Journal of Experimental Psychology: General

Comparative Illusions Are Evidence of Rational Inference in Language Comprehension

Yuhan Zhang, Carina Kauf, Roger P. Levy, and Edward Gibson

Online First Publication, July 21, 2025. https://dx.doi.org/10.1037/xge0001807

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001807

Comparative Illusions Are Evidence of Rational Inference in Language Comprehension

Yuhan Zhang[1], Carina Kauf[2], Roger P. Levy[2], and Edward Gibson[2]
[1] Department of Linguistics, Stanford University
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Sometimes sentences sound acceptable when they are ungrammatical or semantically implausible. In this article, we study "comparative illusion" (CI) sentences where people often rate a sentence like *More people have been to Russia than I have* to be acceptable while in fact it is semantically anomalous. We provide a potential explanation for this language illusion from the noisy-channel framework. We hypothesize that comprehenders make rational inferences over the perceived sentence by entertaining alternative "close" plausible interpretations, where closeness is determined by possible production errors. In four experiments, (a) we identified a linguistic construction that elicits a salient CI illusion effect, (b) we established a range of plausible interpretations of the CI sentence, and (c) we found that the probability for comprehenders to assign a certain plausible interpretation to the CI sentence is proportional to how likely they think that interpretation is to be produced as the CI sentence during noisy language communication. This work contributes to a growing body of literature supporting rational noisy-channel inference during language comprehension.

Public Significance Statement
Sometimes sentences sound acceptable when they are ungrammatical or semantically implausible. One such case is the so-called comparative illusion *More people have been to Russia than I have*, which sounds like a good English sentence, despite the fact that it does not make sense. We show that native speakers make rational inferences over illusion sentences by postulating the plausible alternative intended sentences behind the surface sentence and mentally computing how likely each intended sentence can be represented by the perceived anomalous sentence. We use behavioral experiments to show that comprehenders probabilistically choose the one with the highest likelihood. This study highlights how the domain-general Bayesian inference principle interacts with linguistic knowledge in language processing with uncertainty.

Keywords: language illusion, language processing, psycholinguistics, information theory, rational inference

Supplemental materials: https://doi.org/10.1037/xge0001807.supp

At first glance, sentence (1) often seems acceptable to English speakers, but they often find it difficult to pin down the exact meaning. Sentences like (1) have been said to give rise to an illusion effect: If interpreted literally, they do not make sense, but the sentences seem acceptable initially. This phenomenon is termed the comparative illusion (CI) because the part that tricks people is related to comparative structures.

(1) More people have been to Russia than I have. (CI sentence)

Si On Yoon served as action editor.
Yuhan Zhang https://orcid.org/0000-0002-0550-9581
All the experimental materials, the collected data, and the analysis scripts can be accessed on the Open Science Framework at https://osf.io/89d5b/. A preprint of this article has been made publicly available and can be found at https://osf.io/preprints/psyarxiv/efr3q_v1. The data and ideas of this article have been presented at the conference Architectures and Mechanisms for Language Processing (Basque, Spain) on September 2, 2023, and the Linguistic Society of America Annual Meeting (New York, United States) on January 4, 2024. Edward Gibson and Roger P. Levy received funding from Grant 2121074 from the National Science Foundation.

Yuhan Zhang played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, writing–original draft, and writing–review and editing. Carina Kauf played a lead role in data curation, investigation, methodology, and writing–review and editing and a supporting role in formal analysis, project administration, and visualization. Roger P. Levy played a lead role in funding acquisition, supervision, and writing–review and editing and a supporting role in investigation and methodology. Edward Gibson played a lead role in conceptualization, funding acquisition, methodology, supervision, and writing–review and editing and a supporting role in data curation, formal analysis, investigation, project administration, resources, validation, visualization, and writing–original draft.

Correspondence concerning this article should be addressed to Yuhan Zhang, Department of Linguistics, Stanford University, Margaret Jacks Hall, Building 460 Room 127, Stanford, CA 94305-2150, United States. Email: yuhanczhang@gmail.com

The final verb *have* in (1) introduces an elliptical verb phrase whose antecedent is most naturally understood to be *have been to Russia*, so the sentence should have the same meaning as (1′):

(1′) More people have been to Russia than I have been to Russia.

But *I have been to Russia* does not offer the right kind of information—a quantity or degree—to be compared.

This CI phenomenon was first noticed in Montalbetti (1984) and has been studied in various psycholinguistic studies (Christensen, 2010, 2016; Kelley, 2018; Langsford et al., 2019; Leivada, 2020; O'Connor, 2015; O'Connor et al., 2013; Paape, 2024; Pham, 2022; Wellwood et al., 2018), but it is not well understood. Understanding why illusions like the CI occur may tell us more about how typical language processing works (e.g., Ferreira et al., 2002; Ferreira & Patson, 2007; Goldberg & Ferreira, 2022; Sanford & Sturt, 2002), much as how visual illusions inform us how the visual system works (Coren & Girgus, 2020; Gregory, 1968; Robinson, 1972).

In this article, we propose an account of the surprising part of the CI effect—that the sentences sound acceptable in the first place—based on the noisy-channel theory of human language comprehension (e.g., Gibson et al., 2013; Levy, 2008; Shannon, 1948). We hypothesize that during the comprehension of a CI sentence like (1), comprehenders mentally evaluate not only the literal sentence but also plausible alternative neighboring sentences that a producer might have intended. Since the literal sentence is semantically ill-formed, it may instead be interpreted as one of these alternatives so long as (a) the alternative interpretation is semantically well-formed and (b) the alternative sentence is similar in form to the literal sentence. For sentence (1), for example, a relevant alternative sentence might be (1″) below:

(1″) People have been to Russia more than I have.

According to our hypothesis, the CI effect arises when the well-formed neighboring sentences mislead comprehenders into initially finding sentences like (1) acceptable. If this hypothesis is correct, then factors modulating alternative sentences' plausibility and proximity to the literal sentence should modulate the magnitude of the CI effect (i.e., how likely a native speaker is to find such a sentence acceptable). In this article, we empirically test this key prediction of our hypothesis, using mathematical formalization based on noisy-channel principles of Bayesian probabilistic inference. After first establishing which variants of CI sentence structures like (1) show the greatest illusion effect (Experiment 1), our behavioral results (Experiment 2) reveal three common interpretations. They feature an event comparison interpretation where *more* is used to compare the times or frequencies of the event at issue (e.g., "Students have been to Russia more than I have"), an individual comparison interpretation where the number of people from the two groups is compared (e.g., "There are more students who have been to Russia than just me"), and an event negation interpretation where the subject in the *than* clause has not participated in the said event (e.g., "(Many) students have been to Russia but I have not"). We then ask a separate group of participants to choose the most probable interpretation of the CI sentence (Experiment 3) and replicate the prevalence order observed in Experiment 2. We hypothesize that comprehenders regard these three interpretations as three possible messages intended by a producer who ends up producing them errorfully into the form of the CI sentence. Each intended message can be transformed into the CI sentence via a unique noise corruption profile. According to the noisy-channel theory, the likelihood of each noise corruption predicts the posterior probability for a comprehender to assign the corresponding interpretation to the CI sentence. In Experiment 4, we collect participants' ratings on the likelihood of each of the three noise corruptions. We find that the intended sentence with an event comparison interpretation that shifts the adverbial *more* to the beginning of the sentence has the highest likelihood, followed by the intended sentence with an individual interpretation that substitutes the *than* clause and the intended sentence with an event negation interpretation that substitutes the contrastive clause by a *than* clause. The results in Experiments 3 and 4 show that the probability for comprehenders to reach a certain intended sentence of CI is proportional to how likely that intended sentence is to be produced as the CI sentence, supporting the noisy-channel theory of the CI.

The structure of the article is as follows. In the introduction, we first review existing accounts that aim to address how the illusion arises. Then, we introduce the noisy-channel framework and detail why this framework has the potential as an explanation and how this theory can systematically explain the findings in previous studies. We then report our four experiments that provide supporting evidence for the noisy-channel account of the CI. Finally, we conclude with discussion and suggestions for future research.

Sentence Template Blending Account

The sentence template blending account (named by Wellwood et al., 2018, based on ideas expressed in Townsend & Bever, 2001) claims that the acceptable judgment of sentences like (1) originates from the integration of two sentences similar to it, shown in (2) (words in parentheses are included in the templates). The segments outside of the parentheses can be blended into the canonical CI sentence. This account is based on Townsend and Bever's (2001) analysis-by-synthesis model of sentence processing. This proposal posits that upon hearing a sentence, listeners initiate a hypothesis about its meaning by connecting the content words and phrases according to how these content words and phrases usually form a syntactic template with their statistical distributions. Then listeners generate a more complete syntactic parse to check if the initial lexically and statistically motivated hypothesis is consistent with the syntactically well-formed parse. If these two processes do not match, listeners will be confused by the perceived sentence.

(2a) More people have been to Russia than I (could believe).

(2b) People have been to Russia (more) than I have.

Under this proposal, comprehenders accept (1) because the two sentence templates they rely on are acceptable. An open question for this account is that (2a) and (2b) convey different meanings and there is no prediction as to which meanings result in comprehenders' minds: (2a) expresses that the number of people who have been to Russia exceeds the speaker's expectation; (2b) expresses that the speaker has been to Russia less frequently or for fewer times than a certain group of people. A more complete theory should address what interpretation is driving the illusion effect. Furthermore, Townsend and Bever did not provide empirical evidence to support their idea.

Repair Analysis Account

O'Connor et al. (2013) and O'Connor (2015) proposed a repair analysis account, which claims that comprehenders can flexibly generate various interpretations of CI sentences by making small changes and turning the input into a plausible sentence. For CI sentences like (3a), one possible edit is moving *more* (3b) and the resulting sentence compares the frequency or number of times judges have vacationed in Florida to that of a group of lawyers—we label this reading as the "event comparison interpretation." An alternative type of edit involves the deletion of *the* (3c) or the insertion of *of the* (3d) to make the resulting sentences express a comparison between the number of people in the judge group versus the lawyer group that have vacationed in Florida—we label this reading as the "individual comparison interpretation."

(3a) More judges vacationed in Florida than the lawyers did. (CI sentence)

(3b) Judges vacationed in Florida more than the lawyers did. (*more* moved, event comparison interpretation)

(3c) More judges vacationed in Florida than ~~the~~ lawyers did. (*the* deleted, individual comparison interpretation)

(3d) More of the judges vacationed in Florida than the lawyers did. (insertion of *of the*, individual comparison interpretation)

The proposal in O'Connor (2015) provides clear intuitions about potential edits that link the perceived CI sentence with some interpretations. But she does not provide experimental evidence in support of these ideas.

Event Interpretation Account

The event interpretation account was proposed by Wellwood et al. (2018; an earlier version in Wellwood et al., 2009). It argues that, even though according to the surface form of the CI sentence the comparative morpheme *more* modifies the nominal matrix subject, during processing, participants nevertheless interpret the sentence as a comparison between numbers of events. They hint at a simple syntactic reanalysis process where *more* is interpreted as an adverb (4), which is consistent with O'Connor's (2015) proposal.

(4) ~~More~~ people have been to Russia more than I have.

To empirically test whether this hypothesis was feasible, they designed a verbatim recall task where participants were asked to recall and retype a CI-type sentence after reading it and finishing a working memory task. They expected to see recall errors related to moving *more*, as in (5), but their findings show that the frequency of this error type was lower than errors involving changing the plurality of than-clause subjects (6) and those involving changing a nonrepeatable verb phrase to a repeatable one (7).[1] Therefore, Wellwood et al. concluded that they did not have evidence for this event interpretation account.

(5) More girls ate pizza than I did. (exposure sentence)

a. Girls ate pizza more than I did. (recall from participants)

(6) More girls ate pizza than the boy did. (exposure sentence)

a. More girls ate pizza than boys did. (recall from participants)

(7) More girls began reading the book than the boy did. (exposure sentence)

a. More girls read the book than the boy did. (recall from participants)

In contrast with this finding, our experimental work shows that a comparison of events is the most common interpretation of CI sentences out of the other possibilities and we will elaborate on a noisy-channel explanation for this finding.

