THE ANATOMY OF A COMPARATIVE ILLUSION

ALEXIS WELLWOOD, ROUMYANA PANCHEVA, VALENTINE HACQUARD, COLIN PHILLIPS

Abstract. Comparative constructions like *More people have been to Russia than I have* are generally perceived as acceptable and meaningful by native speakers of English; yet, upon closer reflection, they are judged to be incoherent. This mismatch between initial perception and more considered judgment challenges the idea that we perceive sentences veridically, and interpret them fully; it is thus potentially revealing about the relationship between grammar and language processing. This paper presents the first detailed investigation of these so-called 'comparative illusions'. We test four hypotheses about their source: a shallow syntactic parser, some type of repair by ellipsis, an incorrectly-resolved lexical ambiguity, or a persistent event comparison interpretation. A series of four formal acceptability studies show that speakers are most prone to the illusion when the comparative supports an event comparison reading. A verbatim recall task tests and finds evidence for such construals in speakers' recollections of the sentences. Our results suggest that speakers entertain an interpretation that is initially consistent with the sentence, but fail to notice when this interpretation becomes unavailable at the *than*-clause. We conclude that, rather than illustrating processing in the absence of grammatical analysis, comparative illusions underscore the importance of syntactic and semantic rules in sentence processing.

Draft of manuscript submitted 09/06/2015
Word count (including abstract): 16,795

1. Comparative illusions

Presented with the sentence in (1), native English speakers typically report that it is a perfectly acceptable sentence of their language. Yet, upon closer reflection, these same speakers judge that it has no stable, meaningful interpretation. Sentences of this form have come to be called 'comparative illusions' (CIs) or 'Escher sentences': they have only the appearance of syntactic and semantic well-formedness. CIs are interesting in that they seem to challenge some of our most basic assumptions about language architecture: that we perceive sentences veridically, that we interpret them fully, and that sentence form and meaning are tightly coupled.

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

The phenomenon has been known for some time, but the mismatch between the perception of grammaticality/meaningfulness that characterizes CIs has yet to be systematically investigated. The sentence in (1) was first reported by Montalbetti (1984) as 'the most amazing */? sentence I've ever heard', attributing it to Hermann Schultze. Importantly, linguists and non-linguists alike experience the phenomenon, but, despite much informal discussion in the linguistics community, formal investigation has so far been limited to preliminary results (Fults & Phillips 2004, Wellwood et al. 2009, and O'Connor et al. 2012).

In this paper, we investigate which properties of sentences like (1) are essential for the initial perception of meaningfulness. Grammatically, the problem with CI-type sentences appears to be in the choice of subject in the *than*-clause, since superficially similar comparatives succeed in being both uncontroversially acceptable and meaningful. The meaning of a sentence like (2) just is 'the number of people that have been to Russia exceeds the number of elephants that have'. Yet there is no interpretation of (1) suggested by a similar paraphrase, 'the number of people that have been to Russia exceeds the number of me'.

(2) More people have been to Russia than elephants have.

Given this difference, it is striking that CIs seem to be as acceptable as is reported. And how does the apparent meaningfulness arise? Acceptability alone does not, of course, license the inference that CIs are grammatical: grammaticality and judgments of acceptability often diverge (cf. Lewis & Phillips 2015 for relevant discussion). Garden path sentences (3a) and sentences with multiple center-embedding (3b) are unacceptable, yet nonetheless grammatical (see Bever 1970 and Lewis 1996, respectively). Conversely, in some cases ungrammatical sentences are judged to be acceptable, as in cases of plural attraction (3c) and NPI illusions (3d) (see Bock & Miller 1991; Clifton et al. 1999; Vasishth et al. 2008, Xiang et al. 2009, Parker & Phillips, submitted).

(3) a. The horse raced past the barn fell.
b. The man the woman the child kissed knows jumped.
c. * The key to the cabinets are on the table.
d. * The bills that no senator voted for will ever become law.

Other well-known examples of divergence include grammatical sentences that are perceived to have meanings starkly different from their literal meanings. If a man has a widow, then that man is dead, and no dead man can marry; yet, 30% of respondents answer 'yes' when presented with (4a) (Sanford & Sturt 2002). Similarly, (4b) is literally equivalent to 'All head injuries are trivial enough to ignore', a suggestion to let all such injuries go unattended; nevertheless, speakers routinely understand (4b) as equivalent to 'All head injuries are too important to ignore' (Wason & Reich 1979). In these cases, comprehenders construct and linger on a certain misinterpretation that prevents them from recognizing the error.

(4) a. Can a man marry his widow's sister?
b. No head injury is too trivial to ignore.

CIs potentially present a different sort of case from all of the above examples. Sentences like (1) strike speakers as well-formed, unlike (3a) and (3b). Even extended consideration can leave them seeming well-formed, unlike (3c) and (3d). Despite this, one never arrives at a specific, grammatically-licensed interpretation; there doesn't seem to be a single kind of misinterpretation that speakers can eventually converge on, as there is with (4a) and (4b). Rather, informal reports suggest that speakers tend to believe sentences like (1) are acceptable and have a coherent interpretation, even while they struggle to articulate that interpretation.

CIs thus present a disconnect between apparent well-formedness and shifting-sands interpretation. To understand the phenomenon better, we first need to understand under what conditions it arises, and how far it generalizes. We consider four plausible sources of the CI effect, each differing in which components of language processing or grammar it implicates.

Perhaps the illusion arises due to a shallow syntactic parser, §2.1. Or perhaps ellipsis repairs what would otherwise be a noticeable grammatical defect, §2.2. On the other hand, CIs might arise because of a lexical ambiguity between comparative and additive *more*, §2.3. Finally, the effect could be due to a lingering 'event comparison' interpretation that, while consistent with the syntax of comparatives like (2), is nonetheless ungrammatical for (1), §2.4.

Each account that we consider makes different predictions about which properties of CI-type sentences should make speakers more or less susceptible to the illusion. In §3, we put the predictions of these accounts to the test in four formal acceptability studies, and find evidence only for the event comparison hypothesis. A sentence recall experiment in §4 probes how speakers recast the sentences in production, and finds more direct evidence for this interpretation in how speakers comprehend CIs. The results of these studies are only consistent with the hypothesis that the logical semantics of comparison supports the illusion: speakers initially adopt an event-counting reading that is licensed by the normal syntactic rules in fully grammatical comparatives, but they fail to notice when the interpretation is no longer available.

This study thus informs broad questions about the architecture of sentence processing, and the degree to which grammatical theory informs sentence processing. CIs are not representative of very general, rough-and-ready mechanisms, with little role for traditional syntactic or semantic analysis (e.g. Townsend & Bever 2001, Christianson et al. 2001, Ferreira et al. 2002, Sanford & Sturt 2002). If they were, CI effects should generalize very broadly—not only for comparatives, but far beyond them. In other words, we should see similar illusions frequently. Alternatively, if CIs are tightly linked to well-motivated grammatical mechanisms, then they should be exotic, not generalizing particularly far. This is what we observe: the effect only exists as long as it is possible to construct a plausible event-counting interpretation.

Overall, we find that the parser tracks sentential grammar very closely. It does not involve creating representations that are only partially reflective of syntactic analysis, divorced from semantics (e.g. 'first pass' parsing); nor does it involve abandoning grammatical analysis, and recourse to world-knowledge in the face of significant difficulties (e.g. 'shallow' parsing). Rather, the parser can be 'fooled' by attractive semantic analyses that differ minimally from the problematic syntactic representations it is asked to build (i.e. an analysis that meets the semantic requirements of *more*). Rather than illustrating that processing is possible in the absence of grammatical analysis, comparative illusions underscore the importance of syntactic and semantic rules in sentence processing.

Before turning to our investigation, we briefly address the question of whether examples like (1) are grammatical.

1.1. Are CIs grammatical?

A potential objection to this work is that it is not necessary. CIs have the same sequence of grammatical categories as comparative sentences that are otherwise fine. That is, at a surface level of analysis, the only apparent difference between (1) and (2) is whether a pronoun or a bare plural sits in subject position of the *than*-clause. However, these expressions are both NPs; and if all that matters to grammaticality is the sequence of such categories, then why the fuss about CIs?

Yet, there are reasons to think that these differences matter, and that a surface view of what makes a sentence grammatical should be rejected.

First, there is a syntactic problem with (1). In the syntax tradition going back to at least Bresnan (1973; see also Chomsky 1977), part of the representation of a nominal comparative is a correlate of the matrix clause *more* in the *than*-clause. This correlate is structurally akin to the expression *how many* that appears overtly in degree questions, where nothing but a bare plural is acceptable, contrast (5a) with (5b)-(5d). Indeed, Bulgarian comparatives have an expression *kolkoto* ('how many') that appears overtly in the *ot*-clause, and here no CI-effect is observed, (6a); without a bare plural, the result is categorically unacceptable, (6b).

(5) a. How many elephants have been to Russia?

b. * How many I have been to Russia?

c. * How many the elephant has been to Russia?

d. * How many the elephants have been to Russia?

(6) Poveče amerikanci sa bili v Rusija ...
more americans are been in Russia ...
'More Americans have been to Russia...'

a. ... ot-**kolkoto** slonove sa bili v Rusija
... from-how.many elephant.PL are been in Russia
'than elephants have been to Russia.'

b. * ... ot-**kolkoto** az / slon-ăt / slonove-te
... from-how.many I / elephant-the / elephant.PL-the
'than I / the elephant / the elephants.'

Second, there is a logical semantic problem with (1). Since at least Heim (1985), it has become standard in the degree-theoretic tradition to assume that the semantics of *many* (as part of *more* in the matrix clause, and covertly in the *than*-clause) maps pluralities to their cardinalities. More recently, it has been rendered formally explicit that this is only possible if the expression that *many* combines with is semantically nonsingular (Hackl 2001, Nakanishi 2007, Wellwood et al. 2012). Singular pronouns, singular definite descriptions, and names of individuals all fail to meet this requirement.

A hypothetical contrasting position, in which CIs are deemed to be grammatical, would hold that grammatical generalizations are stated at the level of syntactic categories, and finer-grained distinctions like those just discussed are not important. If the structure underlying (2), a comparison stated between two NP subjects, is legitimate, then one with the same sequence of grammatical categories, (1), should also be legitimate.

Our position is that the classic illusion violates the syntactic and semantic requirements of the comparative construction, and as such should be considered ungrammatical. However, this position will not really affect the discussion that follows. What is crucial is the reported mismatch between speakers' judgments of the acceptability of CI-type sentences, and the fact that they seem to have no coherent sense. This is what we investigate for the remainder of the paper.

2. Plausible sources for comparative illusions

2.1. Syntactic template matching

One attractive way of thinking about why CIs are acceptable exploits a model of sentence processing that implements a template matching procedure. On this view, articulated explicitly by Townsend & Bever (2001), acceptability judgments reflect a two-stage process: a sentence is initially subjected to a relatively superficial matching process that compares it to frequent clause templates, and then it is subjected to more detailed grammatical analysis after a delay, if at all.[1] More generally, the observation is that each of the two clauses alone is perfectly acceptable in some contexts, which is enough to license the inference that CIs are themselves acceptable.

To see how such an account would work, consider the sentences in (7). (7a) expresses a comparison between the number of individuals that have been to Russia and the number that the speaker would have thought have been to Russia. (7b) expresses a comparison between the number of events of people going to Russia and the number of events in which the speaker has been to Russia. From sentences like these, matrix and *than*-clause templates may be extracted; and parsing (1) should just involve matching its matrix and *than*-clauses to such templates.

(7) a. **More people have been to Russia** than I would have thought.
    b. People have been to Russia more **than I have**.

Sentences like (1) should thus satisfy the parser's initial analyses, and so could support judgments of acceptability before any more elaborate analyses are conducted. The implication is that, while CIs may fail at a deeper level of analysis, their success at shallower levels accounts for their apparent acceptability. We state this hypothesis as in (8).

(8) **Syntactic template matching hypothesis**
    CIs reflect the successful matching of a comparative sentence to one or more syntactic templates.

Such an account is quite general; however, based on how it is presented in print (and in personal communication with one of the authors), we can interpret it in two ways. Either (i) a CI is acceptable because each of its clauses is well-formed on its own (the less constrained theory), or (ii) a CI is acceptable just in case there is lexical overlap between the two clause templates against which it is compared (the more constrained theory).

On the less constrained theory, all that should matter for acceptability is that each clause is independently plausible. More generally, it should be possible to arbitrarily combine different clauses without penalty in a wide variety of examples. This does not seem to be the case: the blends in (9)-(10), for example, seem immediately unacceptable, while their constituent clauses are fine in other grammatical contexts. Put differently, if the account is as free as this theory suggests, we would expect to see CI-type effects all the time, while they appear to be relatively rare.

[1] Townsend and Bever have directly addressed comparative illusions and make specific predictions in this direction. One may be able to construct a similar account in the style of 'good enough' processing (e.g. Christianson et al. 2001, Ferreira et al. 2002; see especially Ferreira & Patson 2007 for an overview) or shallow parsing (Sanford & Sturt 2002). The challenge of these accounts is that they don't make specific predictions, so it is hard to test them.

(9) * Mary is too tall as Bill has.
 a. **Mary is too tall** to get on this ride.
 b. Mary has ridden some ride as many times **as Bill has**.

(10) * As many girls have been to Russia than I do.
 a. **As many girls have been to Russia** as boys have.
 b. People go to Russia more **than I do**.

It is hard to tease out the predictions of a theory like (i). Yet, if the account is more constrained as in (ii), then it makes a clear prediction: (11) should not be judged acceptable, since a matrix clause template like (11a) will be available for comparison, but a *than*-clause template like (11b) will not be. Thus, the more constrained theory predicts that speakers should find CI-type sentences with *fewer* to be significantly less acceptable than those with *more*.

