), which can inform future investigations into the mechanisms that may be unique to human sentence processing.

# 5 Conclusion

The current work aimed to disentangle the two most encompassing psycholinguistic accounts of non-literal sentence processing: the rational inference account, which assumes that speakers reconstruct the intended form of an utterance by reasoning about plausible sentence meanings and transmission errors, and the good-enough processing account, which assumes that speakers sometimes selectively ignore information in the stimulus. I have shown that conscious rational inference about intended meanings on the one hand and superficial, good-enough processing on the other can be understood and modeled as competing latent processes within one and the same individual. These two processes also compete with a purely grammar-driven, "pedantic" process that causes ill-formed sentences to be rejected as uninterpretable. All three processes actively trade off with each other while participants read sentences and form judgments about their formal correctness and interpretability. Depending on individual traits, task demands, and the specific error type, processing may be dominated by the "inferencer", "slacker", or "pedant" modes. For instance, the "slacker" mode generally dominates for "depth charge" sentences such as *No test is too difficult to fail*, but highly analytical individuals, as measured by a logic task, may be more likely to completely reject sentences of this type as uninterpretable. For other illusion types, such as inversion sentences (*The mother gave the candle the daughter*), conscious error correction via rational inference dominates, at least when the task is to give explicit judgments. Further investigating individual differences in the context of implemented cognitive models will prove fruitful for uncovering the fine-grained properties of sentences that support or weaken the "inferencer", "slacker" and "pedant" modes within a given reader or listener, and ultimately result in a better understanding of the interaction between sentence processing and other aspects of cognition.

# Acknowledgments

The author would like to thank Shravan Vasishth, Garrett Smith, Kiel Christianson, Adrian Staub, and Kuan-Jung Huang for helpful comments on the paper. Additional thanks go to Himanshu Yadav, Roger Levy, Ted Gibson, and the audiences at AMLaP 2022/2023 and HSP 2023 for fruitful discussions. The experiment was funded by the University of Potsdam.

# References


Bader, M. and Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(8):1286–1311.

Batchelder, W. H. and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1):57–86.

Bellur, S. and Sundar, S. S. (2014). How can we tell when a heuristic has been used? Design and analysis strategies for capturing the operation of heuristics. *Communication Methods and Measures*, 8(2):116–137.

Bhatia, S. and Dillon, B. (2022). Processing agreement in Hindi: When agreement feeds attraction. *Journal of Memory and Language*, 125:104322.

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., and Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, 43(2):83–128.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in Statistics*, pages 201–236.

Bradac, J. J., Martin, L. W., Elliott, N. D., and Tardy, C. H. (1980). On the neglected side of linguistic science: Multivariate studies of sentence judgment. *Linguistics*, 18:967–995.

Brehm, L., Jackson, C. N., and Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, 36(8):959–983.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351.

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

Bürkner, P.-C. (2024). Estimating multivariate models with brms. `https://cran.r-project.org/web/packages/brms/vignettes/brms_multivariate.html`.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., and Pickering, M. J. (2023). Does ChatGPT resemble humans in language use? *arXiv preprint*, 2303.08014.

Cai, Z. G., Zhao, N., and Pickering, M. J. (2022). How do people interpret implausible sentences? *Cognition*, 225:105101.

Chater, N. and Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3(2):57–65.

Chen, S. and Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S, C. and Y, T., editors, *Dual-process theories in social psychology*, pages 73–96. The Guilford Press.

Chen, S., Nathaniel, S., Ryskin, R., and Gibson, E. (2023). The effect of context on noisy-channel sentence comprehension. *Cognition*, 238:105503.

Christianson, K. (2008). Sensitivity to syntactic changes in garden path sentences. *Journal of Psycholinguistic Research*, 37:391–403.

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

Christianson, K., Dempsey, J., Tsiola, A., and Goldshtein, M. (2022). What if they're just not that into you (or your experiment)? On motivation and psycholinguistics. In Federmeier, K., editor, *Psychology of learning and motivation – Advances in research and theory*, pages 51–88. Academic Press.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.

Christianson, K., Luke, S. G., and Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2):538–544.

Cook, P. and Stevenson, S. (2010). No sentence is too confusing to ignore. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 61–69. Association for Computational Linguistics.

Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, 13(1):21–58.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9:519–540.

Dempsey, J., Tsiola, A., Chantavarin, S., Ferreira, F., and Christianson, K. (2023). Nonce word evidence for the misinterpretation of implausible events. *Journal of Cognitive Psychology*, pages 1–19.