Rational Inference and Good-Enough Processing

Paape (2024) investigated CI sentences alongside several other types of sentences of interest in language processing research, with the goal of distinguishing the noisy-channel rational inference account (Gibson et al., 2013; Levy, 2008), which we describe in greater detail in the next section, from the good-enough processing account (Ferreira et al., 2002; Ferreira & Patson, 2007), which posits that language comprehension does not necessarily provide fully accurate or detailed representations of linguistic input, but often just incomplete and underspecified analyses that satisfice for comprehender goals. Paape speculated that rational inference and good-enough processes might both apply within an individual across different trials. In an offline judgment task, Paape asked experimental participants two questions of each sentence: (1) whether they "get it" (i.e., understand what the sentence was intended to mean) and (2) whether the sentence was correct (vs. had an error). The results show that around 50% of the responses for CI sentences were "get it, but incorrect," which Paape took to support a rational inference procedure where participants could recognize the errors and infer a plausible meaning. Around 25% of the responses were "get it, correct," which Paape took as possible support for good-enough processing. In our view, Paape's data provide encouraging initial evidence consistent with a noisy-channel account of the CI. However, we emphasize that the noisy-channel rational inference account does not require that participants notice errors; Paape himself acknowledges this, citing Levy (2008) that "in many cases, these types of correction happen at a level that may be below consciousness." Additionally, Paape does not investigate the nonliteral interpretations arrived at by comprehenders. In this article, we investigate the distribution of these interpretations, which turns out to provide crucial new evidence for the noisy-channel rational inference account.

The Noisy-Channel Hypothesis and the Comparative Illusion

In this article, we propose that the noisy-channel framework (Shannon, 1948)—a mathematical model for rational inference in language processing (Gibson et al., 2013; Levy, 2008; Poppels & Levy, 2016; Ryskin et al., 2018)—integrates intuitions behind the sentence template blending account, the repair analysis account, and

[1] (7a) is still anomalous, but Wellwood et al. (2018) did not provide further explanation.

the event interpretation account and can be used to predict the comprehension behavior of the CI. Under the noisy-channel theory, comprehenders reverse engineer what interpretation could be intended by a producer who ends up producing it with errors in the form of a CI sentence. Comprehenders mentally entertain plausible alternatives of the perceived CI sentence, and they infer a particular interpretation for the CI sentence if that interpretation has a high prior probability to be intended and a high likelihood to be produced into the CI sentence. An example alternative results from the noise operation of "shifting *more*": The intended message of (1) could be "People have been to Russia more than I have," and the producer might produce it in the form of (1). We will show in Experiment 4 that "shifting *more*" is a likely production error and comprehenders also often assign the corresponding interpretation to the CI sentence. In this section, we first introduce the noisy-channel framework and its predictions and then provide supporting evidence from four behavioral experiments.

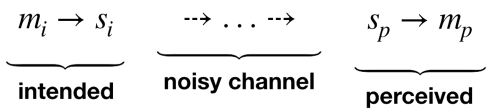
According to the noisy-channel framework, the linguistic input that people process in their daily lives often contains errors. During a conversation—as schematically represented in Figure 1, the speaker intends to convey the message m_i, which can be linguistically encoded as the intended sentence s_i through various modalities such as speaking, writing, or signing. During information transmission, the sentence s_i can be corrupted due to producer or comprehender errors or noise from the environment. In sentence production, for example, words can be deleted, inserted, exchanged, or substituted by other words. Subsequently, the linguistic signal s_p perceived by the listener or the reader could differ from s_i. The literal meaning m_p derived from s_p could thus deviate from the intended message m_i and may result in misunderstanding during communication.

Given this background, to achieve successful communication where m_p maps to the intended m_i, language comprehension can be modeled as rational Bayesian inference (Gibson et al., 2013; Levy, 2008, 2011; Levy et al., 2009; Ryskin et al., 2018; Ryskin, Stearns, et al., 2021). The inference process can be modeled in Equation 1, where, for simplicity, s_i represents the intended sentence, incorporating both the meaning and the form into one symbol:

$$P(s_i|s_p) \propto P(s_i)P(s_p|s_i), \tag{1}$$

$P(s_i|s_p)$ represents the posterior probability that the comprehender assigns the intended sentence s_i to the perceived sentence s_p. This is proportional to the product of prior probability $P(s_i)$, which represents the probability that the producer intends to convey s_i (both its form and meaning), and the noise likelihood term $P(s_p|s_i)$, which represents how likely the intended s_i could be corrupted to and represented by s_p during information transmission. In other words, comprehenders assign an interpretation s_i to the perceived sentence s_p by weighing the prior probability that s_i is intended against the likelihood that s_i is encoded into the perceived s_p.

Figure 1
Schematic Representation Communication in a Noisy Channel



The noisy-channel theory of language processing makes several quantitative predictions. First, the probability of obtaining a nonliteral inference of an implausible sentence is proportional to the prior probability of the inferred meaning, which we operationalize as the plausibility of the meaning in typical world knowledge. For instance, when people encounter the sentence *The mother gave the candle the daughter*, they might entertain (a) the literal interpretation—which is implausible, such that the daughter is given to the candle—and (b) an alternative interpretation associated with the deletion of the preposition *to* from the string *The mother gave the candle to the daughter*, which is much more typical in the world, where the candle is given to the daughter. In addition to tracking meaning plausibility (Gibson et al., 2013; Qian & Levy, 2023), the prior probability of s_i has also been shown to be sensitive to the frequency of the linguistic structure and the discourse context of the intended sentence (Chen et al., 2023; Liu et al., 2020; Poliak et al., 2025; Poliak, Ryskin, et al., 2024).

Second, the nonliteral inference rate is also proportional to the noise likelihood term. Noise distortion of the intended message could take place in multiple ways: as production errors, comprehension errors, or errors because of a noisy environment. The types of distortion vary widely. Distortion can take place at the sound/character level (Poliak, Ryskin, et al., 2024; Ryskin, Bergen, & Gibson, 2021), the morpheme/word level (Gibson et al., 2013; Poppels & Levy, 2016; Qian & Levy, 2023), or the phrasal/structural level (Zhang, Ryskin, & Gibson, 2023). The specific noise operations may involve deletions, insertions, substitutions, or exchanges of parts of the strings.

The noisy-channel theory of language processing has accumulated empirical support from a wide range of phenomena. In addition to explaining interpretations in English sentences across a range of constructions (Gibson et al., 2013; Poppels & Levy, 2016), it explains aspects of sentence interpretation in persons with aphasia (Fedorenko et al., 2023; Gibson et al., 2016; Warren et al., 2017), language illusions and error-correction mechanisms (Qian & Levy, 2023; Ryskin, Bergen, & Gibson, 2021; Zhang, 2024; Zhang, Ryskin, & Gibson, 2023), rational inference across multiple languages (Poliak, Ryskin, et al., 2024; Zhan et al., 2023), and the processing of song lyrics (Poliak, Kimura, & Gibson, 2024).

Within this noisy-channel framework, we investigate whether the comprehension of the CI can be modeled as Bayesian inference over a noisy channel. We model the sentence *More people have been to Russia than I have* as the perceived s_p and comprehender's inferred meaning(s) as s_i. We hypothesize that there are several plausible alternative messages s_i that could be inferred by the comprehender, especially given that previous literature already discussed several choices (O'Connor, 2015; O'Connor et al., 2013; Wellwood et al., 2018). One plausible interpretation is the "event comparison" sentence, which expresses "people have been to Russia more times or more frequently than I have." Another is the "individual comparison" sentence—"there are a larger number of people who have been to Russia than just me." There could be other undiscovered ones in the hypothesis space, as Christensen (2016) discovered more than two plausible interpretations in Danish CI sentences. Therefore, we specifically collected such interpretations in Experiment 2.

Each of the intended s_is would have a different noise corruption profile that transforms them to the corrupted CI sentence s_p during production. For example, when it comes to the canonical CI sentence (1), for the event comparison s_i "people have been to Russia more than I have," the noise edit during production would involve the shift

of the word *more* from later in the sentence to the sentence-initial position. According to Equation 1, we predict that the interpretation that is more likely to be produced into the CI sentence is more likely to be chosen by a comprehender as the intended message given a perceived CI sentence. In other words, a higher noise likelihood $P(s_p|s_i)$ leads to a higher posterior $P(s_i|s_p)$, assuming that the differences in the prior distribution of these alternative interpretations do not outweigh the effect of the noise likelihood term.

Theory Comparison

The noisy-channel explanation of CI integrates existing proposals and, crucially, makes additional predictions regarding CI sentence interpretation. Similar to the ideas expressed in Townsend and Bever (2001), O'Connor (2015), and Wellwood et al. (2018), the noisy-channel account predicts the existence of plausible semantically well-formed alternative interpretations for CI sentences. Within the noisy-channel account, these alternatives correspond to possible *intended* sentences on the part of the language producer. The noisy-channel account makes explicit quantitative predictions that associate the probability of assigning each alternative interpretation to the CI sentence with the likelihood that the sentence expressing that alternative interpretation might be transformed to the CI sentence in the process of errorful speech production. The current article provides new experimental data on the frequencies of possible CI sentence interpretations and their associated noise/error likelihoods to test the quantitative predictions made by the noisy-channel theory.

Experiment Outline

We report four experiments in this article. Experiment 1 consists of an acceptability judgment task where we identified CI sentences that have the highest degree of illusion. We found that illusion sentences with a pronoun than-clause subject (e.g., *More lawyers have vacationed in Florida than I have*) received acceptability scores as high as grammatical controls, while full noun phrase than-clause subjects (e.g., *More lawyers have vacationed in Florida than the clerk has*) were rated less acceptable. We therefore focus on the pronoun cases in the following experiments.

Experiment 2 consists of a paraphrase task that gauged English speakers' interpretation of the CI sentences with a pronoun than-clause subject. After annotation and analysis, we selected the predominant paraphrases for use in later experiments. These interpretations consisted of an event comparison interpretation (e.g., "Lawyers have vacationed in Florida more than I have"), an individual comparison interpretation (e.g., "There are more lawyers who've vacationed in Florida than just me"), and an event negation interpretation (e.g., "(Many) lawyers have vacationed in Florida but I have not").

Experiment 3 consists of a forced-choice task that juxtaposed the three most prominent interpretations from Experiment 2 and asked participants to select the most predominant reading given a CI sentence. The percentage for each interpretation was taken as the proxy for the posterior distribution of $P(s_i|s_p)$. The results show that the event comparison interpretation was the most frequent choice, followed by individual comparison and event negation.

We hypothesize that comprehenders model the perceived CI sentence as the result of language production errors where each of the intended sentences is the intended message, which can be corrupted to a CI sentence via a unique noise profile during production. We motivated the noise operation by independent research in language production and a corpus analysis. We then designed Experiment 4, a noise likelihood rating task initially used by Zhang, Ryskin, and Gibson (2023), to obtain participants' ratings for the likelihood of the noise operation and treat the rating as a behavioral proxy for the noise likelihood term $P(s_p|s_i)$. The task asked how likely each intended sentence out of the three was to be produced as the perceived CI sentence. We labeled the three conditions as "shifting *more*," "than-clause substitution," and "negation-to-comparative transformation," and we found a declining noise likelihood rating in this order. Because this noise likelihood order across the three conditions in Experiment 4 corresponds to the interpretation order in Experiment 3, we conclude that the empirical evidence in our experiments is consistent with the predictions of a noisy-channel explanation of the CI. We discuss the compatibility of the data with alternative theories in the General Discussion section.

Experiment 1

Previous studies have investigated how different linguistic factors modulate the acceptability of CI sentences. First, CI sentences with a repeatable verb phrase tended to receive a higher acceptability rating than those with a nonrepeatable verb phrase (Kelley, 2018; O'Connor, 2015; O'Connor et al., 2013; Wellwood et al., 2018). Repeatable verb phrases denote events that one single individual can participate in multiple times over a relevant time scale. In example (8a), the verb phrase *vacation in Florida* is repeatable because one can easily imagine someone taking a vacation in Florida multiple times. The canonical CI sentence (1) also has a repeatable verb phrase *have been to Russia*. In contrast, nonrepeatable verb phrases denote events that have a natural end point (e.g., Dahl, 1981; Garey, 1957). Usually, the event occurs only once over a relevant time scale. In example (8b), the verb phrase *retired to Florida* is nonrepeatable because the denoted event naturally ends with the retired John settling down in Florida and it is unusual for a person to retire multiple times. Given this, CI sentences as in (9a) have been rated more acceptable than CI sentences as in (9b).

(8a) John vacationed in Florida.

(8b) John retired to Florida.

(9a) More judges vacationed in Florida than the lawyer did.

(9b) More judges retired to Florida than the lawyer did.