(11) Fewer people have been to Russia than I have.
 a. **Fewer people have been to Russia** than I would have thought.
 b. * People have been to Russia fewer **than I have**.

We proceed assuming the more constrained version of the template-matching hypothesis, which predicts that participants will not judge a sentence like (11) to be as highly acceptable as its counterpart with *more*. If participants nonetheless accept CI-type sentences like these at the same rate, then the syntactic template-matching account would have to be made considerably more abstract in order to explain the CI effect.

2.2. Ellipsis repair

A different account links the acceptability of CIs with a process of repair by ellipsis. This proposal differs from the syntactic template-matching approach in that it posits a significant role for grammar in facilitating the illusion, linking the phenomenon with other cases in which successful applications of grammatical processes ameliorate problems elsewhere.

Investigating the possibility that an ellipsis operation facilitates the CI effect, Fults and Phillips (2004) tested CI-type sentences, and found significant degradation in acceptability for those without ellipsis. (Numbers indicate mean ratings on a 1-5 scale.)

(12) a. More people have been to Russia than I have. *3.58*
 b. More people have been to Russia than I have **been to Russia**. *2.90*

Such an account finds plausibility in the many reported cases in which the application of a grammatical rule ‘blinds’ comprehenders to other illicit rule applications. With respect to ellipsis, both the formal syntax (Ross 1969, Lasnik 2001, Merchant 2001, Kennedy 2003) and experimental literature (Frazier & Clifton, Jr. 2011) confirm that sluicing can rescue sentences which would otherwise present robust island violations (e.g. *Mary wants to hire someone who speaks a Balkan language, but I don’t remember which*), and Richards (1997) discusses similar effects with multiple applications of *wh*-movement.

Thus, it may be that some aspect of successfully resolving ellipsis in the *than*-clause of a CI plays a role in its acceptability. We call this the repair-by-ellipsis hypothesis, (13).

(13) **Repair-by-ellipsis hypothesis**
 CIs reflect successful resolution of ellipsis in the *than*-clause.

This hypothesis predicts that CI-type sentences with ellipsis should be judged more acceptable than corresponding sentences with no ellipsis. Yet, there is a potential confound in Fults & Phillips' result: identical material is preferentially deleted in the *than*-clause of a comparative in English (Bresnan 1973). Thus, it could be that simple repetition of the matrix and *than*-clause predicates in (12b) independently reduced their participants' ratings. Nonetheless, the repair-by-ellipsis hypothesis predicts that sentences like (14) with a superficially different VP between the matrix and *than*-clauses should be judged less acceptable than (1).

(14) More people have been to Russia than I have **been to Canada**.

If participants judge sentences like (14) to be as highly acceptable as (1), then an explanation for the CI effect in terms of repair-by-ellipsis is less plausible.

2.3. <u>more</u> ambiguity

Applying the normal interpretive rules to CIs ultimately yields uninterpretability. Yet, the effect could be due to speakers temporarily constructing an alternative interpretation that is coherent, and this accounts for a heightened perception of acceptability. Such an explanation departs from the previous two accounts in implicating semantic processing in the CI effect.

Upon recognizing the incoherence of sentences like (1), speakers often suggest that there is in fact a fully coherent interpretation that could be paraphrased as either of the sentences in (15).

(15) a. **I'm not the only person** that has been to Russia.
b. More people have been to Russia than **just me**.

Such intuitions could suggest that the CI effect arises due to a lexical ambiguity between comparative and 'additive' *more*. To see the difference, consider (16). The additive interpretation in (16a) indicates a quantity in addition to (but not necessarily greater than) a previously-mentioned quantity. The comparative interpretation in (16b) indicates a quantity that is strictly greater than that previously mentioned. Interpreted additively, (1) would be true in any circumstance where there is some number of people who have been to Russia in addition to the speaker. (For detailed discussion of such ambiguities, see Greenberg 2010 and Thomas 2010; cf. Grant's 2013 investigation of similar constructions.)

(16) Mary has worked 10 hours so far on the project. Now she has to work on it **more**.
a. **Additive**: ...some quantity in addition, possibly less than 10 hours.
b. **Comparative**: ...more than 10 hours.

The additive interpretation of *more* is not grammatical when there is an overt *than*-clause in the sentence, however. Thus, the CI effect could reflect parsing the sentence with an additive interpretation via *more people*, and failing to notice when that reading is no longer available, at *than*. This hypothesis is stated as in (17).

(17) **Additive *more* hypothesis**
CIs reflect misinterpretation of comparative *more* as additive *more*.

Such an account predicts that the CI effect should be facilitated just when an additive semantics for *more* is supported. This makes two predictions. First, it must be possible to interpret the subject of the *than*-clause as a member of the set denoted by the matrix subject. On the assumption that no boy belongs to the set of girls, the sentence in (18a) could not mean 'More girls have been to Russia than just that boy.' Second, the comparative quantifier must be *more*: a *just me*-type sentence with *fewer* is uninterpretable, (18b).

(18) a. More **girls** have been to Russia than **that boy** has.
b. * **Fewer** people have been to Russia than **just me**.

The classic illusion in (1) contains a first person subject of the *than*-clause, which could be interpreted as indexing an entity among the set denoted by the matrix subject NP. The additive *more* hypothesis thus predicts that participants should judge sentences like (18a) to be less acceptable than sentences like (1). Additionally, since *fewer* fails to support the additive interpretation, sentences like (18b) should also be less acceptable than sentences like (1).[2]

2.4. Event comparison

The fourth and final account that we consider links the CI effect to the semantics of comparative constructions more generally, rather than to a specific lexical ambiguity. In particular, it ties the effect to a regular process by which a subject nominal comparative can be interpreted in terms of a comparison of events.

This proposal relates to a different suggestion that speakers often make when they encounter CIs: that they express a comparison of numbers of events, just like that of a verbal comparative like (19).

(19) **People** have been to Russia **more than I have**.

A straightforward implementation of this suggestion would be to posit that the CI effect is due to speakers' reanalyzing (1) as in (19). Yet, it is often possible for numerically-quantified noun phrases to be interpreted directly as expressing counts of individual participations in events (Krifka 1990, Barker 1999, Schein forthcoming): for example, (20a) can be true even if the total number of unique individuals is much less than 5 million, so long as there are at least 5 million ridings per week. The only constraint on the availability of this reading is the meaning of the predicate: (20b), unlike (20a), can only be true if the number of unique individuals is 5 million, since it is only possible for a given person to *be on the metro right now* exactly once.

(20) a. 5 million people **ride the metro** each week.
b. 5 million people **are on the metro** right now.

Krifka (1990) locates the event-counting reading in a null determiner ambiguity, whereas Barker (1999) ties it to how the identity conditions on entities are determined for the purposes of counting. For details on these proposals, we refer the interested authors to those works. For our purposes, what is important is that the event comparison reading is available to fully grammatical nominal comparatives. To see this, consider a context like that in (21). Here, the sentence in (22) can be judged true if individuals are counted, (22a), while it can be judged false if events are counted, (22b).

[2] This account makes a further prediction: comparative illusions should only be possible in languages where the comparative and additive morpheme are identical morphophonologically. Greenberg 2010 suggests that not all languages are like English in this respect.

(21) 10 sailboats passed through the lock 10 times each (100 passings), and 5 barges passed through the lock 50 times each (250 passings).

(22) More sailboats passed through the lock than barges did.
 a. **Individual counting**: 10 is greater than 5 ⇒ TRUE
 b. **Event counting**: 100 is not greater than 250 ⇒ FALSE

Thus, the CI effect could arise from speakers analyzing the sentence as a comparison of numbers of events. The event comparison reading is entertained because it is grammatically licensed by the matrix clause, and it persists despite being syntactically unsupported by the *than*-clause (see §1.1). This hypothesis is stated in (23).

(23) **Event comparison hypothesis**
CIs reflect speakers' attempts to compare counts of events.

The event comparison hypothesis predicts that the CI effect should be facilitated just when the semantic properties of the VP support an event-counting interpretation. That is, the predicate must be 'repeatable', as opposed to 'once-only' or 'nonrepeatable' (cf. Nakanishi 2007, Wellwood et al. 2012, Wellwood 2015). (24a) is perfectly interpretable, as Mary may be involved in however many events of running a marathon as she likes, but (24b) is odd, as it suggests that Mary graduated high school multiple times.

(24) a. Mary **ran a marathon** more than John did.
 b. ? Mary **graduated high school** more than John did.

(1) contains a repeatable predicate (*go to Russia*), unlike the CI-type sentence in (25). The event comparison hypothesis thus predicts that a CI-type sentence with a nonrepeatable predicate like that in (25) should be judged as less acceptable than a sentence like (1).

(25) More people **graduated high school** than I did.

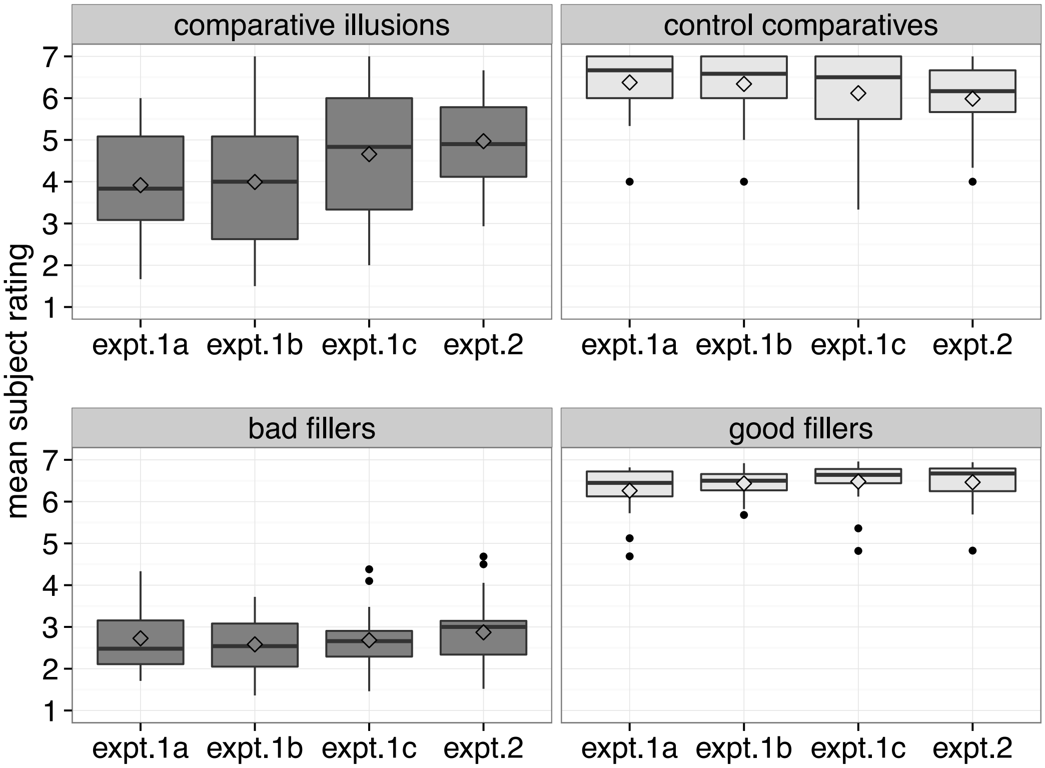
An alternative, attractive version of this hypothesis is that event comparison is the result of syntactic reanalysis, in which *more* is displaced to an adverbial position like in (19). Such a hypothesis could additionally predict a contrast between *more* and *fewer*, since the form of the adverbial quantifier is *less* (e.g. **More people have been to Russia fewer than I have*). Yet, such a prediction depends in part on the background syntax-semantics theory: *less* and *fewer* could be analyzed as morphophonological variants. In light of this possibility, we focus on the semantic version of the event comparison hypothesis until Experiment 3 in §4, when we pit the semantic and syntactic versions against each other.

3. Acceptability judgment studies

In a series of four acceptability judgment studies with 64 unique participants, we investigated the robustness of the CI effect, and which properties are essential to it.

First, how acceptable are CIs once they are put in a formal experimental setting? As we will see directly, the patterns of acceptability responses for sentences like (1) and (2) differed substantially: the mean acceptability ratings for CIs were much lower than were those for control sentences, and the distributions of ratings were much more variable. Importantly, however, the pattern we observed in the ratings for CIs also differed substantially from those

FIGURE 1. Boxplots of mean subject ratings for classic CI-type sentences and controls (top row), and bad and good fillers (bottom row), across acceptability experiments. For each column: diamonds represent the overall mean; heavy lines indicate the median; the upper and lower 'hinges' of the box represent the first quartile (25th percentile) and third quartile (75th percentile); the upper whiskers extend to the highest value within 1.5 times the inter-quartile range of the upper hinges (IQR; the distance between first and third quartiles); the lower whiskers extend to the lowest data point within 1.5 times IQR of the lower hinges; black circles indicate values outside of these ranges (i.e. outliers).



of 'bad' filler sentences, while that for control comparatives did not differ substantially from those of 'good' filler sentences.

For now, we limit our attention to 'comparative illusions' and 'control comparatives' that differ minimally from (1) and (2): subject nominal comparatives with repeatable predicates and *than*-clause ellipsis. CIs, here and below, differ from controls only in the type of *than*-clause subject: CIs have a non-bare plural subject, while controls have a bare plural subject. Our filler sentences were designed to elicit either a low or high rating while having a similar length, degree of syntactic complexity, and, in around one-third of the cases, a similar degree of semantic complexity (i.e. comparative-type meanings) to target items, (26)-(27).