Dentella, V., Murphy, E., Marcus, G., and Leivada, E. (2023). Testing AI performance on less frequent

aspects of language reveals insensitivity to underlying meaning. *arXiv preprint*, 2302.12313.
Drenhaus, H., Saddy, D., and Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 145–165.
Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, 8(11):e81461.
Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, 217(3):108–124.
Erickson, T. D. and Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5):540–551.
Evans, N. J. and Wagenmakers, E.-J. (2019). Evidence accumulation models: Current limitations and future directions. *The Quantitative Methods for Psychology*, 16(2):73–90.
Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203.
Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15.
Ferreira, F., Engelhardt, P. E., and Jones, M. W. (2009). Good enough language processing: A satisficing approach. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, volume 1, pages 413–418. Cognitive Science Society Austin, TX.
Ferreira, F. and Huettig, F. (2023). Fast and slow language processing: A window into dual-process models of cognition [Open Peer commentary on De Neys]. *Behavioral and Brain Sciences*, 46.
Ferreira, F. and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1–2):71–83.
Fortuin, E. (2014). Deconstructing a verbal illusion: The 'No X is too Y to Z' construction and the rhetoric of negation. *Cognitive Linguistics*, 25(2):249–292.
Frank, S. L. and Ernst, P. (2019). Judgements about double-embedded relative clauses differ between languages. *Psychological Research*, 83(7):1581–1593.
Frank, S. L., Ernst, P., Thompson, R. L., and Cozijn, R. (2021). The missing-VP effect in readers of English as a second language. *Memory & Cognition*, 49(6):1204–1219.
Frazier, L. (2015). Two interpretive systems for natural language? *Journal of Psycholinguistic Research*, 44:7–25.
Frazier, L. and Clifton, Jr, C. (2015). Without his shirt off he saved the child from almost drowning: interpreting an uncertain input. *Language, Cognition and Neuroscience*, 30(6):635–647.
Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.
Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.
Gibson, E. and Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248.
Gigerenzer, G. and Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.
Glenberg, A. M., Wilkinson, A. C., and Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10(6):597–602.
Goel, V. and Vartanian, O. (2011). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cognition and Emotion*, 25(1):121–131.
Goldberg, A. E. and Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, 26(4):300–311.
Goldshtein, M. (2021). *Going beyond our means: A proposal for improving psycholinguistic methods*. PhD thesis, University of Illinois Urbana-Champaign.
Hahn, M., Futrell, R., Levy, R., and Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
Hannon, B. and Daneman, M. (2004). Shallow semantic processing of text: An individual-differences account. *Discourse Processes*, 37(3):187–204.

Häussler, J. and Bader, M. (2015). An interference account of the missing-VP effect. *Frontiers in Psychology*, 6:766.

Hayes, B. K., Stephens, R. G., Lee, M. D., Dunn, J. C., Kaluve, A., Choi-Christou, J., and Cruz, N. (2022). Always look on the bright side of logic? testing explanations of intuitive sensitivity to logic in perceptual tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(11):1598–1617.

Heathcote, A. and Matzke, D. (2022). Winner takes all! what are race models, and why and how should psychologists use them? *Current Directions in Psychological Science*, 31(5):383–394.

Huang, K.-J. and Staub, A. (2021a). Using eye tracking to investigate failure to notice word transpositions in reading. *Cognition*, 216:104846.

Huang, K.-J. and Staub, A. (2021b). Why do readers fail to notice word transpositions, omissions, and repetitions? a review of recent evidence and theory. *Language and Linguistics Compass*, 15(7):e12434.

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral studies*, 19(4):577–613.

Kamide, Y. and Kukona, A. (2018). The influence of globally ungrammatical local syntactic constraints on real-time sentence comprehension: Evidence from the visual world paradigm and reading. *Cognitive Science*, 42(8):2976–2998.

Karimi, H. and Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, 69(5):1013–1040.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5):602–616.

Langsford, S., Stephens, R. G., Dunn, J. C., and Lewis, R. L. (2019). In search of the factors behind naive sentence judgments: A state trace analysis of grammaticality and acceptability ratings. *Frontiers in Psychology*, 10:2886.

Lee, P.-S. and Sewell, D. K. (2024). A revised diffusion model for conflict tasks. *Psychonomic Bulletin & Review*, 31(1):1–31.

Leivada, E. (2020). Language processing at its trickiest: Grammatical illusions and heuristics of judgment. *Languages*, 5(3):29.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065.

Levy, R., Bicknell, K., Slattery, T., and 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(50):21086–21090.

Li, J. and Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the n400 and p600 in language processing. *Cognition*, 233:105359.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., Van het Nederend, M. L., Burchert, F., Stadie, N., Caplan, D., and Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, 45(4):e12956.

Logačev, P. and Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, 40(2):266–298.

Meng, M. and Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology*, 74(1):1–28.

Muller, H. E. (2022). *What Could Go Wrong? Linguistic Illusions and Incremental Interpretation*. PhD thesis, University of Maryland, College Park.

Nicenboim, B. and Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, 99:1–34.