Another linguistic factor is the number of the than-clause subject. Than-clause subjects with plural noun phrases (10b) have received higher acceptability ratings than those with singular noun phrases (10a; O'Connor, 2015; Wellwood et al., 2018), but the number effect was not present with pronoun than-clause subjects (11; Wellwood et al., 2018).

(10a) More judges vacationed in Florida than the lawyer did.

(10b) More judges vacationed in Florida than the lawyers did.

(11a) More girls ate pizza than I did.

(11b) More girls ate pizza than we did.

To investigate which configurations of these linguistic factors elicit the strongest illusion effect, Experiment 1 consisted of an acceptability judgment task. Experiment 1A featured first-person pronoun than-clause subjects, and Experiment 1B featured full noun phrase than-clause subjects. Both Experiments 1A and 1B manipulated the repeatability of the verb phrase and the number feature of the than-clause subject. In addition, all the materials used the present perfect tense, which is close to the canonical CI sentence (1), in contrast to the simple past tense that has been used in previous studies (Kelley, 2018; O'Connor, 2015; O'Connor et al., 2013; Paape, 2024; Wellwood et al., 2018).

Method

Participants

In Experiment 1A, 49 participants were recruited from the crowdsourcing platform Prolific (https://www.Prolific.com). Each participant was paid $4 for their participation. We excluded data from those (a) who did not complete at least 90% of all questions, (b) who did not answer at least 75% of the comprehension checks correctly, (c) who gave the same rating across all test trials, and/or (d) who self-identified as nonnative speakers of English or not born in the United States. We analyzed the remaining 45 participants' responses.

In Experiment 1B, a different group of 48 participants was recruited from https://www.Prolific.com. After the same screening procedure, 41 participants remained for the formal analysis.

Throughout the study, we collected data from participants who self-identified as native speakers of English and we reported data from participants only from the United States. The reason was to focus on the use of American English. We did not collect information related to gender, sex, race, or ethnicity of participants because they are irrelevant to the goal of our study.

Materials and Procedure

Both experiments were offline acceptability judgment tasks. Materials were composed of critical items and fillers. The critical materials featured a 2 × 2 manipulation with two additional control conditions. Table 1 shows an example item for Experiment 1A. The first manipulation was the repeatability of the verb phrase. The first author manually created items that varied verb repeatability, and the materials were evaluated by the second and the last authors. Nonrepeatable verb phrases were carefully scrutinized such that an interpretation of experiencing the event multiple times would be semantically odd. The second manipulation concerned the number feature of the than-clause subject: Experiment 1A compared the singular first-person pronoun *I* versus the plural first-person pronoun *we*; Experiment 1B compared singular determiner noun phrases (e.g., *the clerk*) versus their plural counterparts (e.g., *the clerks*). Notably, sentences within the 2 × 2 manipulation were all semantically ill-formed variances of the CI sentence. Please refer to Supplemental Tables S1 and S2 for all critical items in Experiments 1A and 1B.

We constructed the control conditions where *more* was used as an adverb to modify the times or frequency of the event referred to by the verb phrase. The "good control" condition consisted of plausible sentences where the main verb phrase was a repeatable event and was natural to be modified by *more*. The "bad control" condition consisted of implausible sentences where it was not natural to combine nonrepeatable events with *more*. We created 30 items with such a design.

Experiment 1B had a similar design. As shown in Table 2, the 2 × 2 manipulation crossed the repeatability of the verb phrase and the number of the than-clause subject. Different from Experiment 1A, the control conditions both consisted of plausible sentences where the than-clause subjects were bare plural noun phrases. There, the component being compared was the cardinality of the sets of individuals denoted by the matrix subject and the than-clause subject. Thirty items were created with this design.

Experiments 1A and 1B shared the same set of filler items. We designed 64 filler items, which contained either comparative structures or various types of quantifiers. The following 16 constructions appeared four times each: *a few*, *fewer*, *more … than*, *more than*, *many*, *much*, *little*, *less*, *some*, *any*, *a lot of*, *no*, *plenty of*, *enough*, *none*, and *all*. All filler items were grammatical and plausible sentences. An example is *Fewer people have visited Greenland than Canada*.

The procedures for Experiments 1A and 1B were identical. Each participant read a list of 94 sentences in a randomized presentation. Each trial was followed by a yes/no comprehension question (e.g., "Does this sentence mention lawyers and Florida?") that encouraged participants to pay attention to the material. The answer to the comprehension question was counterbalanced such that half of the critical trials were "Yes" and half were "No." Then participants were asked to rate "How natural is the sentence?" on a 7-point fully labeled Likert scale (1 = *extremely unnatural*, 2 = *unnatural*, 3 = *somewhat unnatural*, 4 = *neutral*, 5 = *somewhat natural*, 6 = *natural*, 7 = *extremely natural*).

Table 1
Example Material for Experiment 1A

Pronoun as than-clause subject	Singular subject	Plural subject	Control
Repeatable verbs	More lawyers have vacationed in Florida than I have.	More lawyers have vacationed in Florida than we have.	Many lawyers have vacationed in Florida more than I have. (good)
Nonrepeatable verbs	More lawyers have retired to Florida than I have.	More lawyers have retired to Florida than we have.	Many lawyers have retired to Florida more than I have. (bad)

Table 2
Example Material for Experiment 1B

Noun phrase as than-clause subject	Singular subject	Plural subject	Control
Repeatable verbs	More lawyers have vacationed in Florida than the clerk has.	More lawyers have vacationed in Florida than the clerks have.	More lawyers have vacationed in Florida than clerks have. (good)
Nonrepeatable verbs	More lawyers have retired to Florida than the clerk has.	More lawyers have retired to Florida than the clerks have.	More lawyers have retired to Florida than clerks have. (good)

Results

Figure 2 shows the raw acceptability rating scores for sentences with a pronoun than-clause subject in Experiment 1A.[2] The good control condition and the bad control condition set an upper bound and a lower bound for the acceptability ratings across the six conditions. A visual inspection shows that the CI sentences in the four critical conditions were rated similarly to the good control condition, with slight degradation in the nonrepeatable verb phrase conditions. It is not easy to tell whether the number feature of the than-clause subject plays an important role just from eyeballing.

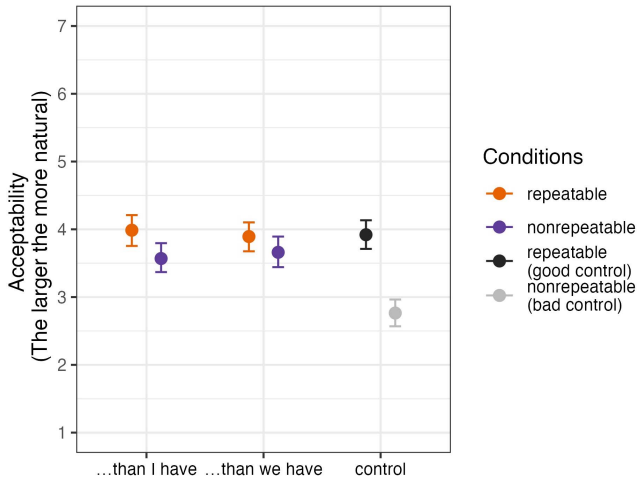
To statistically investigate the effects of verb repeatability and subject number, we ran a Bayesian multilevel cumulative ordinal model, using the *brms* package (Bürkner, 2017; Bürkner & Vuorre, 2019) in R (R Core Team, 2018). This approach models the raw acceptability score for each trial as a standard normal latent dependent variable split into levels with nonequally sized intervals whose boundaries are modulated by fixed effects and group-level random effects. Verb repeatability was entered as a dummy-coded categorical variable (reference level = repeatable), so was the subject number (reference level = singular *I*, the other two levels being "plural *we*" and "control"). The interaction term between these two main effects was also included.[3] To achieve the maximal random effect structure (Barr et al., 2013), the model included a random intercept and random slopes for both the main effects and the interaction effect for both participants and items (see Supplemental Material Section 3 for the full results of all statistical analyses in this article).

We set weakly informative priors for the model parameters: All the intercepts and coefficients were set to have a normal distribution ($M = 0$, $SD = 2$), and the priors for the correlation matrices followed LKJ(2), the correlation prior put forward by Lewandowski, Kurowicka, and Joe and named after them (Lewandowski et al., 2009; Nalborczyk et al., 2019; Nicenboim et al., 2021). The rest of the model parameters were set as the default in *brms*. The model had four sampling chains, each with 4,000 iterations. The first 2,000 samples were taken as a warm-up. To interpret the results, we relied on an $\hat{R}$ that is close to 1.0 to mark the convergence of the sampling chain to the underlying posterior distribution of the target predictor. The model shows that all $\hat{R}$s for the sampling chains for all fixed effects were 1.0, indicating successful convergence. In this article, we use β to represent the posterior mean of the distribution of the estimated coefficients for predictors and CrI to represent the 95% credible interval. These coefficients are on the scale of the standard normal latent variable assumed in cumulative regression to underlie the observed values of the ordinal dependent variable.

The result shows that, with at least 95% probability, (a) sentences with nonrepeatable verb phrases received lower ratings than those with repeatable verb phrases ($\beta = -0.38$, 95% CrI [−0.63, −0.14]) and (b) control sentences with nonrepeatable verb phrases were rated lower compared to sentences with repeatable verbs and singular than-clause subjects ($\beta = -0.82$, 95% CrI [−1.18, −0.46]). For the rest of the condition manipulations, since the 95% credible intervals of the posterior estimates contain zero, we conclude that their effects on sentence acceptability were small or uncertain. Specifically, sentences with *we* were not different from sentences with *I* ($\beta = -0.11$, 95% CrI [−0.32, 0.10]); the same with sentences in the control condition ($\beta = -0.06$, 95% CrI [−0.38, 0.26]). There was no interaction between verb repeatability and the *I*/*we* manipulation ($\beta = 0.18$, 95% CrI [−0.11, 0.47]), either.

Figure 3 shows the acceptability rating patterns in Experiment 1B with noun phrases as the than-clause subject. It is obvious that (a) nonrepeatable verb phrases received lower ratings compared with repeatable verb phrases, (b) singular than-clause subjects received lower ratings than plural than-clause subjects, and (c) sentences in all

Figure 2
Mean Acceptability Ratings When the Than-Clause Subject Was Pronoun (Experiment 1A)



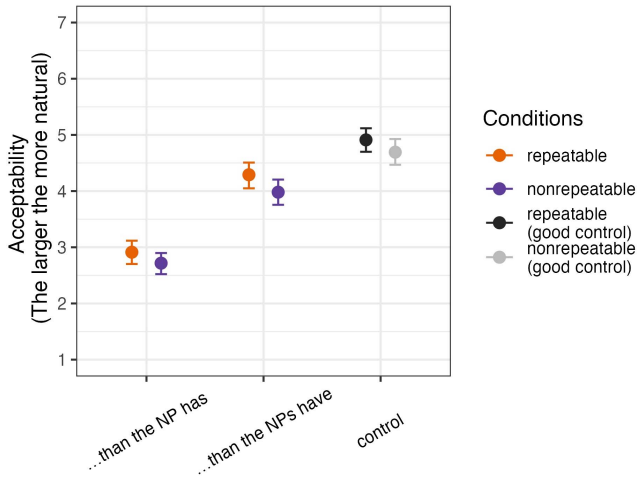
Note. The dots represent the mean of raw acceptability score across conditions in Experiment 1A with 95% bootstrapped confidence intervals. See the online article for the color version of this figure.

[2] Supplemental Figures S1 and S2 show the distribution of the data in Experiments 1A and 1B in boxplots with jittered dots.

[3] Using the leave-one-out cross-validation method (Bürkner & Vuorre, 2019; Vehtari et al., 2017), we chose the statistical model with an interaction term between the two critical manipulations as a better fitted model.

Figure 3

Acceptability Ratings When the Than-Clause Subject Was NP



Note. The dots represent the mean of raw acceptability score across conditions in Experiment 1B with 95% bootstrapped confidence intervals. NP = noun phrase. See the online article for the color version of this figure.

four critical conditions received lower ratings compared to the control conditions with acceptable sentences.

Data in Experiment 1B were analyzed in the same way as in Experiment 1A. In a Bayesian multilevel cumulative ordinal model, the dependent variable was the raw acceptability rating score per trial; the dummy-coded verb repeatability (reference level = repeatable) and the number feature (reference level = singular) were taken as the fixed effect. Using leave-one-out cross-validation, we chose the model without an interaction term between the two main effects. The random effects were maximal structure for both participants and items. The priors and model parameters were the same as those in Experiment 1A.