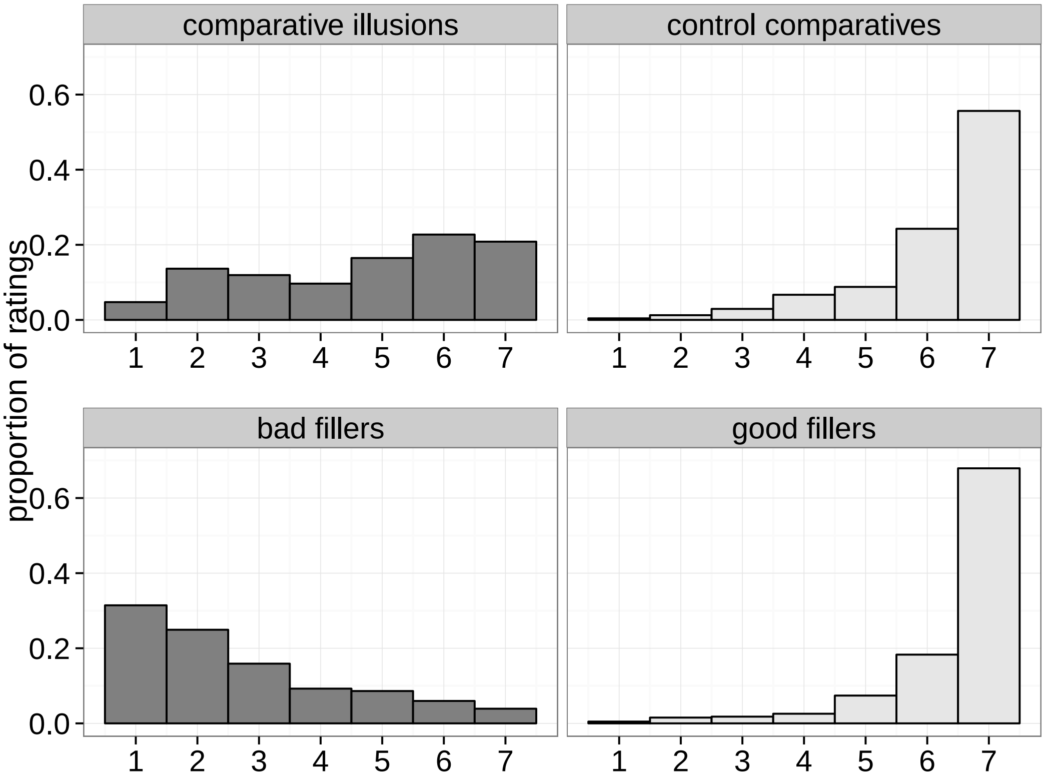
(26) **Examples of 'bad' fillers**

a. A computer program that can be downloaded as many times than you did.

b. Australians will have been to Europe this season to visit the mountains that Uganda.

(27) **Examples of 'good' fillers**

Figure 2. Density plots of rating distributions for classic CI-type sentences and controls (top row), and bad and good fillers (bottom row), across acceptability experiments. Each plot represents the proportion of total responses observed for each scalar value.



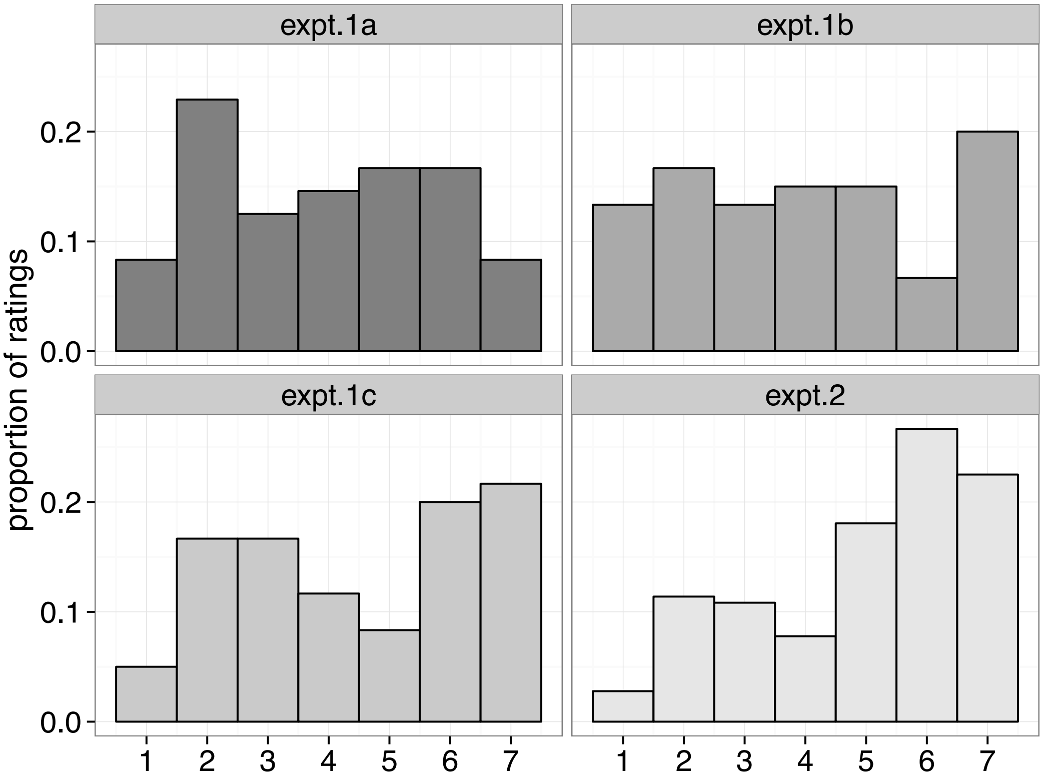
a. Less than 30 percent of the students in the class gave a high rating to the professor.

b. A bartender who works at Sam's favorite bar is known for pouring the best draft beer.

The differences in response to these categories of sentence can be clearly seen by inspecting the plots in Figures 1 and 2. In Figure 1, we see that the mean ratings by subject were an average of 2-3 points lower for CIs than for controls, but 2-3 points higher for CIs than for 'bad' fillers. In addition, the range of averaged responses typically spanned 4-5 points for CIs, while it was generally 2-3 points for control sentences and 'bad' fillers. Except with CIs, participants were fairly consistent in their ratings. The same conclusion is supported by consideration of the density of rating scores in Figure 2. The ratings for CIs were fairly evenly spread along the scale, with a slight bias towards the upper end, while the other categories of sentence clearly tended either towards high acceptability (controls, and 'good' fillers) or low acceptability ('bad' fillers).

Interestingly, however, the fairly flat distribution summarized across experiments in Figure 2 was not observed within each experiment. Figure 3 plots the distributions of raw ratings for CIs across experiments. The distributions were fairly flat in Experiments 1a and 1b, but begin to more closely approximate that of control sentences in Experiments 1c and 2. This variability in ratings was not due to specific items, individuals, or (in general) to order of presentation effects. All of these experiments featured the same basic sets of items; the major difference is that, in Experiments 1c and 2, some of the *than*-clause subjects of CIs

Figure 3. Density plots of rating distributions for classic CI-type sentences by experiment. Each plot represents the proportion of total responses logged for each scalar value.



were plural (i.e. a pronoun or definite description), while all of those in Experiments 1a and 1b were singular.

We discuss the theoretical import of plural subjects later. However, relevant to this overview is that Experiments 1c and 2 also featured the greatest variety of *than*-clause subject types for CIs—including singular and plural proper names, pronouns, and definite descriptions. It is possible that this variety made CI-type sentences less salient, and thus participants were less likely to notice the anomaly. This possibility is supported by the fact that the only one of our acceptability studies that showed order of presentation effects—Experiment 1b—was also the one with the least variety of *than*-clause subject types for CIs, featuring only first person pronouns as in the classic illusion in (1).

In sum, CIs showed much greater variability in acceptability in an experimental setting than did sentences of unambiguous grammatical status. Sometimes they were rated as highly acceptable, other times middling, and other times hardly acceptable.

We turn now to our second question: what factors make participants more likely to assign a high rating? As we will see, only one factor consistently had such an effect: CI-type sentences with repeatable predicates were consistently rated higher than were those with nonrepeatable predicates, supporting the event comparison hypothesis.

3.1. Experiment 1

Experiment 1 had three substudies that were identical in their basic design, execution, and the nature of their participant population. Each was an offline acceptability judgment task with responses recorded on a 7 point scale, where 1 was 'unacceptable' and 7 was 'acceptable'. Participants were University of Maryland undergraduates, all native speakers of American English as determined by a pre-test questionnaire, who received either course credit or $10 for 1 hour of participation. The studies took no more than 30 minutes to complete, and the remaining 30 minutes of participant time were used for unrelated experiments.

Each hypothesis outlined in §2 predicts that specific factors should make speakers more susceptible to CIs. For each of these hypotheses, it is necessary to identify manipulations that selectively impact CIs to the exclusion of fully grammatical controls. Therefore, in all experiments in this series, our primary manipulation was a comparison of illusion and control conditions (the ILLUSION factor), where the illusion conditions were defined by having non-bare plural subjects in their *than*-clauses, and the control conditions were defined by having bare plural subjects in their *than*-clauses. In each study, this factor was crossed with a subset of further factors that specific hypotheses predict should impact acceptability.

The factor QUANTIFIER manipulated the comparative quantifier in the main clause subject position (*more* vs. *fewer*; (28)). This manipulation was used to test two of the four hypotheses about the source of the CI effect. The template matching hypothesis (more constrained version) relies on the ambiguity of *more* as both a determiner and an adverbial. The additive *more* hypothesis relies on the ambiguity of *more* as a lexical item with either a comparative or an additive semantics. Since *fewer* cannot function as an adverbial, and it lacks an additive semantics, both hypotheses predict that CI-type sentences with *fewer* should fail to elicit the CI-effect. (The examples illustrating the factors are simplified for presentation purposes; these are not actual experimental items.)

(28) QUANTIFIER
More/fewer girls ate pizza than the boy did.

The factor ELLIPSIS manipulated whether VP ellipsis had applied in the *than*-clause (ellipsis vs. no ellipsis; (29)). This manipulation was used to test the repair-by-ellipsis hypothesis. That account holds that the acceptability of CIs depends on VP ellipsis, and thus predicts that CI-type sentences with an unelided VP should not elicit the CI effect. However, previous research supporting this prediction (Fults and Phillips 2004) failed to take into account the grammatical preference for deletion in comparatives (Bresnan 1973). In our design, the VP in the 'no ellipsis' conditions differed just enough from the matrix clause VP to circumvent this preference.

(29) ELLIPSIS
More girls ate pizza than the boy {**did**}/{**ate yogurt**}.

The factor PREDICATE TYPE manipulated whether the VP in the comparative was repeatable for a given agent (repeatable vs. nonrepeatable; (30)). This manipulation was used to test the event comparison hypothesis, which holds that the CI effect is due to a persistent event-comparison reading; this interpretation is only grammatically licensed in the matrix clause if it contains a repeatable predicate like *eat pizza*. This account thus predicts that CI-type sentences with nonrepeatable predicates like *graduate high school* should fail to elicit the CI effect.

Figure 4. Schemata for repeatable and nonrepeatable items in Experiments 1a-c, representing 16 unique conditions. Factors represented are predicate type (between items—repeatable, nonrepeatable), quantifier (*more*, *fewer*), illusions (*the boy*) versus controls (*boys*), ellipsis (ellipsis, no ellipsis). subject inclusion was manipulated only within the illusion conditions; those with *girls... the boy* are 'inclusion not possible' trials. *H.S.* abbreviates *high school*, and is used for graphical conciseness; none of our experimental items involved abbreviations.

Sample repeatable item

$$\left\{\begin{array}{c}\text{More}\\ \text{Fewer}\end{array}\right\} \text{ girls ate pizza than } \left\{\begin{array}{c}\text{the boy}\\ \text{boys}\end{array}\right\}\left\{\begin{array}{c}\text{did.}\\ \text{ate yogurt.}\end{array}\right\}$$

Sample nonrepeatable item

$$\left\{\begin{array}{c}\text{More}\\ \text{Fewer}\end{array}\right\} \text{ girls graduated H.S. than } \left\{\begin{array}{c}\text{the boy}\\ \text{boys}\end{array}\right\}\left\{\begin{array}{c}\text{did.}\\ \text{failed out.}\end{array}\right\}$$

(30) predicate type
More girls **ate pizza/graduated high school** than the boy did.

The factor subject inclusion manipulated whether the denotation of the *than*-clause subject could be included in the denotation of the subject NP of the matrix clause ('inclusion possible' vs. 'inclusion not possible'; (31)). This manipulation was used to test the additive *more* hypothesis, which requires the possibility of an inclusion relation in order to license a 'not just me' interpretation of the comparative. In our design, 'inclusion possible' trials usually involved gender-matching the NPs the matrix and *than*-clause subjects, and 'inclusion not possible' involved mismatching these NPs. The additive *more* hypothesis predicts that 'inclusion not possible' trials should fail to elicit the CI effect.

(31) subject inclusion
More **boys** called to complain than **he/she** did.

Two sample items of the type used in Experiments 1a-c are given in Figure 4. The top diagram represents an item with a repeatable VP, and the bottom an item with a nonrepeatable VP. Any path through the figure from left to right corresponds to one condition; the 8 possible paths through each diagram together correspond to 16 experimental conditions. subject inclusion was manipulated only within the illusion conditions; those represented in Figure 4 are all 'inclusion possible' trials.