O'Connor, E. (2015). *Comparative illusions at the syntax-semantics interface*. PhD thesis, University of Southern California.

Orth, W., Yoshida, M., and Sloggett, S. (2021). Negative polarity item (NPI) illusion is a quantification phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(6):906–947.

Paape, D. (2023). When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2:202–218.

Paape, D., Vasishth, S., and von der Malsburg, T. (2020). Quadruplex negatio invertit? The on-line processing of depth charge sentences. *Journal of Semantics*, 37(4):509–555.

Paape, D. and Zimmermann, M. (2020). Conditionals on crutches: Expanding the modal horizon. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 108–126.

Palan, S. and Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Parker, D. and An, A. (2018). Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. *Frontiers in Psychology*, 9:1566.

Parker, D. and Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, 157:321–339.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., and Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, 123(3):335–346.

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In Runner, J., editor, *Experiments at the Interfaces*, pages 147–180. Brill.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Piironen, J. and Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27:711–735.

Pöppel, J. (2023). *Models for Satisficing Mentalizing*. PhD thesis, University of Bielefeld.

Poppels, T. and Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In Papafragou, A., Grodner, D., Mirman, D., and Trueswell, J., editors, *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4):260–281.

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.3.3.

Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2):358.

Rohaut, B. and Naccache, L. (2017). Disentangling conscious from unconscious cognitive processing with event-related EEG potentials. *Revue neurologique*, 173(7-8):521–528.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., and Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80:491–513.

Ryskin, R., Futrell, R., Kiran, S., and Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, 181:141–150.

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., and Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, 158:107855.

Sanford, A. J., Leuthold, H., Bohan, J., and Sanford, A. J. (2011). Anomalies at the borderline of awareness: An ERP study. *Journal of Cognitive Neuroscience*, 23(3):514–523.

Sanford, A. J. and Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, 6(9):382–386.

Schlesinger, I. M. (1975). Why a sentence in which a sentence in which a sentence is embedded is embedded is difficult. *Linguistics*, 13(153):53–66.

Schwarz, F. and Zehr, J. (2021). Tutorial: Introduction to PCIbex – An Open-Science Platform for Online Experiments: Design, Data-Collection and Code-Sharing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Selten, R. (1990). Bounded rationality. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 146(4):649–658.

Shin, U., Yi, E., and Song, S. (2023). Investigating a neural language model's replicability of

psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology*, 14.
Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138.
Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, 1:161–176.
Solcz, S. (2011). *Not All Syllogisms Are Created Equal: Varying Premise Believability Reveals Differences Between Conditional and Categorical Syllogisms*. PhD thesis, University of Waterloo.
Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.26.8.
Stella, M. and Engelhardt, P. E. (2022). Use of parsing heuristics in the comprehension of passive sentences: evidence from dyslexia and individual differences. *Brain Sciences*, 12(2):209.
Stupple, E. J., Ball, L. J., Evans, J. S. B., and Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, 23(8):931–941.
Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36:201–216.
Teodorescu, A. R. and Usher, M. (2013). Disentangling decision models: from independence to competition. *Psychological Review*, 120(1):1–38.
Tessler, M. H., Tenenbaum, J. B., and Goodman, N. D. (2022). Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, 14(3):574–601.
Trippas, D., Kellen, D., Singmann, H., Pennycook, G., Koehler, D. J., Fugelsang, J. A., and Dubé, C. (2018). Characterizing belief bias in syllogistic reasoning: A hierarchical Bayesian meta-analysis of ROC data. *Psychonomic Bulletin & Review*, 25:2141–2174.
Trippas, D., Pennycook, G., Verde, M. F., and Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning*, 21(4):431–445.
Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712.
Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2019). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.1.0.
Vehtari, A., Gelman, A., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.
Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432.
Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142 – 228.
von der Malsburg, T. and Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, 28(10):1545–1578.
Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.
Wason, P. C. and Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, 31(4):591–597.
Wellwood, A., Pancheva, R., Hacquard, V., and Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, 35(3):543–583.
Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language & Communication*, 18:47–67.
Xiang, M., Grove, J., and Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology*, 4:708.
Yadav, H., Paape, D., Smith, G., Dillon, B. W., and Vasishth, S. (2022). Individual Differences in Cue Weighting in Sentence Comprehension: An Evaluation Using Approximate Bayesian Computation. *Open Mind*, 6:1–24.
Yadav, H., Smith, G., Reich, S., and Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, 129:104400.
Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.
Zhang, Y., Gibson, E., and Davis, F. (2023a). Can language models be tricked by language illusions? Easier with syntax, harder with semantics. *arXiv preprint*, 2311.01386.
Zhang, Y., Ryskin, R., and Gibson, E. (2023b). A noisy-channel approach to depth-charge illusions. *Cognition*, 232:105346.