The results show that, with at least 95% probability, sentences with nonrepeatable verbs were rated less acceptable than sentences with repeatable verbs ($\beta = -0.22$, 95% CrI [−0.41, −0.03]); sentences with a plural noun phrase than-clause subject were rated more acceptable than those with a singular noun phrase ($\beta = 1.21$, 95% CrI [0.94, 1.48]); sentences in the control condition were rated even higher ($\beta = 1.83$, 95% CrI [1.46, 2.22]).

In another Bayesian multilevel cumulative ordinal regression model, we tested the effect of verb repeatability and subject number by collapsing the critical 2 × 2 data in Experiments 1A and 1B. The raw rating score across the two experiments was the dependent variable; verb repeatability (reference = nonrepeatable), than-clause subject form (reference = pronoun), and the subject number (reference = singular) were entered as dummy-coded main-effect variables, with an interaction term crossing the subject form and number. Random intercepts for both participants and items, random slopes for repeatability and subject number for participants, and random slopes for repeatability, subject form, and subject number with an interaction term of the latter two for items were entered as the random effects.[4] The rest of the model parameters were the same as the previous two.

The result shows that, with at least 95% probability, (a) having repeatable verb phrases increased the acceptability ($\beta = 0.27$, 95% CrI [0.15, 0.39]), (b) having a noun phrase than-clause subject reduced the acceptability ($\beta = -0.97$, 95% CrI [−1.49, −0.44]), and (c) there was no main effect of plural subjects ($\beta = -0.01$, 95% CrI [−0.21, 0.18]) but an interaction effect existed where sentences with a plural noun phrase than-clause subject were rated higher ($\beta = 1.24$, 95% CrI [0.96, 1.52]).

Discussion

Experiment 1 provides a picture of the acceptability landscape of CI sentences. We see that CI sentences with a pronoun than-clause subject triggered a stronger illusion effect than those with a noun phrase than-clause subject since the former was rated similarly to the grammatical and plausible controls but the latter received apparently lower ratings than the good controls. Furthermore, the repeatability of verb phrases played a major role in modulating the acceptability ratings across critical items, replicating the findings in previous studies (O'Connor, 2015; O'Connor et al., 2013; Wellwood et al., 2018). The finding that the number feature of the than-clause subject was significant in the noun phrase case but not in the pronoun case also replicates the findings of Wellwood et al. (2018).

Since the pronoun case—the case for which the CI was originally reported—elicited a stronger illusion effect, the following experiments will focus on explaining these specific sentence types.

Experiment 2

Now that we have shown that the strongest case of illusion involves a repeatable verb phrase and a pronoun than-clause subject, we will seek to identify the range of possible interpretations before we attempt to work out the inference mechanism in the noisy-channel theory. Although no previous study has elicited English speakers' interpretations of canonical CI sentences, Christensen (2016) elicited interpretations of CI sentences using a paraphrase paradigm in Danish. The English equivalents of the prevalent interpretations are in (12): an "excluding me" reading (12a), a "more people than just me" reading (12b), and a "more times than me" reading (12c). Experiment 2 thus gauged the available interpretations of CI sentences in English.

(12a) Some people have been to Russia except me.

(12b) More people have been to Russia than (just) me.

(12c) Some people have been to Russia more (often) than I have.

Method

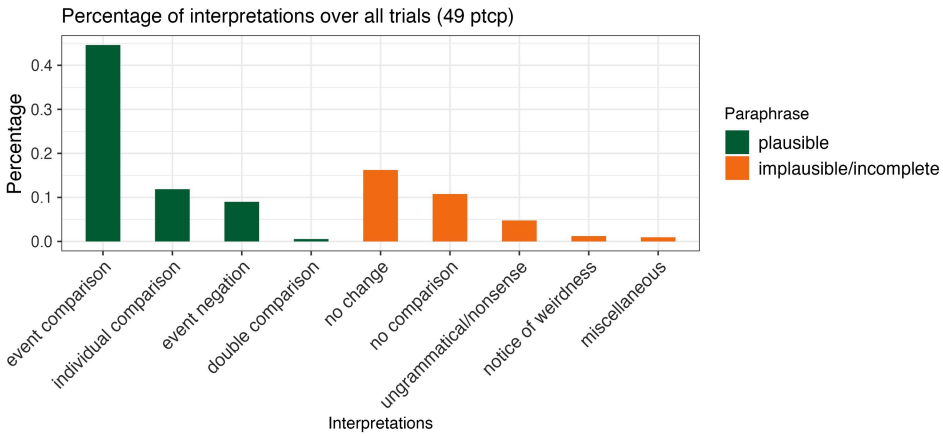
Participants

A total of 58 participants who self-identified as native speakers of English were recruited from https://Prolific.com. Each received a $2 payment for their participation. Only those who (a) finished at least 90% of all questions, (b) answered at least 90% of the comprehension questions correctly, and (c) provided grammatical paraphrases to more than 2/3 of the trials were included in the analysis, resulting in 49 participants for the formal analysis.

[4] Through leave-one-out cross-validation, we excluded a more complicated model that entered a three-way interaction because it had poorer goodness of fit.

Figure 4

Percentage of Different Categories of Interpretations Across All Trials



Note. ptcp = participant. See the online article for the color version of this figure.

Materials and Procedure

The critical materials were the 30 items that were extracted from Experiment 1 and that featured repeatable verb phrases with a pronoun than-clause subject (e.g., *More lawyers have vacationed in Florida than I/we have*). The only within-subjects condition manipulation was the pronoun number feature (*I* vs. *we*). We chose this subset of sentences from Experiment 1 because they elicited the strongest illusion effect.

To mitigate learning and fatigue effects, we created two lists, each containing 15 items. For each sentence, the pronoun than-clause subject was randomly chosen between *I* and *we*. The number feature of the than-clause subject was counterbalanced across each list. No filler items were included in this experiment.

The experiment instructions were as follows:

> In this study, we are looking into how native speakers of English use and interpret the word "more." "More" can be used in different positions in a sentence and mean different things. For example, in the sentence *I walk more than I run*, it can mean "the frequency of me walking is higher than the frequency of me running." It can also mean "the total time duration from walking is longer than running." In another sentence *I read more books than journals*, the meaning is simpler, which is "the number of books I read is larger than the number of journals I read." We provide 15 sentences including the word "more" in the survey and ask you to try your best to provide a clear and accurate paraphrase that is similar to the ones above. Please type in your answer. We would not be able to pay those who just copy and paste the original sentence.

For each trial, the paraphrase instruction was "What does this sentence mean exactly?" After the participant finished typing in their paraphrase, a yes/no comprehension question followed to encourage participants to be attentive (e.g., "Does this sentence mention COVID-19 and vaccines?"). The answers to the comprehension question were counterbalanced between "Yes" and "No."

Results and Analysis

The analysis took place in two steps: paraphrase coding and statistical summarization. The first step involved a manual coding of each paraphrase. The first two authors went through 868 paraphrase trials and assigned interpretable labels. They first did a pilot coding session with 100 trials and settled on a coherent coding schema before finishing the rest of the paraphrases. The Cohen's κ statistic for the interrater reliability for all trials was 0.73, achieving substantial agreement (Landis & Koch, 1977). The two coders later reconvened to settle the disagreement and reached consensus on the labels for all trials.

Figure 4 displays the percentage of different categories of interpretations across all trials. The plausible categories are color-coded in green, and the implausible or incomplete categories are in orange. Altogether, there are nine categories, out of which four types emerged as plausible ones. We labeled them as "event comparison" (41.65%), "individual comparison" (11.34%), "event negation" (8.35%), and "double comparison" (1.07%). The event comparison interpretation was the most dominant. There were five categories of implausible or incomplete paraphrases: "no change" (20.64%), "no comparison" (10.86%), "ungrammatical/nonsense" (3.94%), "notice of weirdness" (1.19%), and "miscellaneous" (0.95%).

Table 3 displays one paraphrase example for each category. For "event comparison," *more* was used to compare the frequency or times of events occurring for the two groups.[5] For "individual comparison," *more* was used to compare the cardinality of different groups. For "event negation," participants inserted a contrastive reading by conveying that the than-clause subject has not participated in the events that the matrix subject has experienced. The "double comparison" interpretation constituted a smaller proportion where people applied *more* twice during the interpretation. "No change" means the syntactic structure of the paraphrase was almost the same as the original CI sentence. "No comparison" means in the paraphrase

[5] In our experimental items, there were five repeatable verbs (*support the union*, *enjoy the Harry Potter series*, *hope for the end of the COVID-19 pandemic*, *enjoy high school*, *love alcohol*) that can be modified by an adverbial *more* (e.g., *Americans have enjoyed high school more than I have*), but the interpretation is not precisely a comparison of event count or frequency, but more of a comparison of degrees, since these verb phrases describe more of a status, rather than an event. For simplicity, we grouped paraphrases with *more than* within these items into the "event comparison" category, because in both these specific cases and the canonical cases *more* modifies the verb phrase, rather than the matrix subject.

Table 3
Example Paraphrase for Each Category

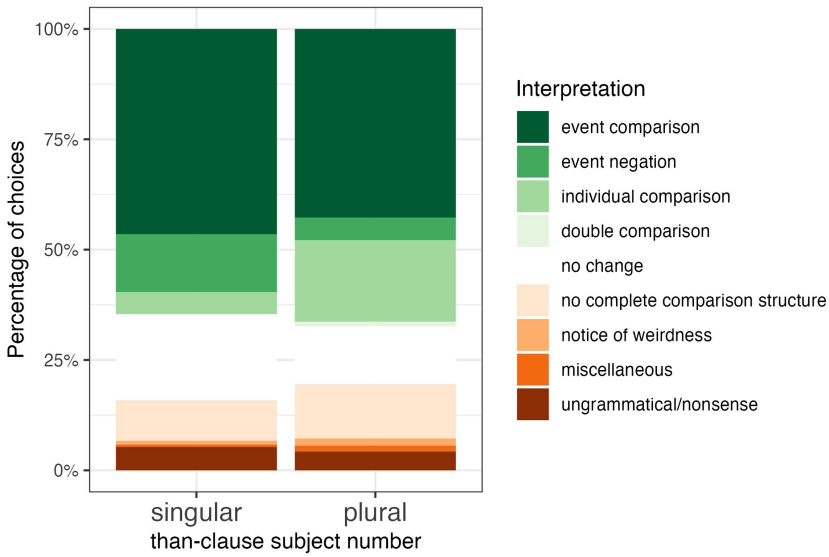
Label	CI sentence	Paraphrase
Event comparison	More students have been to Russia than I have.	Students have been to Russia *more (times)* than I have.
Individual comparison	More Americans have toured Antelope Canyon than we have.	More Americans have toured Antelope Canyon than *non-Americans*.[a]
Event negation	More economists have talked about the recession in 2022 than I have.	The speaker most likely *did not* talk about the 2022 recession.
Double comparison	More engineers have traveled to San Francisco than we have.	*More* people with the profession of Engineer travel to San Francisco *more regularly*.
No change	More teenagers have used Tiktok than I have.	More teenagers use Tiktok than I do.
No comparison	More residents of Massachusetts have gone skiing than we have.	The people of Massachusetts have gone skiing.
Ungrammatical/nonsense	More residents of Massachusetts have gone skiing than I have.	There are a higher number of people skiing than him.
Notice of weirdness	More lawyers have vacationed in Florida than I have.	Doesn't make sense honestly.
Miscellaneous	More Brazil fans have followed the 2022 World Cup than we have.	We do not follow the 2022 world cup the way Brazil does.

Note. Texts in italics highlight part of the paraphrase that is crucial to determine the label. CI = comparative illusion.
[a] Participants assigned different interpretations to "we": Some treated "we" as referring to groups of people that have the opposite characteristics to the matrix subject (e.g., Americans vs. non-Americans); some interpreted "we" as belonging to the same group as the matrix subject; some just stuck with "we." It remains to understand why there were different interpretations of the first-person pronoun.

the comparative structure *more … than* was (partially) dropped. "Ungrammatical/nonsense" means the sentence could not be interpreted due to an ungrammatical error or a semantic incoherence, nor did the sentence preserve the same structure, which made it qualified for the category of "no change." "Notice of weirdness" means in the paraphrase participants explicitly mentioned that the original material was weird. "Miscellaneous" encompasses all the other kinds of paraphrases.