Table 1 summarizes the predictions of the four accounts relative to these factors, with '>' representing the prediction that the factor heading that column should yield higher acceptability for CI-type sentences on the left-hand level as opposed to the right-hand level. Apart from subject inclusion, the effects of these manipulations are predicted to be interactions, affecting the illusion conditions but not the control conditions. As noted above, subject inclusion was tested only within the illusion conditions.

Sets of items were distributed across 8 lists in a Latin Square design, and were then combined with filler sentences for roughly a 1/3 ratio of experimental to filler sentences,

Table 1. Predicted interactions by hypothesis and factor. Each hypothesis (apart from the additive *more* hypothesis) predicts an interaction between the factor illusion and the factor listed at the top of each column. '>' indicates the direction of the interaction, such that the illusion conditions should be more acceptable on the left-hand factor loading than on the right-hand loading, as compared to the control conditions. The additive *more* hypothesis predicts that the illusion conditions in which subject inclusion is possible should be more acceptable than the illusion conditions where it is not possible. A '-' indicates that the hypothesis makes no predictions for that factor.

	QUANTIFIER	ELLIPSIS	SUBJECT INCLUSION	PREDICATE TYPE
Hypothesis	*more*—*fewer*	ellipsis — no ellip.	possible—not poss.	repeat—nonrep.
Template matching	>	-	-	-
Repair by ellipsis	-	>	-	-
Additive *more*	>	-	>	-
Event comparison	-	-	-	>

creating 8 questionnaires. The number of items differed across experiments in the series; specific details are given below. Fillers were designed to approximate the complexity of the experimental items, and were evenly split between those that should elicit lower and higher ratings (see discussion of (26) and (27) above). Approximately one-third of the total number of fillers had comparative forms (i.e. non-subject nominal and verbal comparatives, equatives, or superlatives; see (26a) and (27a) for examples), which were included to help mask the experimental items.

3.1.1. Experiment 1a

In Experiment 1a we tested the predictions of the four hypotheses about the source of the CI effect, as summarized in Table 1.

We manipulated the factors illusion, quantifier, and ellipsis within items, and the factor predicate type between items. The factor subject inclusion was manipulated within the illusion conditions. In this experiment, CI-type sentences featured only singular proper names, 3rd person pronouns, and definite descriptions as *than*-clause subjects. We made this choice because 3rd person expressions provided a minimal contrast with the matrix bare plural NP, and because it would not be possible to use first person pronouns and still manipulate subject inclusion within the illusion conditions. First person pronouns could in principle always pick out a member of the group denoted by the matrix clause subject.

Each questionnaire consisted of 48 experimental trials and 144 filler trials, and there were 16 participants.

Results Unless otherwise stated, all statistics reported below are the result of linear mixed effects regressions with maximal random effects terms (i.e. including random intercepts and slopes by subject and item; Barr et al. 2013). All analyses were conducted using R's lmer4 package (Bates et al. 2014).

We found a reliable main effect of illusion, with the control conditions rated more highly than the illusion conditions (means: control 5.88, illusion 3.11; $\beta = 2.77, \mathrm{SE} = .28, \chi^2(1) = 31.95, p < .001$). Recalling our discussion in the introduction to §3, these means for the illusion conditions should be interpreted with caution; they reflect a mixture of high and low scores, as well as scores for many items that our experimental manipulations predicted should lead to lower ratings.

There was a reliable main effect of QUANTIFIER, with the *more* conditions rated more highly than the *fewer* conditions (*more* 4.61, *fewer* 4.38; $\beta = .24, \mathrm{SE} = .1, \chi^2(1) = 5.77, p < .02$). Importantly, there was no interaction of ILLUSION and QUANTIFIER ($\beta < .01, \mathrm{SE} = .22, \chi^2(1) < .01, p = .98$). Replacing *more* with *fewer* was associated with a small decline in ratings in the illusion conditions (*more* 3.22, *fewer* 2.99) and in the control conditions (*more* 5.99, *fewer* 5.76). These results are not consistent with the syntactic template matching hypothesis or the additive *more* hypothesis, both of which predicted a substantial difference between the comparative quantifiers specifically within the illusion conditions.

There was no main effect of ELLIPSIS (ellipsis 4.55, no ellipsis 4.43; $\beta = .11, \mathrm{SE} = .13, \chi^2(1) = .71, p = .4$), and no interaction of ILLUSION and ELLIPSIS ($\beta = -.06, \mathrm{SE} = .23, \chi^2(1) = .06, p = .8$). Within the illusion conditions, ratings were similar for the 'ellipsis' and 'no ellipsis' conditions (ellipsis 3.18, no ellipsis 3.04) as well as within the control conditions (ellipsis 5.92, no ellipsis 5.83). These results fail to support the repair-by-ellipsis hypothesis, which predicts that CI-type sentences, but not controls, should receive higher ratings with *than*-clause ellipsis.

There was a reliable main effect of PREDICATE TYPE, in which the repeatable conditions were rated more highly than the nonrepeatable conditions overall (repeatable 4.72, nonrepeatable 4.27; $\beta = .45, \mathrm{SE} = .11, \chi^2(1) = 12.7, p < .001$). We also found an interaction of ILLUSION and PREDICATE TYPE ($\beta = -.57, \mathrm{SE} = .23, \chi^2(1) = 5.6, p < .02$). The nonrepeatable illusion conditions were rated lower than the repeatable illusion conditions (nonrepeatable 2.74, repeatable 3.47) whereas controls did not show this effect (nonrepeatable 5.79, repeatable 5.96). This interaction supports the event comparison hypothesis.

Turning to SUBJECT INCLUSION, we compared ratings within the illusion conditions for trials that supported an additive interpretation versus those that did not. There was a slight increase in ratings for the 'inclusion not possible' over the 'inclusion possible' conditions, but this was not significant (inclusion not possible 3.21, inclusion possible 3.02; $\beta = -.2, \mathrm{SE} = .23, \chi^2(1) = .72, p = .4$). Nonetheless, this pattern fails to support the additive *more* hypothesis, which predicted a substantial difference in the opposite direction.

Discussion This study aimed to determine which of the four hypotheses presented in §2 accounts for the CI effect. Each hypothesis predicted that specific factors would impact the acceptability of CI-type sentences over and above any effects on control sentences.

Our results provide support only for the event comparison hypothesis. This hypothesis predicts that the factor PREDICATE TYPE would reliably impact the acceptability of CI-type sentences, such that those with repeatable predicates would be rated more highly than CI-type sentences with nonrepeatable predicates. This was observed in Experiment 1a.

In light of the long-standing claim that CIs sound highly acceptable, one potentially surprising aspect of our results is that the illusion conditions received much lower mean ratings than did sentences in the control conditions. However, as discussed in some detail in the introduction to §3, the mean rating score obscures the fact that CIs were often as much or more likely to receive a high rating as a low rating.

Nonetheless, it is possible that two features of Experiment 1a could have artificially decreased the ratings for CI-type sentences. First, the *than*-clause subjects in our illusion conditions featured only singular third person pronouns and definite descriptions, differently from the classic illusion in (1). We included this feature of the design in order to be able to manipulate the factor SUBJECT INCLUSION. Yet, third person pronouns and definite descriptions require discourse antecedents in normal usage, which were not accessible in our experiment.

Second, the experiment was 192 trials long, which may have provided participants ample time to notice the anomaly. If this explanation were correct, however, we might expect that the average ratings for CI-type sentences would decline more over time than the average ratings for control sentences. This possibility was not supported by a linear regression analysis: average ratings within the illusion and control conditions remained stable between the first and second halves of the experiment (ORDER × ILLUSION $\beta = .001, \mathrm{SE}\ = .002, \chi^2(1) = .67, p = .41$).

The lack of order of presentation effects in this experiment suggests that the variability we observed in the rating scores for sentences like the classic CI (cf. Figure 1 and 2 above) reflects a probabilistic process: if participants fail to notice the anomaly of the CI on a given trial, they will rate it higher. Otherwise, they will rate it lower.

3.1.2. Experiment 1b

We considered the possibility that the choice of *than*-clause subjects in our stimuli in Experiment 1a (3rd person pronouns and descriptions) may have artificially decreased the ratings for CI-type sentences. The experiment also seemed longer than necessary. Modifying these aspects of study, Experiment 1b tested the predictions of the four hypotheses about the source of the CI effect, in a potentially more favorable experimental context.

All of the *than*-clause subjects in the illusion conditions of this experiment were first person singular pronouns, as in the classic illusion in (1). As before, we manipulated the factors ILLUSION, QUANTIFIER, and ELLIPSIS within items, and PREDICATE TYPE between items. The factor SUBJECT INCLUSION was not manipulated: all of the matrix clause NPs in our experimental items were human-denoting NPs (as in Experiment 1a), and so all of the illusion conditions could in principle support an additive interpretation.

We reduced the number of our experimental items from 48 to 40, and the number of filler sentences from 144 to 100. There were 24 participants.

Results As in Experiment 1a, we found a reliable main effect of ILLUSION, with the control conditions receiving higher ratings than the illusion conditions (control 5.74, illusion 3.26; $\beta = 2.48, \mathrm{SE}\ = .26, \chi^2(1) = 38.5, p < .001$). As before, it is important to flag that the mean ratings of the illusion conditions do not reflect the overall acceptability of CIs; there was a high degree of variability in responses, and many of our manipulations were designed to result in lower ratings.

There was a small but reliable main effect of QUANTIFIER, with the *more* conditions rated higher than the *fewer* conditions overall (*more* 4.6, *fewer* 4.4; $\beta = .19, \mathrm{SE}\ = .08, \chi^2(1) = 4.52, p = .03$). This could reflect a general dispreference for negative quantifiers. More importantly, there was no interaction between QUANTIFIER and ILLUSION ($\beta < -.01, \mathrm{SE}\ = .15, \chi^2(1) < .01, p = .98$), with rating scores for the illusion conditions (*more* 3.36, *fewer* 3.17) different by the same margin as for the control conditions (*more* 5.84, *fewer* 5.64).

These results are consistent with Experiment 1a, and fail to support the syntactic template-matching hypothesis or the additive *more* hypothesis.

We found a reliable main effect of ELLIPSIS, in which the 'ellipsis' conditions were rated more highly than the 'no ellipsis' conditions overall (ellipsis 4.64, no ellipsis 4.36; $\beta = .29, \mathrm{SE} = .09, \chi^2(1) = 8.42, p < .01$). This was likely due to the shorter sentence length in the ellipsis conditions. However, we did not find an interaction between ILLUSION and ELLIPSIS ($\beta = -.09, \mathrm{SE} = .15, \chi^2(1) = .36, p = .55$): the illusion conditions (ellipsis 3.43, no ellipsis 3.09) differed to approximately the same extent as the control conditions (ellipsis 5.86, no ellipsis 5.62). These results are consistent with Experiment 1a in failing to support the repair by ellipsis hypothesis.

We found a reliable main effect of PREDICATE TYPE, with the repeatable conditions rated more highly than the nonrepeatable conditions overall (repeatable 4.81, nonrepeatable 4.19; $\beta = .61, \mathrm{SE} = .12, \chi^2(1) = 20.62, p < .001$). We also found an interaction between ILLUSION and PREDICATE TYPE ($\beta = -.94, \mathrm{SE} = .23, \chi^2(1) = 14.03, p < .001$), with the repeatable illusion conditions rated much more highly than the nonrepeatable illusion conditions (repeatable 3.82, nonrepeatable 2.70) unlike the control conditions (repeatable 5.82, nonrepeatable 5.67). These results are consistent with Experiment 1a in supporting the event comparison hypothesis.

We did not test for effects of SUBJECT INCLUSION in this experiment, since only first person singular pronouns were used as *than*-clause subjects.

Discussion The results of Experiment 1b, like those of Experiment 1a, showed that only the factor PREDICATE TYPE reliably had an effect that selectively impacted the acceptability of CI type sentences. These results are predicted by the event comparison hypothesis.

The mean ratings that we obtained for CI-type sentences remained fairly low. Even among the repeatable illusion conditions, the average rating was higher than in Experiment 1a, but perhaps still not as high as we might have expected given informal reports (Expt.1a 3.47, Expt.1b 3.82). However, behind these means is a wide variability in ratings that, we have suggested, reflects the probability of noticing the anomaly on a given trial: if the anomaly is detected, the CI receives a lower rating; otherwise, it receives a higher rating.

Previously, we considered whether lack of variety in *than*-clause subject types within the illusion conditions could make participants more aware of the CI anomaly. If so, we might expect average scores to have declined over the course of the experiment more for CI-type sentences than for control sentences in Experiment 1b. This happened: a linear regression analysis uncovered a reliable interaction between ORDER and ILLUSION (ORDER × ILLUSION $\beta = .008, \mathrm{SE} = .002, \chi^2(1) = 17.5, p < .001$), with the ratings for the illusion conditions decreasing more over time than did ratings for the control conditions.

This order of presentation effect contrasts with the lack of such an effect in Experiment 1a. In Experiment 1a, the overall lower means for CI-type sentences could have resulted from the fact that many of our items used 3rd person pronouns and definite descriptions, without providing appropriate discourse antecedents. In Experiment 1b, we presented participants only with sentences that more closely tracked the classic CI in (1), with 1st person singular pronouns. However, the repetitive use of this form likely increased the salience of CI-type sentences, leading to substantially decreased scores as the experiment progressed.

In the next experiment, we used a much wider variety of *than*-clause subject types, to see if this would maximally decrease the salience of the CI-type sentences, and thus raise their overall acceptability ratings.

3.1.3. Experiment 1c

Since limiting the range of *than*-clause subject types could have affected participants' likeliness of noticing the CI anomaly, this experiment featured a wide variety of subject types. Our goal was, again, to test the predictions of the four hypotheses about the source of the CI effect in a potentially more favorable experimental context.