Figure 5 displays the distribution of each paraphrase type for singular versus plural than-clause subjects. For both conditions, the event comparison interpretation was ranked the first, followed by "event negation" and "individual comparison." Figure 6 represents

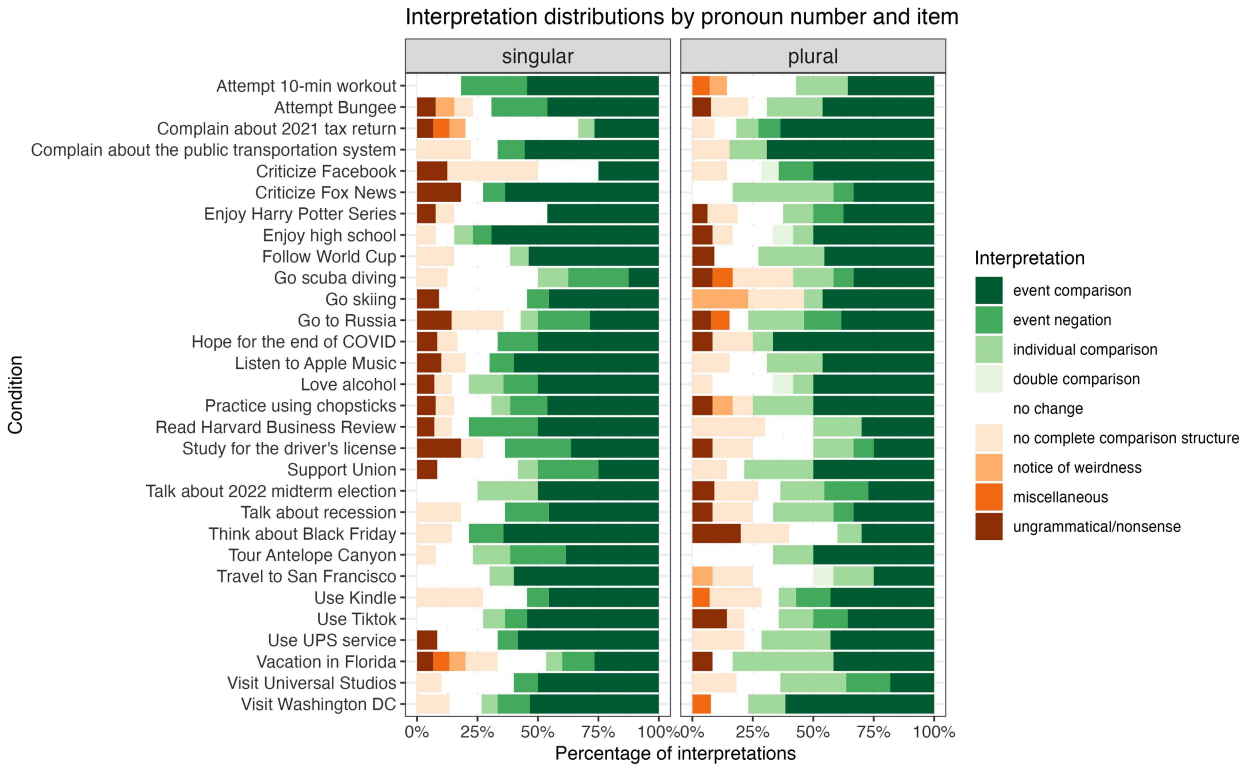
Figure 5
Percentage of Paraphrase Categories by the Number of the Than-Clause Subject in Experiment 2



Note. The figure displays the percentage of each paraphrase category for comparative illusion sentences that differ by *I*/*we*. The green bars represent the plausible paraphrases, and the orange bars represent the implausible ones. The white bar represents the paraphrases that are the same as comparative illusion sentences. Plausible paraphrases take up more than 50%. See the online article for the color version of this figure.

Figure 6

The Distribution of Paraphrase Categories by the Number of the Than-Clause Subject and Item



Note. The figure shows the distribution of paraphrase categories by the number feature of the than-clause subject in comparative illusion sentences and across the 30 items. The green bars represent the plausible paraphrases, the orange bars represent the implausible ones, and the white bar represents the "no change" paraphrase. UPS = United Parcel Service. See the online article for the color version of this figure.

the distribution of by-item paraphrase categories. The prevalence of the "event comparison" is robust across items.

Discussion

Experiment 2 consisted of a paraphrase task that explored English speakers' interpretation of CI sentences. The results confirm the existence of multiple plausible interpretations and demonstrate that the event comparison interpretation is the most prevalent among others. This broad range of interpretations empirically supports intuitions from the literature. It aligns with the intuitions in the sentence template blending account (Townsend & Bever, 2001), the repair analysis account (O'Connor, 2015), and the event interpretation account (Wellwood et al., 2018) in showing that comprehenders reach an alternative interpretation, which is different from the perceptual CI sentence. Using a different method compared to the recall task in Wellwood et al. (2018), Experiment 2 confirms that the comparison of events is indeed a common interpretation of the CI sentence. More than 60% of the paraphrases are sensible, which could be viewed as aligning with the behavioral patterns in Paape (2024) where more than 50% of the CI trials received a "get it" response. The 20% "no change" responses probably result from these participants' failure to notice the anomaly, perhaps due to these participants' inattentiveness to the task (eight participants out of 49 had more than 50% of their trials fall into the "no change" category; see Supplemental Figure S3 for the paraphrase distribution across participants).

Because the event comparison interpretation, the individual comparison interpretation, and the event negation interpretation were the most common plausible paraphrases, we take these three as the plausible interpretations (i.e., s_i) for the CI sentence.

Interestingly, the interpretation distribution in English differs from the distributions in Danish (Christensen, 2016). In English, the event comparison interpretation had a higher percentage than the individual comparison and the event negation interpretations. In Danish, the order was reported to be the opposite (i.e., event negation > individual comparison > event comparison). Of course, the morphosyntactic characteristics of the comparative structure in Danish are different from those in English. In English, *more* could be used both as the comparative for the quantifier *many* and as the comparative for the adverb *much*. But in Danish (13), *flere* ("more") is unambiguously used as the comparative for the nominal quantifier *mange* ("many"), while *mere* ("more") is the adverbial comparative for *meget* ("much").

(13a) Flere folk har været i Rusland end jeg har. (illusion)

More people have been in Russia than I have

"More people have been to Russia than I have."

(13b) Folk har været I Rusland mere end jeg har

People have been in Russia more than I have

"People have been to Russia more than I have."

The difference in the syntactic distribution of *more* in Danish and English might modulate the distribution of possible interpretations of the CI sentence. It is therefore not possible to transfer the availability of these readings from Danish to English.

Experiment 3

The noisy-channel theory predicts that the posterior probability of inferring a certain interpretation given the perceived CI sentence is proportional to how likely that interpretation is to be produced into the CI sentence. In Experiment 2, three prominent paraphrases of the CI sentences emerged: "event comparison," "event negation," and "individual comparison" interpretations. Experiment 3 consisted of a forced-choice comprehension task, where each participant was presented with an illusion sentence and all three interpretations. They were asked to choose the interpretation that best matches their initial interpretation of the CI sentence. Here, the illusion sentence is treated as s_p and the three interpretations as s_i. The percentage of each interpretation approximates a distribution of s_is given s_p. We took this distribution as a proxy for the distribution of $P(s_i|s_p)$ in that the ranking of the percentage monotonously correlates with the posterior probability.

Method

Participants

Sixty participants who self-identified as native English speakers finished the task from https://Prolific.com. Each was paid $6 for their participation. We analyzed those who successfully completed the five English sentence completion trials, who finished at least 90% of all trials, who completed the comprehension questions correctly more than 75% of the time, and who self-identified as from the United States. Fifty participants were included in the final analysis.

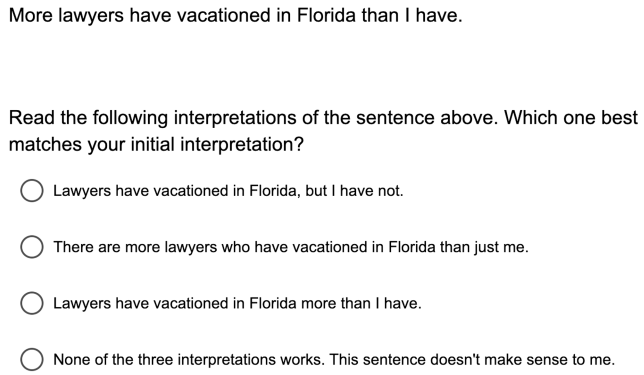
Materials and Procedure

To get a behavioral proxy for the posterior distribution, we used a forced-choice task. We presented the illusion sentence together with the three interpretations. The experimental instruction was as follows: "Read the following interpretations of the sentence above. Which one best matches your initial interpretation?" In addition to the three interpretations, there was also a fourth option: "None of the three interpretations works. This sentence doesn't make sense to me."

The 30 critical items were the same as items in Experiment 2, where the verb phrase was repeatable and the number of the than-clause subject varied between *I* and *we*. Figure 7 shows one example trial. Each of the three interpretations had a consistent syntactic structure across all items. The "event comparison" interpretation treated *more* as an adverb (e.g., *Lawyers have vacationed in Florida more than I/we have*). The "individual comparison" interpretation featured a "there be" construction; *more* modified the cardinality of the subject group; the following "than just me/us" highlighted the reading that there are more people who have experienced the event than just the speaker or the group that the speaker belongs to (e.g., *There are more lawyers who have vacationed in Florida than just me/us*). The "event negation" interpretation used the coordinator *but* to connect two clauses and expressed that the speaker (group) has not experienced the event at stake (e.g., *Lawyers have vacationed in Florida, but I have not*).

Figure 7

An Example Trial for Experiment 3



Note. The interpretations were in the order of "event negation," "individual comparison," "event comparison," and nonsense. In an actual trial, the order of the first three interpretations was randomized.

There were 60 filler items where the sentences to be interpreted were all grammatical and plausible. Each sentence corresponded with four interpretation choices, three of which featured "event comparison," "individual comparison," and "event negation," respectively, and the last one being the nonsense option. Out of the 60 sentences, 20 had the expected answer as "event comparison" (e.g., *College students have been listening to podcasts more than I have*); 20 had "event negation" (e.g., *Youtubers have earned money with their videos but we have not*); 20 had "individual comparison" (e.g., *More computer scientists than philosophers have collaborated with colleagues*). Within the "event comparison" and "event negation" groups, 10 items had the first-person pronoun *I* as the than-clause subject and 10 items had *we*. Within the "individual comparison" group, both the matrix subject and the than-clause subject were plural noun phrases. The syntactic templates for the four options were the same as those for the critical items.

Each participant finished a randomized list of 90 forced-choice trials. The order of the three meaningful choices was randomized within a trial. At the beginning of the task, participants performed five English sentence completion trials. We used these trials to filter out participants whose English skills were not fluent.

Results

We considered answers to the filler items as comprehension checks. In theory, each filler sentence had a unique answer. For example, the sentence *College students have been listening to podcasts more than I have* should be interpreted as "event comparison" and participants should select the choice "college students have been listening to podcasts more than I have" (in the exact structure). Indeed, for this item, 41 participants out of 60 chose the "event comparison" option (three chose "event negation," four chose "individual comparison," and 12 chose "nonsense"). However, participants often unexpectedly chose the "event comparison" option for fillers with an intended "individual comparison" interpretation. For example, for the item *More senior citizens than young adults have*

undergone surgeries, 18 out of 60 participants chose the "event comparison" option while 41 of them chose the "individual comparison" option. The former group might interpret this item as "senior citizens have undergone surgeries more times or more frequently than young adults have"; the supposed comparison between the cardinalities of the two groups was coerced into a comparison of the frequency/times of events. This could make sense given the following context: 60 out of 100 senior citizens have undergone surgeries while only 30 out of 100 young adults have done so. On average, one senior citizen has undergone surgeries for 0.6 times, compared to 0.3 times for an average young adult. Therefore, it seems reasonable for the "individual comparison" fillers to have an event comparison interpretation.

For each of the 20 "individual comparison" fillers, an average of 19 participants out of 60 chose the "event comparison" option, compared to an average of 36 participants for the "individual comparison" option. Given that participants only confused these two choices for this specific "individual comparison" condition and they overwhelmingly chose the expected interpretation for the other fillers (see Supplemental Figure S7 for the choice distribution across filler conditions), we adopted a softer exclusion criterion by treating the choice "event comparison" for the "individual comparison" filler condition as acceptable. Applying the 75% threshold of comprehension question accuracy, 50 participants remained in the analysis.[6]

As a result, responses from 1,498 trials were entered into the analysis. Figure 8 represents the percentage of different interpretations for singular and plural first-person pronoun than-clause subjects. In both conditions, the event comparison interpretation was the most predominant, followed by individual comparison and event negation.

To statistically investigate the ranking of the four interpretations and whether the number feature of the than-clause subject played a role, we fitted the forced-choice data into a Bayesian multilevel multinomial regression model. The response variable was the four categories of interpretations ("event comparison," "event negation," "individual comparison," and "no sense," reference = "event comparison"). The independent variable was the dummy-coded categorical variable representing the number feature of the than-clause subject (reference level = "singular"). The random effects included a random intercept and a random slope for the number feature condition for both items and participants.

In the model specification, the model family was set to be categorical, which was specifically used when the response variable was categorical with more than two levels. There were four sampling chains each with 2,000 iterations to generate the posterior distributions of the estimated parameters. The prior setting was the default in *brms*. All $\hat{R}$s were 1.0, indicating successful convergence. We relied on Bürkner (2024) to interpret the model output. While the reference response category was event comparison, taken as φ, each of the other category was κ and η_κ represents the linear equation. Then the exponentiated e^{η_κ} marks the ratio between the probability of the response variable being κ, $P(y = \kappa)$, and the probability of the reference category φ, $P(y = \varphi)$.

First of all, there were substantial differences between the probability of the event comparison interpretation and each of the other three interpretations. In the singular pronoun condition, the probability of choosing the individual comparison interpretation was 0.06 times the probability of choosing the event comparison interpretation ($\beta = -2.89$, 95% CrI [−3.91, −2.01]). The probability of choosing the event negation interpretation was 0.005 times that of choosing the baseline event comparison group ($\beta = -5.21$, 95% CrI [−7.15, −3.76]). The probability of choosing nonsense was 0.02 times that of event comparison ($\beta = -4.19$, 95% CrI [−5.86, −2.83]). As for the effect of a plural than-clause subject, the results indicate that it reduced the probability of event negation ($\beta = -2.63$, 95% CrI [−7.49, −0.10]) compared to its probability in the singular subject condition. This is not the case for the other two types of responses (plural and individual comparison: $\beta = 0.05$, 95% CrI [−0.52, 0.70]; plural and no sense: $\beta = -0.18$, 95% CrI [−1.23, 0.66]). Even so, the relative order of the three interpretations in the plural subject condition does not differ from that in the singular subject condition.

Discussion

Experiment 3 used a forced-choice selection task that collected English speakers' choices over the most salient interpretation of the CI sentence out of three plausible alternatives, approximating a behavioral measurement for the posterior probability $P(s_i|s_p)$. We found that the "event comparison" interpretation was by far the most dominant, followed by "individual comparison" and then "event negation." The plurality of the pronoun than-clause subjects did not affect the preference ranking of the three interpretations.

Experiment 4

Experiment 4 explores the possibility that the comprehension patterns we observed in previous experiments are results of noisy-channel interpretation processes. During this process, comprehenders take the CI sentence as linguistic input that might contain errors and rationally infer its intended meaning by estimating how likely a certain intended meaning can be linguistically produced to a CI sentence. That is, speakers might intend to convey a message involving a comparative meaning but end up producing the syntactically acceptable but semantically ill-formed CI sentence. In turn, comprehenders infer the intended sentence given the perceived CI sentence based on their estimation of the likelihood of production error.

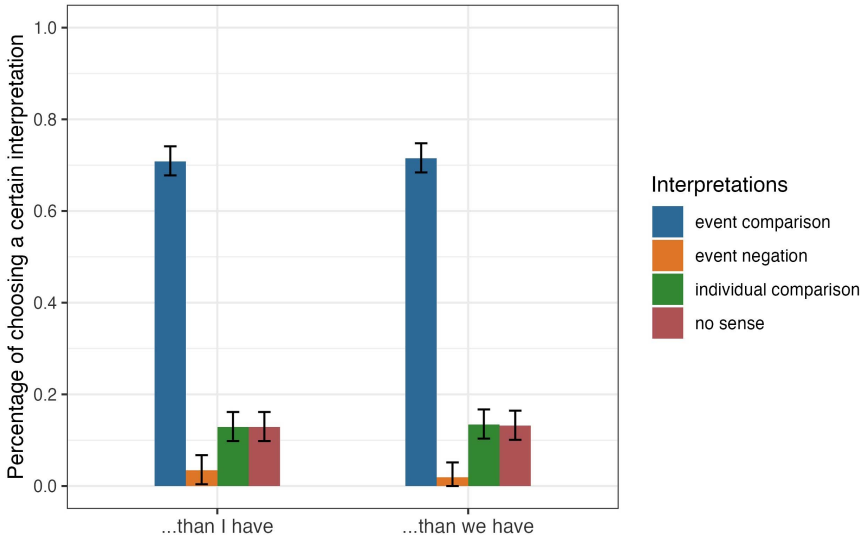
Based on Experiments 2 and 3, we hypothesize three profiles of noise corruption that transform an intended sentence to the CI sentence. They are labeled as "shifting *more*," "than-clause substitution," and "negation-to-comparative substitution" as seen in (13), with each locus of possible error underlined.[7]

[6] Fillers that were expected to be interpreted as "event comparison" did not trigger confusion—for each "event comparison" filler, an average of 51 out of 60 participants chose the "event comparison" option and the rest of the participants chose other wrong options at an equal chance. If we adopt a stricter criterion, such as choosing "event comparison" for "individual comparison" fillers was treated as incorrect, then 36 participants remained. The percentage distribution in Figure 8 for 50 participants and the alternative for 36 participants in Supplemental Figure S8 are almost identical, with similar statistical results. Please see Supplemental Material Section 4 for an analysis based on the stricter criterion.

[7] The intended sentences are different from the experimental materials in Experiment 3. In Experiment 3, the intended interpretations were more structurally distant from the CI sentence, whereas in Experiment 4, the intended sentences were made to approximate the form of the CI sentence as much as possible. This is because in Experiment 3, we aimed to motivate participants to distinguish different interpretations by recognizing different structures, and in Experiment 4, we aimed to prevent additional noncritical words in the intended sentence from reducing the production error likelihood.

Figure 8

The Percentage Distribution of Choosing Different Interpretations for Singular and Plural First-Person Pronoun Than-Clause Subjects



Note. The error bars represent the 95% confidence intervals calculated by the *MultinomialCI* package (Version 1.2) in R adapted from Sison and Glaz (1995). See the online article for the color version of this figure.

(13a) Shifting *more*—"event comparison"

Intended: Lawyers have vacationed in Florida more than I have.

Produced: More lawyers have vacationed in Florida than I have.

(13b) Than-clause substitution—"individual comparison"

Intended: More lawyers have vacationed in Florida than just me.

Produced: More lawyers have vacationed in Florida than I have.

(13c) Negation-to-comparative substitution—"event negation"

Intended: More lawyers have vacationed in Florida but I haven't.

Produced: More lawyers have vacationed in Florida than I have.

The intended "shifting *more*" sentence (13a) features a shift of *more* in its syntactic position—the adverb *more* in the comparative structure is produced as a noun phrase modifier at the beginning of the sentence. Language production errors in the form of word shifts are frequently reported and characterized by function words such as quantifiers and adverbs (Bock, 2011; Garrett, 1975, 1980). For example, a shift error (together with a morphological error) involving a comparative structure appears in Fromkin (1971) where the intended sentence *He's a far better man than anyone here* is produced as *He's a farther man than anyone better here*. Both the than-clause substitution (13b) and the negation-to-comparative substitution (13c) sentences involve substitutions, which are prevalent speech errors (Bock, 2011; Fay, 1981; Stemberger, 1982). In (13b), *just me* is substituted by *I have*, which forms a syntactic ellipsis. In (13c), the contrastive structure *but I haven't* is substituted by *than I have*.

Research in language production and noisy-channel theory has shown that production errors are more likely to happen when (a) there are few edits involved; (b) given a substitution, the substituting part is more frequent than the segment to be substituted; and (c) the intended and produced structures are similar at phonological, syntactic, and semantic levels (Dell & Reich, 1981; Gibson et al., 2013; Harley & MacAndre, 2001; Kittredge et al., 2008; Koranda et al., 2022; Zhang, Ryskin, & Gibson, 2023). Out of the three intended sentences, the "shifting *more*" condition involves the simplest edit—a position exchange of the same word *more*—while preserving the comparative structure, compared to the other two conditions where the substituting words change the syntactic structure more drastically. The than-clause substitution condition replaces *just me* by *I have*, involving two substitutions; the negation-to-comparative substitution condition replaces *but I haven't* with *than I have*, involving substitutions of two words as well.

A corpus search based on the Corpus of Contemporary American English (Davies, 2008) with 1 billion word tokens revealed that sentences with the structure *more … than* outnumbered those with *more than* and *more … but*. In the corpus analysis, we used the software package spaCy (https://spacy.io/) to tokenize and parse the corpus (1,179,317,004 tokens) into 76,352,408 sentences. We combined keyword search and the dependency structure of sentences to filter target sentences with the structure *more … than*, *more than*, and *more … but*. One predominant unwanted structure (e.g.,

more than five people came where *more than* modifies the quantity) was deleted. We ended up with 222,873 sentences with the structure *more … than* (0.29%), 104,241 sentences with the structure *more than* (0.14%), and 4,489 sentences with the structure *more … but* (5.88e−03%). This suggests that it is plausible that a speaker might intend to produce a *more than* structure but be attracted to produce the more frequent *more … than* structure.

We therefore predict that the intended sentence in the "shifting *more*" condition is more likely to be produced into the CI sentence. Since it is hard to find this exact error in corpora or to cause people to produce such errors in the lab, we use comprehenders' judgments on language production in a noise likelihood rating task to collect an approximant for the noise likelihood term.

Method

Participants

Sixty-three participants were recruited from https://Prolific.com. Each received a $6 payment for their participation. We analyzed those who self-identified as native speakers of English from the United States, whose five initial English completion trials were grammatical and plausible, who passed the comprehension check questions more than 75% of the time for all relevant trials, and who finished at least 90% of all trials. Fifty-nine participants contributed to the final analysis.

Materials and Procedure

The noise likelihood rating task followed the design in Zhang, Ryskin, and Gibson (2023). Participants first read the following instruction:

> People make speech errors all the time when they intend to convey ideas in spoken sentences, especially when they are distracted or speaking fast. These errors include but are not limited to deletions, insertions, exchanges, and substitutions of certain words and structures. In the following task, you are required to read the intended sentence, think about its meaning and then think about how likely it is that someone would say the produced sentence when they meant to say the intended sentence, and respond using the rating scale.

Then participants were given a pair of sentences (the first sentence was labeled "intended" and the second "produced") and were asked how likely it is for a speaker to utter the "produced" sentence. For each sentence pair, the participants gave their rating on a fully labeled 7-point Likert scale (1 = *extremely unlikely*, 2 = *strongly unlikely*, 3 = *somewhat unlikely*, 4 = *intermediate*, 5 = *somewhat likely*, 6 = *strongly likely*, 7 = *extremely likely*; see Supplemental Figure S5 for a visualization of an example trial on the screen).

There were 30 critical items in this study. Each item featured three pairs of sentences corresponding to three conditions: "shifting *more*," "than-clause substitution," and "negation-to-comparative substitution." The produced sentence remained the same across conditions. The critical materials also differed in whether the than-clause subject was a singular or plural first-person pronoun. We chose a joint presentation of the three pairs of sentences on the experiment screen, following Marty et al. (2020), to increase the sensitivity of this rating task.

There were 60 filler items. Twenty were depth-charge illusion sentences taken from Zhang, Ryskin, and Gibson (2023), 10 were simple sentences in active voice, 10 were simple sentences in passive voice, 10 were double object constructions, and 10 were ditransitive sentences with a prepositional object. The latter 40 items were taken from Gibson et al. (2013). Each filler had three conditions where the intended sentence deviated from the produced sentence in three ways (see Supplemental Material Section 5 for details of the filler design). We designed a comprehension check in the 40 simple-sentence fillers. There, while two conditions had simple edits from the intended sentence to the produced sentence, the third condition had an impossible intended sentence (e.g., *The man held the woman* as the intended sentence for the produced *The ball kicked the girl*) and participants were expected to choose "extremely unlikely."