Each hypothesis predicts different factors to impact the acceptability of CI-type sentences but not control sentences (the factor ILLUSION). In this experiment, a wide variety of non-bare plural subject types defined the illusion conditions: singular and plural first and third person pronouns and definite descriptions, as well as proper names. As in Experiments 1a,b, we manipulated the factors QUANTIFIER and ELLIPSIS within items, and PREDICATE TYPE between items. We manipulated SUBJECT INCLUSION within the illusion conditions, as in Experiment 1a, such that half of the items would plainly support an inclusion relation with the matrix clause subject NP, and half would not.

Questionnaires consisted of 40 experimental trials and 140 fillers. There were 24 participants.

Results Consistent with Experiments 1a,b, we found a main effect of ILLUSION, with the control conditions rated more highly than the illusion conditions (control 5.82, illusion 3.77; $\beta = 2.04, \mathrm{SE} = .2, \chi^2(1) = 49.17, p < .001$). As with the previous two experiments, we advise caution against considering the means for the illusion conditions as revealing the true acceptability of CIs: they mask a high degree of variability in the responses, and many of our manipulations were designed to decrease the acceptability of CI-type sentences.

As in Experiment 1b, we found a main effect of QUANTIFIER, with the *more* conditions rated more highly than the *fewer* conditions overall (*more* 4.98, *fewer* 4.6; $\beta = .38, \mathrm{SE} = .11, \chi^2(1) = 9.94, p < .01$). However, consistent with Experiments 1a-b, we found no interaction between ILLUSION and QUANTIFIER ($\beta = .02, \mathrm{SE} = .16, \chi^2(1) = .02, p = .9$): with the illusion conditions (*more* 3.95, *fewer* 3.59) differing along this dimension approximately as much as the control conditions (*more* 6.01, *fewer* 5.62). These results fail to support the syntactic template matching hypothesis or the additive *more* hypothesis.

Also as in Experiment 1b, we found a main effect of ELLIPSIS, with the 'ellipsis' conditions rated more highly than the 'no ellipsis' conditions (ellipsis 4.89, no ellipsis 4.69; $\beta = .2, \mathrm{SE} = .1, \chi^2(1) = 3.49, p = .06$). Also consistent with the previous studies, we found no interaction between ELLIPSIS and ILLUSION ($\beta = -.13, \mathrm{SE} = .15, \chi^2(1) = .66, p = .42$), with the difference in the illusion conditions (ellipsis 3.9, no ellipsis 3.64) being approximately equal to the difference in the control conditions (ellipsis 5.88, no ellipsis 5.75). These results fail to support the repair by ellipsis hypothesis.

As in both previous studies, we found a main effect of PREDICATE TYPE, with the repeatable conditions rated more highly than the nonrepeatable conditions overall (repeatable 5.03, nonrepeatable 4.56; $\beta = .47, \mathrm{SE} = .15, \chi^2(1) = 8.26, p < .01$). In this experiment, though, the interaction between ILLUSION and PREDICATE TYPE was less reliable ($\beta = -.5, \mathrm{SE} = .32, \chi^2(1) = 2.42, p = .12$), although the effect was in the predicted direction: the illusion conditions with repeatable predicates were rated more highly than those with nonrepeatable predicates (repeatable 4.13, nonrepeatable 3.41), in contrast to the control conditions (repeatable 5.93, nonrepeatable 5.7). This difference—.72 in the illusion conditions, and .23 in the control conditions—is consistent with the event comparison hypothesis.

As in Experiment 1a, we compared within the illusion conditions those trials that supported an inclusion relation versus those that did not. Unlike in that study, we found a reliable effect of SUBJECT INCLUSION ($\beta = -.71, \mathrm{SE}\ = .3, \chi^2(1) = 5.42, p < .02$), yet not in the direction predicted by the additive *more* hypothesis: participants rated the illusion conditions lower when an additive interpretation was supported than when it was not (inclusion possible 3.46, not possible 4.18). These results are in the same unexpected direction as in Experiment 1a, and fail to support the additive *more* hypothesis.

Discussion Experiment 1c sought to test the predictions of four hypotheses about the source of the CI effect, in an experimental context in which the *than*-clause subject types varied to a high degree. As in Experiments 1a-b, only the factor PREDICATE TYPE reliably impacted the acceptability of CI-type sentences: the repeatable illusion conditions were rated more highly than the nonrepeatable illusion conditions. This result is predicted by the event comparison hypothesis.

The mean ratings for repeatable CIs overall were highest in this experiment (Expt.1a 3.47, Expt.1b 3.82, Expt.1c 4.13). The major difference between Experiment 1c and the previous experiments was in the variety of the *than*-clause subject types within the illusion conditions: Experiment 1a featured only singular proper names, third person pronouns, and definite descriptions, Experiment 1b featured only first person singular pronouns, while Experiment 1c featured all of these types as well as plural variants.

The higher overall means in this experiment are consistent with the idea that a greater variety of *than*-clause subjects can decrease the salience of the CIs: if participants are less likely to notice the anomaly, they are less likely to assign it a lower rating. Support for this possibility comes from the fact that there was no order of presentation effect in this experiment, unlike in Experiment 1b: a linear regression analysis revealed no interaction between order of presentation and the ILLUSION factor (ORDER × ILLUSION $\beta = -.001, \mathrm{SE}\ = .002, \chi^2(1) = .39, p = .53$). As in Experiment 1a, the average ratings for both the illusion conditions and the control conditions remained fairly constant across the experiment.

However, the higher overall means are also consistent with another possibility. This experiment included plural subjects of the *than*-clause in CI-type sentences, which could themselves allow for satisfaction of the event-counting reading. That is, even with a nonrepeatable predicate like *graduate high school*, a plurality of events can be inferred if the subject is plural: one event for each member of the plural subject; cf. (32). Any non-bare plural subject NPs in the *than*-clause of a subject comparative is equally ungrammatical according to the theory discussed in §1.1; yet, if such phrases are plural, they can nonetheless support the inference of a plurality of events.

(32) a. **The girl** graduated high school. [one event]

b. **The girls** graduated high school. [multiple events]

3.2. Interim discussion

Our primary interest in Experiments 1a-c was testing what could be responsible for the CI-effect: the perception that sentences like (1) are acceptable and meaningful, but ultimately seem to have no coherent sense. We tested four factors that were predicted to affect the acceptability of CIs substantially more than that of fully grammatical controls. The results of these manipulations are summarized in Table 2.

Table 2. Means and interaction effects in Experiments 1a-c. The effects of the factor predicate type were in the direction predicted by the event comparison hypothesis. The factor subject inclusion was tested only within the illusion conditions in Experiments 1a and 1c (hence '-' for the control conditions in those experiments); its effect in Experiment 1c was in the opposite direction predicted by the additive *more* hypothesis. The chi-squared column provides effect sizes; '*' indicates that the effect has a p-value of less than .05, and '**' indicates a p-value of less than .01.

	Experiment 1a			**Experiment 1b**			**Experiment 1c**		
Factors	control	illusion	χ^2	control	illusion	χ^2	control	illusion	χ^2
fewer	5.76	2.99	$< .01$	5.65	3.17	$< .01$	5.62	3.59	$< .1$
more	5.99	3.22		5.84	3.36		6.01	3.95	
ellipsis	5.92	3.18	$< .1$	5.86	3.43	< 1	5.88	3.90	< 1
no ellip.	5.83	3.04		5.62	3.09		5.75	3.64	
inclusion	-	3.02	< 1				-	3.46	5.4*
no inclus.	-	3.21					-	4.18	
nonrep.	5.79	2.74	5.6*	5.67	2.70	14.0**	5.70	3.41	2.4
repeat	5.96	3.47		5.82	3.82		5.93	4.13	p=.12

The event comparison hypothesis predicted an interaction between the factors illusion and predicate type. If the CI-effect requires that the predicate be 'repeatable' for a given agent, then CI-type sentences with such predicates should be judged more acceptable than those with nonrepeatable predicates. The only consistently reliable interaction effect that we found in Experiments 1a-c was due to the difference between repeatable and nonrepeatable predicates, supporting the event comparison hypothesis.

The syntactic template matching hypothesis predicted an interaction between the factors illusion and quantifier. If perceiving a CI-type sentence as acceptable involves matching templates that lexically overlap a determiner *more* and an adverbial *more*, then we should have found substantially decreased acceptability for CI-type sentences with *fewer* in the illusion conditions compared to the control conditions, since *fewer* does not have an adverbial use. Yet, comparatives in general tended to receive higher ratings with *more* as opposed to *fewer*, an effect likely due to *fewer* being a negative quantifier and so incurring additional processing costs.

The repair-by-ellipsis hypothesis predicted an interaction between the factors illusion and ellipsis. If the CI effect requires ellipsis in the *than*-clause, then we should have found substantially decreased acceptability in the illusion conditions without ellipsis as compared to the control conditions. However, we failed to find such a pattern; instead, sentences with ellipsis tended to be rated more highly overall. This could have been due to the fact that sentences with ellipsis are shorter, and thus easier to process than comparable sentences without ellipsis.

Finally, the additive *more* hypothesis predicted an effect of the factor SUBJECT INCLUSION within the illusion conditions. If the illusion of acceptability requires that the *than*-clause subject be a possible member of the denotation of the matrix subject (hence permitting a 'just me'-type reading), then the illusion conditions where inclusion was possible should have been rated more highly than those where inclusion was not possible. In fact, any effects we found were in the same unpredicted direction (Experiments 1a,c). This hypothesis also predicted an interaction between the factors ILLUSION and QUANTIFIER, since *fewer* lacks the requisite additive semantics; yet, this effect was not observed.

CI-type sentences were not rated as highly as their fully grammatical and interpretable counterparts, but they were most acceptable across Experiments 1a-c when the understood predicate was repeatable. The mean ratings for CIs was overall highest in Experiment 1c, when a variety of subject-types (singular and plural) appeared in the *than*-clause. One possible explanation for this is that variety decreased the salience of the CIs, and thus made participants less likely to notice the anomaly on a given trial.

Another possibility that we explore next is that the higher ratings in Experiment 1c were due in part to the inclusion of plural *than*-clause subjects. With a plural subject, a plurality of events is possible even with a nonrepeatable predicate: consider *the girls ate pizza* (repeatable; multiple events) and *the girls graduated high school* (nonrepeatable; multiple events). Given the anti-singular semantic requirements of *more*, and given that a plural subject can itself support an event-counting interpretation, this might in turn support heightened acceptability.

In our final judgment study, we tested for effects of properties of the *than*-clause subject directly.

3.3. Experiment 2

In Experiment 2, we tested whether plurality in the subject NP of the *than*-clause in CI-type sentences helps explain the increased acceptability ratings observed for the illusion conditions in Experiment 1c. The event comparison hypothesis explains the effect of the factor PREDICATE TYPE as in terms of speakers analyzing CI-type sentences as a comparison between pluralities of events. However, a plural *than*-clause subject may itself lend plausibility to this interpretation. Experiment 2 thus tested the prediction that the acceptability of CI-type sentences will be directly impacted by the number features of the subject, but not necessarily by other nominal features.

This study had a 12 condition, 2 x 6 design manipulating the factors PREDICATE TYPE and SUBJECT TYPE. The factor PREDICATE TYPE was manipulated between items, and SUBJECT INCLUSION within items. We varied Person (1st versus 3rd), Sort (pronouns versus definite descriptions), and Number (singular versus plural). It was not possible to manipulate these dimensions factorially due to conditions that had to be omitted. The combination of 1st person and definite description (singular or plural) is not possible, and 3rd person plural pronouns are independently unacceptable: speakers naturally interpret *they* in *More girls ate pizza than they did* as anaphoric to *girls*, which sounds contradictory. The bare plural subject condition was included as a fully acceptable control. A guide to the conditions is given in Figure 5.

In light of Experiments 1a-c, we expected a main effect of each of the factors SUBJECT TYPE and PREDICATE TYPE, such that the control condition would be rated more highly

Figure 5. Schema for items in Experiment 2, representing 12 unique conditions. Factors represented are predicate type (repeatable, nonrepeatable), and subject type. The bare plural *boys* marks the control condition.

$$\text{More girls} \left\{ \begin{array}{c} \text{ate pizza} \\ \text{graduated high school} \end{array} \right\} \text{than} \left\{ \begin{array}{c} \text{I} \\ \text{we} \\ \text{the boy} \\ \text{the boys} \\ \text{he} \\ \text{boys} \end{array} \right\} \text{did}$$

than the illusion conditions (i.e. those with non-bare plural subjects), and the repeatable conditions would be rated more highly than the nonrepeatable conditions. We also expected an interaction between the factors subject type and predicate type, in which the repeatable illusion conditions would be rated more highly than the nonrepeatable illusion conditions.

We planned comparisons between subsets of the illusion conditions to test for which properties of the *than*-clause subject would impact acceptability. The Person comparison contrasted the *he* and *I* conditions (both pronominal and singular). The Sort comparison contrasted the *he* and *the boy* conditions (both third person and singular). Two Number comparisons contrasted the *I* and *we* conditions (both pronominal and first person), and the *the boy* and *the boys* conditions (both definite descriptions and third person). Of these comparisons, we expected that only Number would have an effect on the acceptability of the illusion conditions: CI-type sentences with plural subjects would be rated more highly than those with singular subjects.

More generally, the event comparison hypothesis predicts that, within the illusion conditions, the more 'plurals' there are, the more highly a CI-type sentence should be rated. This reflects the possibility that higher ratings are assigned probabilistically on the basis of whether an event-counting reading is supported. This hypothesis thus predicts that sentences with repeatable predicates and plural subjects should be rated the highest of any of the illusion conditions, followed by sentences with either a repeatable predicate or a plural subject, followed by sentences with a nonrepeatable predicate and a singular subject.

We distributed 36 sets of items across 6 lists in a Latin Square fashion. These were combined with 108 filler sentences to create 6 questionnaires. Fillers were designed to be evenly split between sentences that should elicit a low rating and those that should elicit a high rating. Acceptability judgments were recorded on a 7 point scale where 1 is 'unacceptable' and 7 is 'acceptable'. Participants were 24 University of Maryland undergraduates, all native speakers of American English, who received either course credit or $10 for 1 hour of participation. The present study took no more than 30 minutes to complete, and the remaining 30 minutes of participant time were used for unrelated experiments.

Results We first report whether Experiment 2 replicated the major effects from Experiments 1a-c, and then whether any nominal features of the *than*-clause impact the acceptability of CI-type sentences.

First, we found that the control condition was rated more highly than the illusion conditions overall (control 5.86, illusions 4.62; $\beta = 1.24, \mathrm{SE} = .19, \chi^2(1) = 24.29, p < .0001$). Second, we found a main effect of PREDICATE TYPE (repeatable 5.14, nonrepeatable 4.51; $\beta = .48, \mathrm{SE} = .14, \chi^2(1) = 9.58, p < .01$). Third, we found a marginal interaction between ILLUSION and PREDICATE TYPE ($\beta = -.45, \mathrm{SE} = .25, \chi^2(1) = 3.15, p = .08$), wherein the illusion conditions with nonrepeatable predicates were rated substantially lower than those with repeatable predicates (nonrepeatable 4.27, repeatable 4.97), in contrast to the control conditions (repeatable 5.99, nonrepeatable 5.74). These results replicate those of Experiments 1a-c.

Next, we turn to our comparisons between subsets of the illusion conditions.

Comparing the *I* and *he* conditions (Person comparison), we found a marginal effect (*I* 4.56, *he* 4.22; $\beta = -.35, \mathrm{SE} = .18, \chi^2(1) = 3.41, p = .06$), in which the first person singular pronouns (as in the classic example) were rated as more highly acceptable than those with third person singular pronouns. As noted above, this could be due to the lack of discourse antecedents for third person pronouns, which independently decreases acceptability.

Comparing *the boy* and *he* conditions (Sort comparison), we found that they were not reliably different (*the boy* 4.08, *he* 4.22; $\beta = .14, \mathrm{SE} = .19, \chi^2(1) = .52, p = .47$), suggesting that Sort (pronominal or definite) does not significantly impact the acceptability of CI-type sentences.

For the first Number comparison, we compared the *I* and *we* conditions. We did not find a reliable difference between them (*I* 4.56 , *we* 4.99; $\beta = .42, \mathrm{SE} = .26, \chi^2(1) = 2.59, p = .11$). However, we did find a reliable difference between the *the boy* and *the boys* conditions (*the boy* 4.08, *the boys* 5.26; $\beta = 1.17, \mathrm{SE} = .26, \chi^2(1) = 14.83, p < .001$). These results suggest that Number (singular or plural) does render CIs more acceptable.

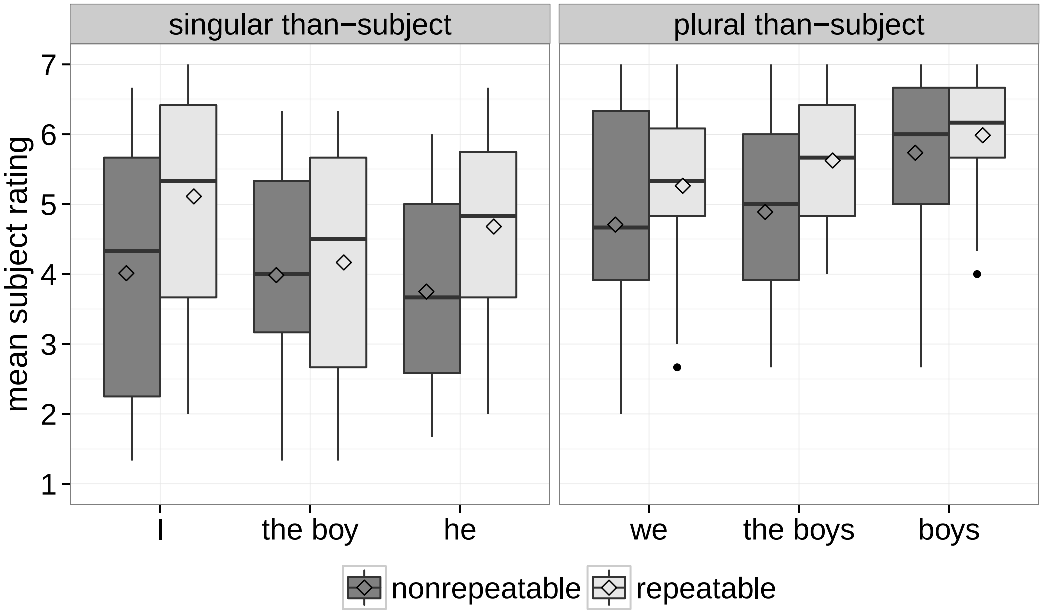
Probing the results further, we conducted a linear regression within the illusion conditions and found a main effect of PREDICATE TYPE (repeatable 4.97, nonrepeatable 4.27; $\beta = .69, \textsc{se} = .14, \chi^2(1) = 17.67, p < .001$), and a main effect of Number (plural 5.12, singular 4.28; $\beta = .84, \textsc{se} = .2, \chi^2(1) = 13.1, p < .001$). This pattern is predicted by the event comparison hypothesis. There was no interaction between Number and PREDICATE TYPE ($\beta = -.09, \textsc{se} = .21, \chi^2(1) = .18, p = .67$).

The results are visualized in Figure 6. In this experiment, as in the previous experiment, we observed a high degree of variability in responses in the illusion conditions, as well as, to some extent, in the control condition. In this figure, the impact of the factors REPEATABILITY and subject plurality are clear.

Discussion This experiment investigated which features of the subject NP in the *than*-clause would impact the acceptability of a CI-type sentence, and how that interacts with the repeatability of the predicate. Plurality of the subject NP and repeatability of the predicate both significantly affected the ratings: the repeatable illusion conditions were consistently rated more highly than were the nonrepeatable illusion conditions, and CI-type sentences with plural subjects were consistently rated more highly than were those with non-plural subjects.

The most highly-rated illusion conditions combined plural subjects and repeatable predicates (mean: 5.44); next highest were the conditions with plural subjects and nonrepeatable predicates (4.8), approximately equaling the conditions with singular subjects and repeatable predicates (4.65); finally, these were followed by the conditions with singular subjects

Figure 6. Boxplots of mean subject ratings by subject plurality and repeatability in Expt. 2. For each column: diamonds indicate the overall mean; heavy lines indicate the median; the upper and lower hinges represent the first and third quartiles; the upper whiskers extend to the highest value within 1.5 times the inter-quartile range of the upper hinges, and the lower whiskers extend to the lowest data point within 1.5 times the inter-quartile range of the lower hinges; black circles represent outlying values.



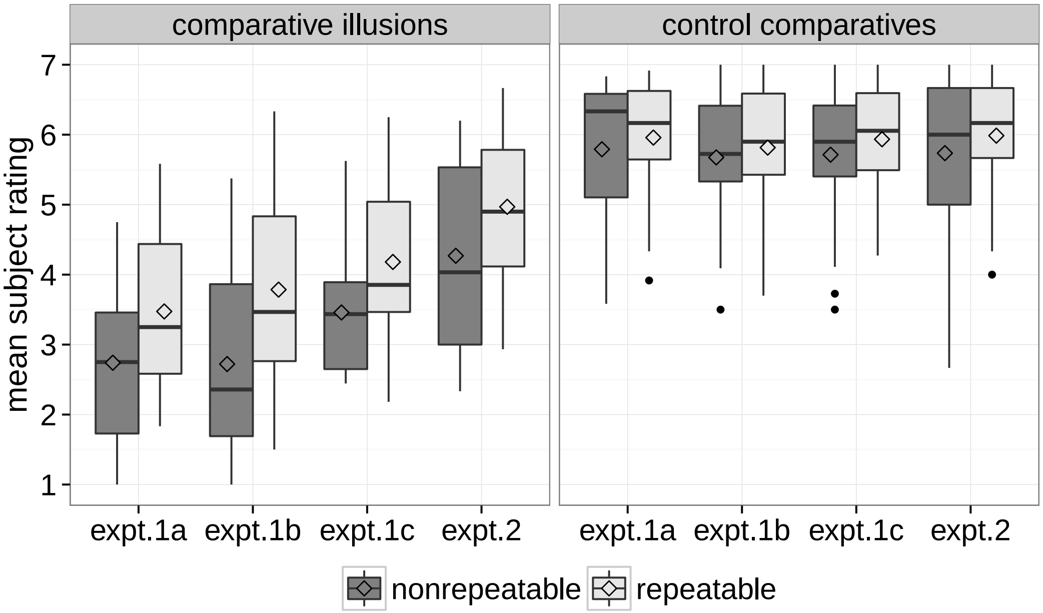
and nonrepeatable predicates (3.92). In fact, CI-type sentences with repeatable predicates and plural subjects reached nearly the level of acceptability of controls.

These results uniformly support the event comparison hypothesis, in which the CI-effect is connected to the interpretation of comparative quantification. The more 'plural' a CI-type sentence is (and thus, the more compatible with the anti-singular semantics of the comparative quantifier), the more likely participants are to judge the sentence as acceptable. Importantly, this effect is not expected under any of the other hypotheses presented in §2. In particular, the syntactic template-matching account does not except fine-grained semantic factors to matter for acceptability, since acceptability is assessed before interpretation.

An important question raised by our acceptability results is: why are even the most obstinately 'singular' sentences (i.e. those with singular *than*-clause subjects and nonrepeatable VPs) still receiving a fairly high average rating (3.92/7)? If a higher acceptability rating for a CI-type sentence depends on its supporting an event-counting reading, the fact that these sentences fail to provide such support might be expected to lead to extremely low acceptability.

We think this is due to how we manipulated repeatability in our items. That is, we know of events like 'graduating high school' that they are once-only, but this is not enforced grammatically. This can be seen by inspecting a sentence like that in (33), which transparently bears the unlikely interpretation (i.e. unlikely given what we know about the nature of such

Figure 7. Boxplots of mean subject ratings by experiment, showing the effects of repeatability in Experiments 1-2.



events). Hence, it is possible that speakers sometimes allow for a predicate like *graduate high school* to support the event-counting reading, which leads them to sometimes consider even our most 'singular' items as more highly acceptable.

(33) Mary graduated high school three times.

3.4. Looking forward

Experiments 1 and 2 showed that the semantic dimension of 'repeatability' in the verb phrase positively impacted the acceptability of CI-type sentences like (1) substantially more than it impacts the acceptability of controls like (2) (this is summarized in Figure 7). Experiment 2 showed that a related dimension in the subject noun phrase of the *than*-clause—plurality—similarly positively impacted the acceptability of CI-type sentences. These effects are predicted by the event comparison hypothesis, and none of the effects predicted by the other hypotheses were borne out.

The interest in the classic illusion in (1) is that it sounds like a well-formed sentence of the language even while one acknowledges that it lacks any clear sense. On the grammatical theory discussed in §1.1, a sentence of this form is predicted to lack a syntactically-licensed interpretation: there is no way to link the covert *how many* with the embedded subject (it can't be hosted by pronouns, proper names, and definite descriptions), and it is not licensed in the verb phrase by the normal rules of ellipsis. Nonetheless, our acceptability judgment data suggest that an interpretation in terms of a comparison of events is, at least temporarily, entertained.

In general, we observed high variability in the rating scores that our experimental participants assigned to CI-type sentences, but not to control sentences. We have hypothesized that this reflects a probabilistic process by which participants are sometimes 'fooled' into thinking that the CI-type sentence is acceptable, just in case they are able to maintain an event-counting reading. Such readings are possible only when the verb phrase of the comparative is repeatable, or its *than*-clause subject is plural. However, participants sometimes notice that this reading isn't licensed syntactically, and thus assign it a lower rating.

While we find these data compelling, acceptability judgment studies remain a fairly indirect method of assessing interpretation. For instance, it could be that repeatable predicates and plural *than*-clause subjects just make CI-type sentences 'sound' better, without playing any interesting role in how participants are interpreting the sentences; perhaps our participants aren't interpreting them at all. Such tasks cannot definitely tell us whether processing CIs importantly involves use of those features (repeatability, plurality) that we have found impact the judgments.

Alternatively, if the acceptability data indeed reveal that our participants rate CI-type sentences more highly because they are entertaining an event-counting interpretation—one that is grammatically licit up to a certain point—then it should be possible to get more direct evidence for that interpretation. Thus, in our last experiment, we investigate CIs in production.

4. Sentence recall

Our acceptability judgment studies supported the event comparison hypothesis: the CI effect arises because speakers are sometimes able to construct an interpretation that satisfies the logical semantics of the comparative construction, while failing to notice when that interpretation is no longer supported by the syntax.

From the perspective of interpretation, the event-counting reading is supported just in case the verb phrase is repeatable or the *than*-clause subject is plural. Our acceptability data reflected the possibility that, in general, the persistence of this reading is probabilistic, and more likely to occur when the stimuli provide more 'plural' expressions (i.e. repeatable predicates, plural *than*-clause subjects). In other words, the more opportunities presented by a CI-type sentence to satisfy the event-counting interpretation, the less likely participants are to notice when that interpretation is no longer supported.