Participants read a randomized list of all 90 items. Within each item, the order of three sentence pairs was randomized. Once an item was completed, participants could not return to it. As in the other surveys in this article, participants also performed five sentence completion trials at the beginning of the study, which helped us filter out nonfluent speakers of English.

Results

Figure 9 displays the distribution of the noise likelihood rating for the three conditions by the than-clause subject number feature, after the exclusion of inattentive participants. A visual inspection suggests that the "shifting *more*" condition was the one that participants thought was the most likely to be produced into a CI, among the three options, followed by the "than-clause substitution" condition and the "negation-to-comparative substitution" condition. The pattern is the same in both singular and plural versions (see Supplemental Figure S6 for a different visualization).

In order to evaluate our hypotheses quantitatively, we fitted a Bayesian multilevel cumulative ordinal model. The raw rating was taken as the dependent variable; the number feature of the than-clause subject (reference = singular) and the noise profiles (reference = shifting *more*) were taken as fixed effects; no interaction effects were included based on leave-one-out cross-validation. The random effects included random intercepts and random slopes of both the number feature and the noise profile condition for subjects and items. The priors and other model parameters were set as the same as those in Experiment 1.

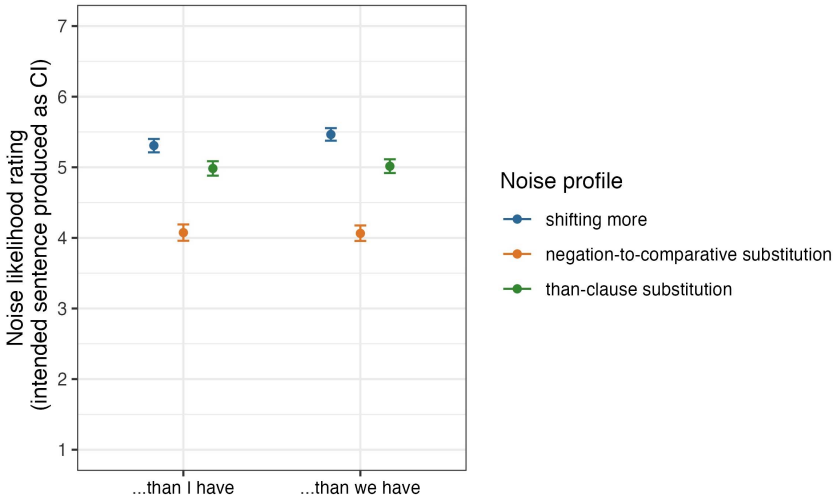
The results show that participants thought that it would be more likely for the intended event comparison sentence to shift *more* to sentence-initial position to form a CI sentence than for an event negation sentence to substitute the contrastive clause to a than-clause ($\beta = -1.28$, 95% CrI [−1.63, −0.94]). Similarly, participants thought that it would be more likely for the intended event comparison sentence to shift *more* than for the intended individual comparison sentence to substitute the than-clause part ($\beta = -0.39$, 95% CrI [−0.74, −0.02]). There was no or small main effect of the than-clause subject number feature, or the effect was small and uncertain ($\beta = 0.03$, 95% CrI [−0.06, 0.12]).

Discussion

In Experiment 4, we constructed a noise likelihood rating experiment to approximate how likely comprehenders think others would produce the CI sentence given each of three intended sentences. The results show that the intended sentence with the event comparison interpretation (e.g., *Students have been to Russia more*

Figure 9

Noise Likelihood Ratings for the Three Types of Intended Sentences to Be Produced as the Comparative Illusion Sentence



Note. The error bars represent the 95% bootstrapped confidence intervals. CI = comparative illusion. See the online article for the color version of this figure.

than I have) with an underlying "shifting *more*" operation was the most likely, compared to (a) the individual comparison sentence (e.g., *More students have been to Russia than just me*) that substitutes the *than* clause structure and (b) the event negation sentence (e.g., *Students have been to Russia but I haven't*) that substitutes the entire contrastive clause. The number feature of the than-clause subject pronoun (i.e., *I* vs. *we*) did not affect this pattern. This result relies on comprehenders' judgment on how people in general produce sentences and is consistent with our prediction that more structural edits lead to a lower likelihood rate.

Taking this behavioral measurement as a proxy for the noise likelihood term, $P(s_p|s_i)$, and the percentage of choice from Experiment 3 as a proxy for the posterior probability, $P(s_i|s_p)$, we show that for a certain plausible interpretation of the CI sentence, it undergoes specific noise operations; the fewer and simpler operations are involved, the higher the likelihood of the language production error, and the more probable the corresponding intended sentence is to be inferred given the CI sentence.

Transparency and Openness

We report all data exclusion criteria, all experimental manipulations, and all measures in the study, and we strictly follow the Journal Article Reporting Standards (Kazak, 2018). All the experimental materials, the collected data, and the analysis scripts can be accessed via the Open Science Framework platform (https://osf.io/89d5b/). Experimental materials in Experiment 1 can be found in Supplemental Material Section 1. Data were analyzed using R Version 4.4.2 (R Core Team, 2024) and Python Version 3.9 (van Rossum & Drake, 2009).

General Discussion

Through four experiments, we provide a noisy-channel explanation for why comprehenders initially regard the semantically incoherent sentence *More people have been to Russia than I have* as acceptable, a phenomenon called "comparative illusion." According to the noisy-channel framework of language processing (Gibson et al., 2013; Levy, 2008; Shannon, 1948), we model the anomalous sentence as a string whose form is corrupted during transmission and thus deviates from the intended message. We hypothesize that during language comprehension, comprehenders would rationally infer an intended message by postulating multiple possible interpretations of this signal and choosing the one that is most likely to be produced and transformed into the target string.

First, Experiment 1 identified the sentence structures that gave rise to the strongest CI effect among variations that have been studied in previous research. These were sentences with a pronoun in the *than* clause such as *More students have been to Russia than I/we have*. In Experiment 2, we used a paraphrase task to map out the likely interpretations of the CI sentences. Our manual annotation revealed three prevalent readings: (a) comparing the times/frequency of events (e.g., "People have been to Russia more than I have," which we labeled as "event comparison"), (b) comparing the cardinality of the two groups of individuals who participated in that event (e.g., "There are more people who have been to Russia than just me," labeled as "individual comparison"), and (c) denying the participation of the event by the than-clause subject (e.g., "Students have been to Russia but I have not," labeled as "event negation"). In Experiment 3, we used a forced-choice selection task where participants were asked to choose one interpretation out of the three that best matched their initial understanding of the CI sentence. The "event comparison" interpretation was the most probable one, followed by "individual comparison" and "event negation." We took the percentage distribution among the three interpretations as a behavioral proxy for the posterior probability $P(s_i|s_p)$. In Experiment 4, we hypothesized that comprehenders model the CI sentence as the result of a production error and each of the three intended sentences can be falsely produced to the CI sentence via unique noise operations.

The three noise operations are "shifting *more*," "than-clause substitution," and "negation-to-comparative transformation," which correspond to the intended sentences with an event comparison interpretation, an individual comparison interpretation, and an event negation interpretation, respectively. We used a noise likelihood rating task asking participants to rate the likelihood of each noise operation. The results show that the "shifting *more*" condition was the most likely, followed by the "than-clause substitution" and the "negation-to-comparative transformation" conditions. We took the noise likelihood rating as a behavioral proxy for the noise likelihood term $P(s_p|s_i)$. Assuming that any differences in prior probabilities of these three intended sentences are outweighed by their differences in the noise likelihood, the behavioral data are consistent with the noisy-channel hypothesis that the posterior probability of $P(s_i|s_p)$ is proportional to the noise likelihood $P(s_p|s_i)$, suggesting that the comprehension of CI sentences is rational inference over a noisy channel (Gibson et al., 2013; Levy, 2008).

The rational noisy-channel account of the CI synthesizes and complements previous accounts with concrete predictions supported by empirical data from our experiments. First, the existence of multiple possible interpretations revealed in Experiment 2 aligns with the intuitions of the repair analysis account (O'Connor, 2015). It also incorporates ideas from the sentence template blending account (Townsend & Bever, 2001), specifically the hypothesizing of plausible alternatives, even though our account assumes that probabilistic inference places a distribution over these alternatives while Townsend and Bever assume these templates are blended in comprehenders' minds. Second, the overall preference for the event comparison interpretation in Experiments 2 and 3, coupled with the high likelihood of shifting *more* in Experiment 4, supports the event comparison intuition in O'Connor (2015) and Wellwood et al. (2018). We used a different method compared to the recall task in Wellwood et al. (2018) and confirmed that moving *more* is indeed possible in the comprehension of CI sentences. The unique contributions of our approach are as follows: (a) We searched for possible interpretations of the CI sentence exhaustively in a bottom-up experimental approach, avoiding potential interpretation biases from the researchers; (b) we also showed that the likelihood of noise edits through which the intended sentence is linguistically represented as the perceived sentence quantitatively predicts the posterior probability for a comprehender to derive that intended sentence from the CI sentence during language comprehension.

Our experimental data are also potentially compatible with the "good enough" processing account (Ferreira et al., 2002; Ferreira & Patson, 2007; Karimi & Ferreira, 2016). On the one hand, according to the "good enough" account, comprehenders make use of heuristics to achieve a "good enough" interpretation of a complicated sentence and may not realize the discrepancy between their interpretation and the sentence's literal meaning. In the CI, one possible heuristic could be the prevalent adverbial use of *more* drives the event comparison interpretation. However, the "good enough" account does not specify how comprehenders prefer one interpretation over other alternatives of the perceived string. On the other hand, one could also argue that comprehenders' initial interpretation is underspecified—they only vaguely register a comparative meaning but do not go deeper to analyze what information is being compared. Nevertheless, it is worth noting that the noisy-channel theory might be considered as a formally precise rendering of one version of a "good enough" account: The original article introducing the idea speculates that "good enough" processing leads to a "reanalyzed structure [that] does not match the input string" (Christianson et al., 2001, p. 396) but is nevertheless consistent with principles of grammatical structure and semantic interpretation. Future studies could investigate these issues in greater depth—for example, future experiments could allow comprehenders to choose multiple interpretations and provide confidence ratings for each choice, instead of the forced choice design in Experiment 3, to investigate the extent to which multiple interpretations coexist.

There are several directions for future work. First, our noisy-channel predictions need to incorporate the role of priors in the explanation of the comprehension patterns. While it could be true that these three interpretations might share similar prior probabilities $P(s_i)$, there could of course be differences in priors across different items. For example, Figure 6 shows that for the CI sentence *More girls have gone scuba diving than I have*, the event negation interpretation ("I haven't gone scuba diving") was the most common. This might be because fewer of our experimental participants have gone scuba diving compared to activities such as taking a vacation in Florida. Our current research assumes that any systematic differences in the prior probabilities of the alternative interpretations are outweighed by differences in the likelihood. Future research could investigate whether this assumption holds stable across different items and whether such differences are reflected in CI interpretation probabilities.

Second, our focus is on materials with pronouns as the subject of the *than* clause. It remains to be better understood why materials with pronouns are more likely to be misinterpreted than the ones with full noun phrases in the *than* clause, such as *More lawyers have vacationed in Florida than the clerk has* (see Zhang et al., 2025, for a preliminary explanation). One possibility is that in everyday language use, people are more likely to compare themselves with an observation than compare two groups of people. If there is a higher frequency of first-person pronouns in comparative structures, comprehenders might be more likely to be attracted by this common use and consider the CI sentence as plausible. Future studies could conduct corpus analyses for investigation.

Third, our experiments that investigate the noise model are limited in that they asked participants what they think other people might produce in certain situations, but these experiments do not investigate the cognitive processes of actual language production. It would be valuable in future work to investigate language production in order to explore (a) the likelihood that people utter the CI structure in the first place and (b) the frequency of the three noise profiles that transform the intended meaning to the CI sentence. For example, future studies might utilize the retyping method (Ryskin et al., 2018) to investigate whether errorful production of CI sentences can be elicited in a more direct way.

Another future direction that could provide a more complete understanding of such materials involves an understanding of how such interpretations are arrived at online. Besides, it is important to map out the cross-linguistic landscape of CI. It is already a puzzle why the verb repeatability effect was not found in Danish and Swedish data (Christensen, 2016). It might be due to the morphosyntactic distinctiveness of the nominal comparative morpheme and the adverbial comparative morpheme in these languages, but more rigorous testing across languages is needed.