In Experiment 3, we sought more direct evidence for the role of the event-counting reading, by turning to a different type of task—verbatim sentence recall—that places very different demands on speakers. This type of task can potentially be highly useful, since it can tell us not only what participants do when they are asked to produce anomalous CI-type sentences, but it could provide clues as to how the sentences are being interpreted, if at all. We build upon a paradigm developed by Potter & Lombardi (1990), asking: (i) how good are participants at recalling CI-type sentences? And, to the extent that they are reasonably successful, (ii) is there evidence for the event-counting reading in the forms that they are able to recall?

Normally, producing a sentence involves (at least) mapping an intended meaning to some syntactic and phonological form. It has long been observed, however, that production of a previously-presented sentence from memory has strikingly different profiles depending on whether the recall is (roughly) short- or long-term: short-term recall is fairly high-fidelity with respect to the form of the previously-presented sentence, while long-term recall often returns the 'gist', or suitable paraphrase of the meaning of that sentence. This contrast has usually been taken as evidence for two distinct production processes: short-term, verbatim recall, that depends on a stored surface representation of the form; and long-term recall, that depends on the normal processes involved in language production.

Potter & Lombardi (1990) hypothesized, in contrast, that a single set of mechanisms is used for language perception and production: the pathway from perception to production always involves the normal process of storing a meaning, and assigning a form to that meaning at the point of recall. On their theory, the observed differences between short- and long-term recall are due to how active or accessible specific lexical items are at the point of production. In cases of extremely short-term verbatim recall, the words in a target sentence will be more active than any potential competitors. However, activation is fleeting; as the time increases between the presentation of the target sentence and its recall, other words are likely to be more active than the words used in the target.

Potter & Lombardi (1990) found evidence for this hypothesis by manipulating the activation of competitor words at the point of recall in a verbatim sentence recall task. Following the visual presentation of a sentence, a 'list-probe' task required participants to consider a list of 5 words, then judge (yes/no) to whether a subsequently-presented word had appeared in the list. Immediately afterward, participants recalled the initial sentence aloud; the result was that participants recalled the sentence with the lure word replacing its near-synonym on 27% of trials in which it was present in the list. This finding is not compatible with the possibility that speakers rely on surface-based representations, since there would be no explanation for how a new word came to be incorporated into such a representation. (Potter & Lombardi found the same effect even when the list-probe task occurred prior to the presentation of the target sentence, suggesting that the first result didn't merely amount to the difference between short- and long-term memory.)

We explained the patterns in our acceptability data in terms of participants' being 'fooled' by CI-type sentences just in case they maintained an event-counting interpretation. Such a meaning is consistent with the logical semantic requirements of the comparative, but not with the syntax of the CI. Nevertheless, if participants can store an event-counting interpretation when they encounter a CI, then there should be evidence for that meaning in a sentence recall task. In contrast, if they are not able to store this or any other meaning for the sentence, then recall should be difficult.

Our extension of this methodology represents, to our knowledge, the first time that a sentence recall task has been used to probe the production of syntactically and semantically anomalous sentences. This could help shed light on what choices speakers make in situations where they are asked to find a meaning for a sentence that literally doesn't have one.

4.1. Experiment 3

We investigated whether semantic plurality (repeatability in the VP, plurality in the *than*-clause subject NP) is the dimension relevant to how speakers interpret CI-type sentences, by investigating how they are produced in a verbatim sentence recall task.

There are two dimensions along which we can make predictions as to how acceptability could pattern with recall in this task. So far, we have hypothesized that the degree to which a given CI-type sentence supports the event-counting reading correlates with the likelihood that a participant will assign the sentence a higher rating. Moving to a production-based study, this could predict that those CI-type sentences which better support the event-counting reading should be easier to recall than those that do not, since a regeneration of form at recall is possibly only if a meaning can be stored for it.

Further, we assume that participants will attempt to recover a meaning for a target sentence if it is at all possible. Moreover, participants should try harder in a situation where the sentence is harder to interpret, or less acceptable. If so, we predict that those CI-type sentences that do not support the event-counting reading should nevertheless sometimes be recalled as though they do. This possibility is supported by the fact that even obstinately 'singular' CI-type sentences (those with nonrepeatable VPs, and singular subject NPs in the *than*-clause) sometimes received higher ratings. Since the grammatical properties of nonrepeatable predicates in English do not absolutely impose a 'singular' interpretation, it may be that speakers at times construe such predicates as though they were repeatable.

If this line of reasoning is correct, then we expect to find a pattern of 'changes', or errors, in production, that correlates with the patterns we saw in the acceptability data: more errors on the illusion conditions than the control conditions (a main effect of ILLUSION), more errors on the nonrepeatable conditions than on the repeatable conditions (a main effect of PREDICATE TYPE), and overall the most errors on the nonrepeatable illusion conditions.

Alternatively, it could be that the CI effect straightforwardly involves syntactic reanalysis, in which a CI-type sentence is recalled with *more* in an adverbial rather than determiner position. This possibility was raised in the brief discussion of the syntactic version of the event comparison hypothesis in §2. If this alternative is correct, then we should find that participants displace the comparative quantifier in recall more in the illusion conditions than in the control conditions.

4.1.1. Design

As in the acceptability experiments, we manipulated the factors ILLUSION (illusion, control) and PREDICATE TYPE (repeatable, nonrepeatable). In addition, we created two types of items that differed in which parts of the sentence determined the repeatability of the predicate.

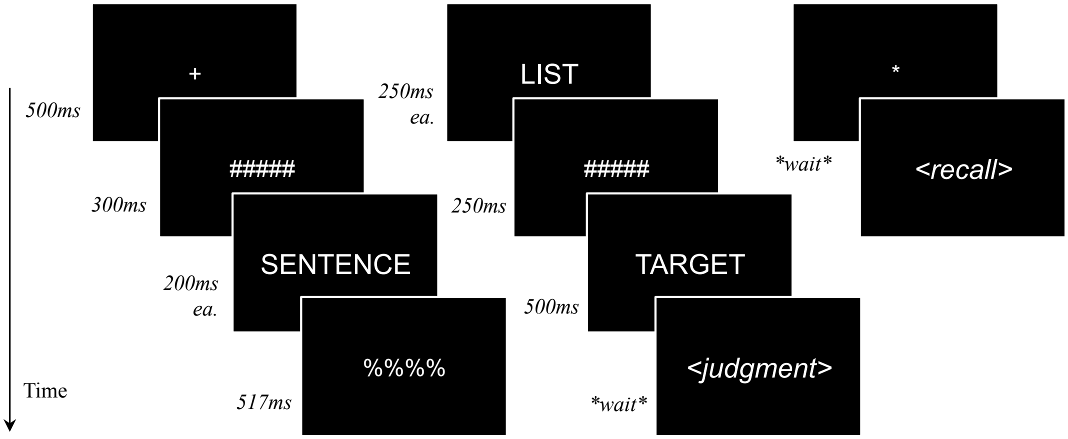
In one set of items, we manipulated repeatability through the aspect of the predicate (ASPECT items). These sentences were classified as nonrepeatable if they had an initiative or terminative aspectual verb introducing their VP, and repeatable if they had a continuative aspectual verb or a form of *be*, (34). (These are simplified examples intended to illustrate the relevant contrasts.)

(34) ASPECT contrast
 a. Mary {**started, finished**} reading the book. [nonrepeatable VP]
 b. Mary {**continued, was**} reading the book. [repeatable VP]

OBJECT items were classified as nonrepeatable if they had an ordinal modifier, and repeatable otherwise, (35).

(35) OBJECT contrast
 a. Mary ate her {**first, last**} cupcake. [nonrepeatable VP]
 b. Mary ate her {**tasty, strawberry**} cupcake. [repeatable VP]

FIGURE 8. Schematic presentation of procedure from Experiment 3.



4.1.2. Procedure

Our procedure follows that laid out in Potter & Lombardi (1990) exactly, and is summarized in Figure 8. First, a fixation cross appeared for 500ms, followed by a visual mask for 300ms. In the Sentence phase, the words of a sentence appeared in rapid serial visual presentation mode (RSVP), with a presentation duration of 200ms each, followed by a visual mask for 517ms. In the Distractor phase, a list of five words appeared RSVP for 250ms per word, ending with a visual mask for 250ms. Next, a capitalized word appeared for 500ms, at which point participants were asked to judge whether that word was in the immediately preceding list, pressing F for 'yes' or J for 'no'. After making this judgment, the Recall phase began, signalled by a visually-presented asterisk. Participants were given as much time as needed in the Recall phase. The experiment proper was preceded by 6 practice trials to insure familiarity with the procedure.

4.1.3. Stimuli

All of our experimental sentences were between 11 and 17 (mean 14.2) words long, in order to ensure that verbatim recall would be somewhat difficult. Sample items from the experiment are given in Figure 9. In all of our items, *more* was preceded by an unrelated adverbial phrase; we included this aspect of the design to avoid having *more* occur first in the sentence, which might independently reduce the likelihood that participants would displace *more* to an adverbial position. This design feature provides the best environment for testing the syntactic version of the event comparison hypothesis. All of our illusion conditions had singular *than*-clause subjects.

Lists of words for the distractor task were constructed out of sets of 5 words matched for character length (3-7 characters per word), and we minimized their phonological and semantic similarity to each other, and to the elements of the sentence they were paired with. The target word was present in the list of words on only half of the trials, for an expected 50/50 split in 'yes' and 'no' responses.

24 sets of 4 items were distributed across 4 lists in a Latin Square design, and then combined with 90 filler sentences. Fillers contained no ungrammatical or anomalous sentences,

Figure 9. Schemata for experimental items in Experiment 3, each representing 4 unique conditions. In Aspect items, predicate type was manipulated at the point of an aspectual verb (repeatable *continued*; nonrepeatable *began*). In Object items, it was manipulated at the point of the verbal object (repeatable *a charming haiku*; nonrepeatable *their first haiku*). *W&P* is an abbreviation of *War and Peace*. The abbreviation is simply to accommodate the sentence graphically; there were no abbreviated items in the experiment.

Aspect item

$$\text{Last year more young people} \left\{ \begin{array}{c} \text{continued} \\ \text{began} \end{array} \right\} \text{reading W\&P than} \left\{ \begin{array}{c} \text{the old man did.} \\ \text{old men did.} \end{array} \right\}$$

Object item

$$\text{In English class more girls wrote} \left\{ \begin{array}{c} \text{a charming} \\ \text{their first} \end{array} \right\} \text{haiku than} \left\{ \begin{array}{c} \text{the boy did.} \\ \text{boys did.} \end{array} \right\}$$

and were comprised of 36 comparative-type sentences (e.g. equative, superlative, etc.) and 54 non-comparative-type sentences. The order of presentation was randomized within each list for each participant. The experiment was implemented in DMDX (Forster & Forster 2003).

4.1.4. Error coding

We coded overall failure of recall, as well as two broad categories of errors: movement and non-movement. To illustrate these, we use simplified examples (i.e. not actual experimental items) to make the relevant difference between target and recall for each error type as transparent as possible. We discuss at the end of this section the types of errors that occurred but which were not coded.

Recall failure A trial was coded as a recall failure if the response failed to contain a comparative sentence. This included complete silence, an utterance like ‘I forget’, or, for example, ‘Boys did something’ for a target like *More boys did X than girls did.*

Movement errors A response was classified as a movement error if the nominal determiner *more* was recalled in an adverbial or direct object position (36). The syntactic version of the event comparison hypothesis predicts that the comparative quantifier should be displaced more in the illusion conditions than in the control conditions. As far as more specific predictions, it is not entirely clear. We might expect *more* would be moved to an adverbial position at a higher rate in the repeatable illusion conditions than in the nonrepeatable illusion conditions, since in the latter case the result would be ungrammatical. However, we might expect that *more* would be displaced to the direct object position at a higher rate in the nonrepeatable illusion conditions, since in this case there is more motivation to correct the representation. (Note that, in some cases, participants produced *more* both within the subject phrase and in an adverbial or direct object position; these errors were coded as movement errors.)

(36) **Moving *more* error**
More girls ate pizza than I/boys did. →
Girls ate pizza **more** than I/boys did. [adverbial recall]
Girls ate **more** pizza than I/boys did. [DO recall]

Non-movement errors A response was classified as a non-movement error if the target sentence was recast along one of the following dimensions, which are important in light of the semantic version of the event comparison hypothesis.

NP number error. This type of error involved recalling the subject of the *than*-clause in a different number (singular or plural) than the target sentence. Within the illusion conditions, this renders a singular NP subject as plural, (37). Within the control conditions, this involves rendering the plural NP subject of the *than*-clause as singular, (38). The semantic version of the event comparison hypothesis predicts more NP number errors in illusion trials than in control trials. This error was coded both for Aspect and Object items. Note that not all of our items had definite descriptions in the *than*-clause; for the purposes of coding, this error ignores whether or not the determiner was retained on those trials. (We discuss results pertinent to pluralizing and deleting the determiner, resulting in a fully grammatical sentence, in the discussion.)

(37) **NP number error** [singular → plural]
More girls ate pizza than the **boy** did. →
More girls ate pizza than (the) **boys** did.

(38) **NP number error** [plural → singular]
More girls ate pizza than **boys** did. →
More girls ate pizza than {the/a/some} **boy** did.

VP number error. This type of error involved recalling the verb phrase with a different repeatability status (repeatable or nonrepeatable) than the target sentence, and was only coded for the Aspect items. Within the nonrepeatable conditions, it modifies the VP so that it is potentially repeatable: this involved changing an initiative or terminative verb to a copular or continuative verb, (39). Within the repeatable conditions, it modifies the VP so that it is nonrepeatable: this involved recalling a VP with a continuative or copular verb with an initiative or terminative verb, (40). The semantic event comparison hypothesis predicts more NP number errors on nonrepeatable trials than on repeatable trials.