Last, with the recent advancement of pretrained large language models (LLMs), the CI sentences have been used to test their

linguistic abilities (Dentella et al., 2023; Hu et al., 2024; Zhang, Gibson, & Davis, 2023). These studies show that (a) LLMs can distinguish grammatical from ungrammatical sentences when the than-clause subject features a noun phrase and (b) CI sentences with plural noun phrases receive higher probabilities than those with a singular noun phrase. These are impressive results that align with human beings. But researchers have not yet tested whether LLMs will experience an illusion effect with the canonical comparative sentence in (1), or probed LLMs' paraphrases for comparative structures (although it has been shown that prompt-based evaluation tends to deviate from LLMs' real linguistic capacities; Hu & Levy, 2023; Kauf et al., 2024). Knowing whether LLMs can be tricked by the most canonical CI, or whether they can produce plausible interpretations via inferences, would provide unique insights into whether they can be investigated against the hypothesis of being cognitive models of language processing (Zhang, Gibson, & Davis, 2023).

Overall, the noisy-channel approach offers a candidate explanation for the CI. Together with existing work that applies the same approach to explain the depth-charge illusion (Zhang, Ryskin, & Gibson, 2023)[8] and the negative polarity illusion (Zhang & Gibson, 2024), this work supports the noisy-channel approach as a promising approach to explain language comprehension phenomena given uncertain linguistic input.

Constraints on Generality

Our study focuses on explaining the illusory comprehension behavior of sentences such as *More people have been to Russia than I have* by neural typical native adult speakers of (American) English. The empirical evidence that supports an illusory experience comes from the acceptability judgment task in Experiment 1 with 30 critical items. The follow-up examination hinges on the replicability of the acceptability patterns in Experiment 1. We only focus on the early phase of the comprehension mechanism—why comprehenders initially find the anomalous sentence acceptable. It remains to be explored whether there are heterogeneous stages of comprehension that could involve error detection or meaning reconstruction. Furthermore, our noisy-channel account focuses on explaining English materials from neural typical native adult speakers in the United States. Future studies are necessary to bring the theory to explain cross-linguistic comprehension patterns or comprehension patterns of different populations.

[8] With this said, we acknowledge that the depth-charge illusion has also received substantial investigation under a superficial, good-enough processing account (Paape, 2023, 2024; Paape et al., 2020).

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bock, K. (2011). How much correction of syntactic errors are there, anyway? *Language and Linguistics Compass*, *5*(6), 322–335. https://doi.org/10.1111/j.1749-818X.2011.00283.x

Bürkner, P.-C. (2024, April 2). *Interpreting results from categorical() with brm—Interfaces/brms* [Computer software]. The Stan Forums. https://discourse.mc-stan.org/t/interpreting-results-from-categorical-with-brm/4120

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Chen, S., Nathaniel, S., Ryskin, R., & Gibson, E. (2023). The effect of context on noisy-channel sentence comprehension. *Cognition*, *238*, Article 105503. https://doi.org/10.1016/j.cognition.2023.105503

Christensen, K. R. (2010). Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex. *Brain and Cognition*, *73*(1), 41–50. https://doi.org/10.1016/j.bandc.2010.02.001

Christensen, K. R. (2016). *The dead ends of language: The (mis)interpretation of a grammatical illusion. Let us have articles betwixt us—Papers in historical and comparative linguistics in honour of Johanna L. Wood*. Aarhus University.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407. https://doi.org/10.1006/cogp.2001.0752

Coren, S., & Girgus, J. (2020). *Seeing is deceiving: The psychology of visual illusions*. Routledge.

Dahl, Ö. (1981). On the definition of the telic-atelic (bounded-nonbounded) distinction. In P. Tedeschi & A. Zaenen (Eds.), *Tense and aspect* (pp. 79–90). Brill. https://doi.org/10.1163/9789004373112_006

Davies, M. (2008). *Corpus of Contemporary American English (COCA)* [Data set]. https://doi.org/10.7910/DVN/AMUDUW

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 611–629. https://doi.org/10.1016/S0022-5371(81)90202-4

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(51), Article e2309583120. https://doi.org/10.1073/pnas.2309583120

Fay, D. (1981). Substitutions and splices: A study of sentence blends. *Linguistics*, *19*(7–8), 717–750. https://doi.org/10.1515/ling.1981.19.7-8.717

Fedorenko, E., Ryskin, R., & Gibson, E. (2023). Agrammatic output in nonfluent, including Broca's, aphasia as a rational behavior. *Aphasiology*, *37*(12), 1981–2000. https://doi.org/10.1080/02687038.2022.2143233

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52. https://doi.org/10.2307/412187

Garey, H. B. (1957). Verbal aspect in French. *Language*, *33*(2), 91–110. https://doi.org/10.2307/410722

Garrett, M. F. (1975). The analysis of sentence production. *Psychology of Learning and Motivation*, *9*, 133–177. https://doi.org/10.1016/S0079-7421(08)60270-4

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production* (Vol. 1, pp. 177–220). Academic Press. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2325810

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United*

States of America, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., & Kiran, S. (2016). A rational inference approach to aphasic language comprehension. *Aphasiology*, *30*(11), 1341–1360. https://doi.org/10.1080/02687038.2015.1111994

Goldberg, A. E., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311. https://doi.org/10.1016/j.tics.2022.01.005

Gregory, R. L. (1968). Visual illusions. *Scientific American*, *219*(5), 66–76. https://doi.org/10.1038/scientificamerican1168-66

Harley, T. A., & MacAndre, S. B. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research*, *30*(4), 395–418. https://doi.org/10.1023/A:1010421724343

Hu, J., & Levy, R. (2023). *Prompting is not a substitute for probability measurements in large language models*. arXiv preprint. https://doi.org/10.18653/v1/2023.emnlp-main.306

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(36), Article e2400917121. https://doi.org/10.1073/pnas.2400917121

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040. https://doi.org/10.1080/17470218.2015.1053951

Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E., & Ivanova, A. A. (2024). *Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models*. arXiv preprint. https://doi.org/10.48550/arXiv.2403.14859

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kelley, P. (2018). *More people understand Eschers than the linguist does: The causes and effects of grammatical illusions* [PhD thesis, Michigan State University]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/2041968142/abstract/4F76E1837E1F488CPQ/1

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463–492. https://doi.org/10.1080/02643290701674851

Koranda, M. J., Zettersten, M., & MacDonald, M. C. (2022). Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science*, *33*(9), 1440–1451. https://doi.org/10.1177/09567976221089603

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Langsford, S., Stephens, R. G., Dunn, J. C., & 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*, Article 2886. https://doi.org/10.3389/fpsyg.2019.02886

Leivada, E. (2020). Language processing at its trickiest: Grammatical illusions and heuristics of judgment. *Languages*, *5*(3), Article 29. https://doi.org/10.3390/languages5030029

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the conference on empirical methods in natural language processing (EMNLP '08)* (pp. 234–243). Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613749

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. *Proceedings of the 49th annual meeting of the Association for Computational Linguistics* (pp. 1055–1065). Association for Computational Linguistics. https://aclanthology.org/P11-1106/

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21086–21090. https://doi.org/10.1073/pnas.0907664106

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2020, September 3–5). *Structural frequency effects in noisy-channel comprehension* [Conference session]. The 26th Architectures and Mechanisms for Language Processing Conference. https://tedlab.mit.edu/tedlab_website/researchpapers/Liu_Ryskin_Futrell_Gibson_Penn_Ling_2021.pdf

Marty, P., Chemla, E., & Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics*, *5*(1), Article 72. https://doi.org/10.5334/gjgl.980

Montalbetti, M. M. (1984). *After binding: On the interpretation of pronouns* [PhD thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/15222

Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A., & Bürkner, P. C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nicenboim, B., Schad, D., & Vasishth, S. (2021). *Introduction to Bayesian data analysis for cognitive science*. Chapman and Hall/CRC Press.

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

O'Connor, E., Pancheva, R., & Kaiser, E. (2013). Evidence for online repair of Escher sentences. *Proceedings of Sinn und Bedeutung*, *17*, 363–380. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/350

Paape, D. (2023). The role of incremental and superficial processing in the depth charge illusion: Experimental and modeling evidence. *Journal of Semantics*, *40*(1), 93–125. https://doi.org/10.1093/jos/ffad003

Paape, D. (2024). How do linguistic illusions arise? Rational inference and good-enough processing as competing latent processes within individuals. *Language, Cognition and Neuroscience*, *39*(10), 1334–1365. https://doi.org/10.1080/23273798.2024.2387226

Paape, D., Vasishth, S., & von der Malsburg, T. (2020). Quadruplex negatio invertit? The online processing of depth charge sentences. *Journal of Semantics*, *37*(4), 509–555. https://doi.org/10.1093/jos/ffaa009

Pham, L. (2022). *Acceptability of comparative illusions as a function of interactions between repeatability of a verb phrase and active production task*. University of Amsterdam.

Poliak, M., Kimura, H., & Gibson, E. (2024). Mis-heard lyrics: An ecologically-valid test of noisy channel processing. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46, pp. 3782–3788). Cognitive Science Society.

Poliak, M., Malik-Moraleda, S., & Gibson, E. (2025). Rational language comprehension depends on priors about both meaning and structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001470

Poliak, M., Ryskin, R., Braginsky, M., & Gibson, E. (2024). It is not what you say but how you say it: Evidence from Russian shows robust effects of the structural prior on noisy channel inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(4), 637–649. https://doi.org/10.1037/xlm0001244

Poppels, T., & Levy, R. P. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 378–383). Cognitive Science Society.

Qian, P., & Levy, R. P. (2023). *Comprehenders' error correction mechanisms are finely calibrated to language production statistics*. PsyArXiv. https://doi.org/10.31234/osf.io/e3v5b

R Core Team. (2018). *R: A language and environment for statistical computing* (Version 3.4.4) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Robinson, J. O. (1972). *The psychology of visual illusion*. Hutchinson.

Ryskin, R., Bergen, L., & Gibson, E. (2021). *Agreement errors are predicted by rational inference in sentence processing*. PsyArXiv. https://doi.org/10.31234/osf.io/uaxsq

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

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, Article 107855. https://doi.org/10.1016/j.neuropsychologia.2021.107855

Sanford, A., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386. https://doi.org/10.1016/S1364-6613(02)01958-7

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, *90*(429), 366–369. https://doi.org/10.1080/01621459.1995.10476521

Stemberger, J. P. (1982). Syntactic errors in speech. *Journal of Psycholinguistic Research*, *11*(4), 313–345. https://doi.org/10.1007/BF01067585

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. MIT Press. https://doi.org/10.7551/mitpress/6184.001.0001

van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

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

Wellwood, A., Pancheva, R., Hacquard, V., Fults, S., & Phillips, C. (2009). *The role of event comparison in comparative illusions* [Poster presentation]. The 22nd Annual CUNY Sentence Processing Conference, Davis, CA, United States.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, *35*(3), 543–583. https://doi.org/10.1093/jos/ffy014

Zhan, M., Chen, S., Levy, R., Lu, J., & Gibson, E. (2023). Rational sentence interpretation in mandarin Chinese. *Cognitive Science*, *47*(12), Article e13383. https://doi.org/10.1111/cogs.13383

Zhang, Y. (2024). *The rational processing of language illusions* [Doctoral dissertation, Harvard University Graduate School of Arts and Sciences]. ProQuest Dissertations & Theses. https://www.proquest.com/dissertations-theses/rational-processing-language-illusions/docview/3130530410/se-2

Zhang, Y., & Gibson, E. (2024, May 16–18). *A memory-based account of robust negative polarity illusion effects* [Conference session]. Talk at The 37th Annual Conference on Human Sentence Processing, Ann Arbor, MI, United States. https://doi.org/10.17605/OSF.IO/8WSYA

Zhang, Y., Gibson, E., & Davis, F. (2023). Can language models be tricked by language illusions? Easier with syntax, harder with semantics. *Proceedings of the 27th conference on computational natural language learning (CoNLL)* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.1

Zhang, Y., Ryskin, R., & Gibson, E. (2023). A noisy-channel approach to depth-charge illusions. *Cognition*, *232*, Article 105346. https://doi.org/10.1016/j.cognition.2022.105346

Zhang, Y., Wang, E., & Shain, C. (2025, March 27–29). *Distance to plausible alternatives predicts acceptability ratings in comparative illusion* [Poster presentation]. The 38th Annual Conference on Human Sentence Processing, College Park, MD, United States. https://doi.org/10.17605/OSF.IO/EXN5K

Received May 1, 2024
Revision received April 30, 2025
Accepted May 14, 2025 ■