(39) **VP number error** [nonrepeatable → repeatable]
More girls {**began, finished**} reading the book than the boy did. →
More girls {**continued, were**} reading the book than the boy {did/was}.

(40) **VP number error** [repeatable → nonrepeatable]
More girls {**continued/were**} reading the book than the boy {did/was}. →
More girls {**began, finished**} reading the book than the boy did.

Modifier deletion error. This type of error involved deleting an adjective (critical or non-critical) in the direct object position of the matrix clause, and was coded for only within the Object items. For the nonrepeatable conditions, deletion of this adjective critically renders the VP potentially repeatable, (41). For the repeatable conditions, deletion of the adjective has no effect on the repeatability of the predicate, (42). The semantic event comparison hypothesis predicts more modifier deletion errors on nonrepeatable trials than on repeatable trials.

(41) **Modifier deletion error** [nonrepeatable → repeatable]
More girls ate their **first** strawberry cupcake than the boy did. →
More girls ate their/a strawberry cupcake than the boy did.

(42) **Modifier deletion error** [no effect on repeatability]
More girls ate a **tasty** strawberry cupcake than the boy did. →
More girls ate their/a strawberry cupcake than the boy did.

We did not code for errors that were irrelevant to the hypotheses under consideration. For example, participants often substituted lexical items that were semantically similar (e.g. *assignment → paper*, *hockey fan → basketball fan*, etc.), more rarely with functional expressions (e.g. *a glass → one glass*, *drank → didn't drink*), and they deleted non-critical adjectives (i.e. those not in the matrix clause VP; for example, *than the young Spaniard did → than the Spaniard*).

4.1.5. Participants

34 University of Maryland undergraduates participated in this task, all native speakers of American English as determined in a pre-test questionnaire. The study took no more than 30 minutes to complete, and the remaining 30 minutes of participant time were used for unrelated experiments. 10 participants were excluded for failure to successfully follow task instructions (3 participants) or due to technical problems that lead to failure to record responses (7 participants). We report the results of 24 participants, for a total of 552 verbal responses to our experimental items that were recorded and coded.

4.1.6. Results

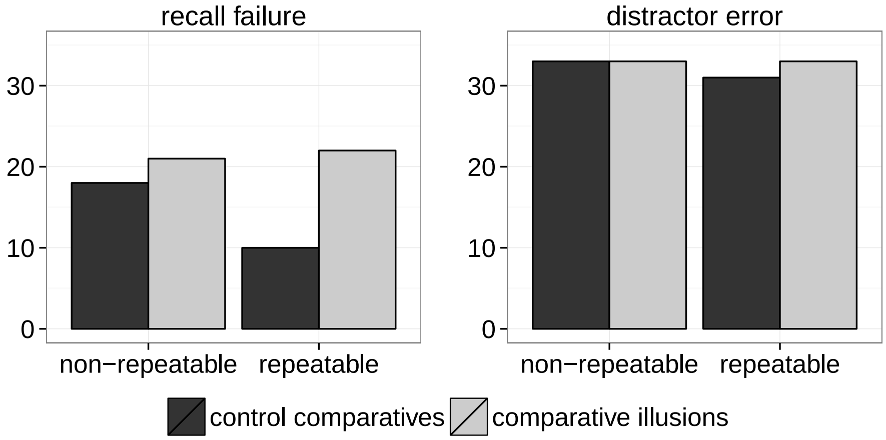
We investigated whether overall rate of recall, distractor task accuracy, movement errors, or non-movement errors would distinguish between illusion and control production targets, and which would correlate qualitatively with our acceptability data.

The statistics we report below are the result of two types of analysis, either logistic or linear mixed effects regressions, with maximal random effects terms where possible (Barr et al. 2013). We used logistic regressions when considering binary response data (e.g. a particular type of error occurred on a given trial, or not). We used linear regressions when considering summed response data (e.g. the number of total errors on a given trial). We proceed with the logistic analysis as the default; when we present the results of a linear analysis, we indicate this explicitly.

First, we examined the rate of global failure in recall. We found an effect of the factor ILLUSION ($\beta = .68, \text{SE} = .31, \chi^2(1) = 5.12, p = 0.024$), due to greater failure to recall on the illusion trials than on the control trials (see Figure 10). We found no effect of the factor PREDICATE TYPE ($\beta = .32, \text{SE} = .37, \chi^2(1) = .7, p = .4$), as participants failed to recall the nonrepeatable and repeatable targets at around the same rate. There was no interaction between the factors ILLUSION and PREDICATE TYPE ($\beta = -.9, \mathrm{SE} = .61, \chi^2(1) = 2.18, p = .14$).

Error rates in the distractor task were the same in the illusion and control conditions (Figure 10). Excluding those trials on which participants failed at recall, we found no difference in distractor error by the factor ILLUSION ($\beta = .13, \mathrm{SE} = .23, \chi^2(1) = .32, p = .57$) or by PREDICATE TYPE ($\beta = .1, \mathrm{SE} = .26, \chi^2(1) = .15, p = .7$), nor was there any interaction between these factors ($\beta = -.17, \mathrm{SE} = .48, \chi^2(1) = .13, p = .72$).[3]

FIGURE 10. Counts of recall failure and distractor task errors in Experiment 3.



Turning to the counts of errors within successful recall trials, we found a main effect of the factor ILLUSION: a total of 260 errors were identified on 232 illusion recall trials, but a total of 113 errors on 246 control recall trials (Figure 11; linear: $\beta = .69, \mathrm{SE} = .10, \chi^2(1) = 25.32, p < .001$). We found no effect of the factor PREDICATE TYPE (linear: $\beta = .06, \mathrm{SE} < .1, \chi^2(1) = .38, p = .5$) and no interaction between the factors ILLUSION and PREDICATE TYPE (linear: $\beta = .1, \mathrm{SE} = .13, \chi^2(1) = .8, p = .4$).

Within these errors, we found marginally more movement errors on the illusion trials than on the control trials (illusion 44, control 27; $\beta = .9, \mathrm{SE} = .48, \chi^2(1) = 3.43, p = .064$; Figure 12, first panel). There was no effect of the factor PREDICATE TYPE for this type of error ($\beta = -.36, \mathrm{SE} = .5, \chi^2(1) = .5, p = .48$), and no interaction between this factor and ILLUSION ($\beta = -.2, \mathrm{SE} = .9, \chi^2(1) < .1, p = .8$).

We observed a similar pattern for non-movement errors, except there were many more errors of this type for the illusion targets than for the control targets (Figure 12, second panel; control 86, illusion 216; linear: $\beta = .61, \mathrm{SE} = .097, \chi^2(1) = 24.04, p < .001$). There were, however, no significant effects of PREDICATE TYPE at this level of categorization (linear: $\beta = .1, \mathrm{SE} = .08, \chi^2(1) = 1.5, p = .2$), and no interaction between the factors PREDICATE TYPE and ILLUSION (linear: $\beta = .1, \mathrm{SE} = .1, \chi^2(1) = .87, p = .35$).

Next, we turn to the specific subtypes of non-movement errors.

With respect to *than*-clause subject number errors (NP number), we found many errors in the illusion conditions (nonrepeatable 61, repeatable 53), and virtually none in the control conditions (nonrepeatable 1, repeatable 0; Figure 13, first panel). This resulted in a main

[3] This and subsequent analyses exclude an additional 3 trials on which DMDX failed to record the participants' responses to the distractor task.

Figure 11. Counts of errors logged on successful recall trials by condition in Experiment 3.

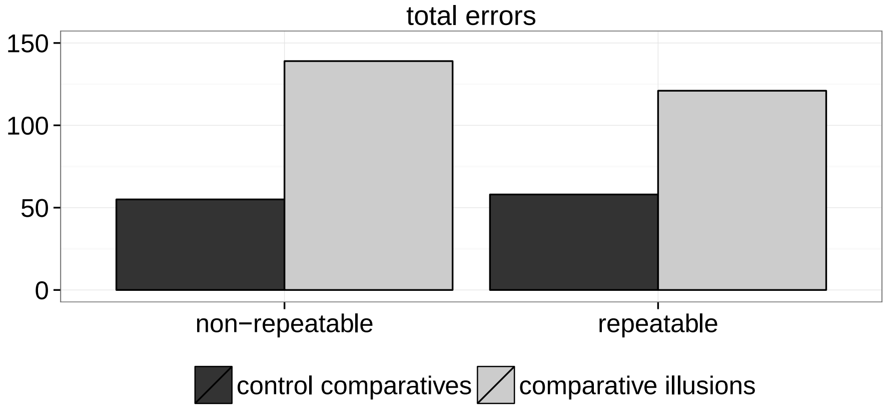
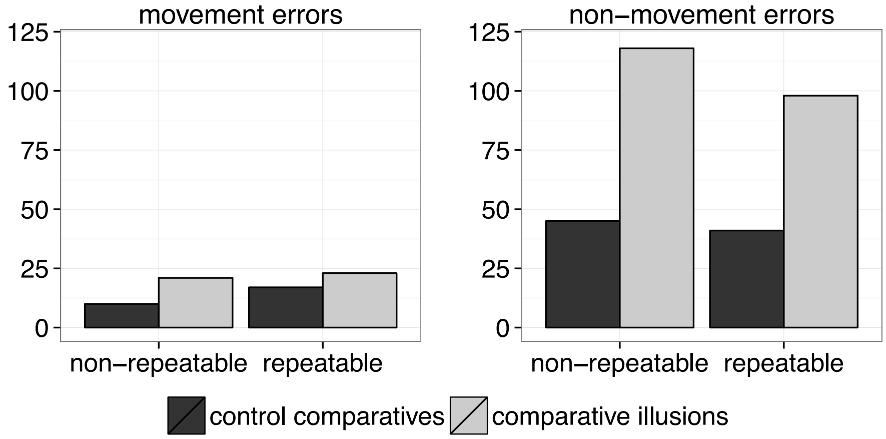


Figure 12. Counts of movement and non-movement errors by condition in Experiment 3.

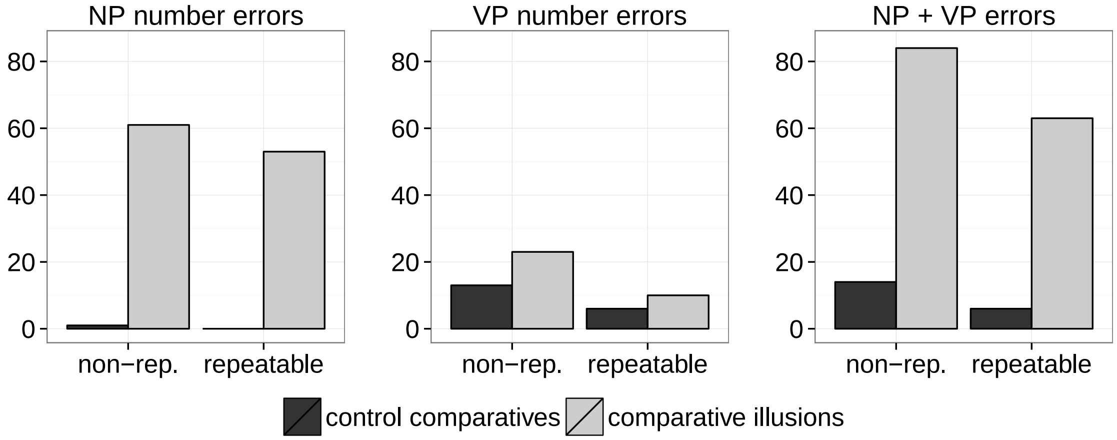


effect of the factor illusion ($\beta = 8.78, \mathrm{SE} \ = 1.9, \chi^2(1) = 231.6, p < .001$). The paucity of errors in the control conditions violated the assumptions of the logistic regression, and we could not analyze the factor predicate type on the full dataset. Comparing just within the illusion conditions, however, we found no effect of this factor on NP number errors ($\beta = .4, \mathrm{SE} \ = .3, \chi^2(1) = 1.7, p = 0.2$).

Turning to repeatability errors (VP number), coded only within the Aspect items, we found a main effect of the factor illusion ($\beta = 1.1, \mathrm{SE} \ = .5, \chi^2(1) = 4.7, p = .03$), with more errors made on illusion than control targets (illusion 33, bare plural 19; Figure 13, second panel). We also found a main effect of predicate type ($\beta = 1.53, \mathrm{SE} \ = .68, \chi^2(1) = 5.0, p = .03$), with more errors made on nonrepeatable targets than on repeatable targets (nonrepeatable 36, repeatable 16). There was no interaction between the factors illusion and predicate type ($\beta = .16, \mathrm{SE} \ = 1.04, \chi^2(1) < .1, p = .9$).

Inspecting Figure 13, it appears qualitatively that the error pattern considered over the conjunction of NP and VP errors (third panel) matches what we observed in our acceptability

Figure 13. Counts of number errors in Experiment 3. NP number errors were pluralizing for illusion targets, and singularizing for control targets; VP number errors made repeatable targets nonrepeatable, and nonrepeatable targets repeatable.



studies: the most errors were observed in the nonrepeatable illusion condition, followed by the repeatable illusion condition; and there was a marginal difference between the nonrepeatable and repeatable control conditions. Statistically, this difference is reflected in a main effect of illusion (linear: $\beta = .58, \mathrm{SE} = .09, \chi^2(1) = 25.2, p < .001$) and of predicate type (linear: $\beta = .12, \mathrm{SE} = .06, \chi^2(1) = 3.56, p = .059$), but not in an interaction (linear: $\beta = .1, \mathrm{SE} = .08, \chi^2(1) = 1.6, p = .2$).

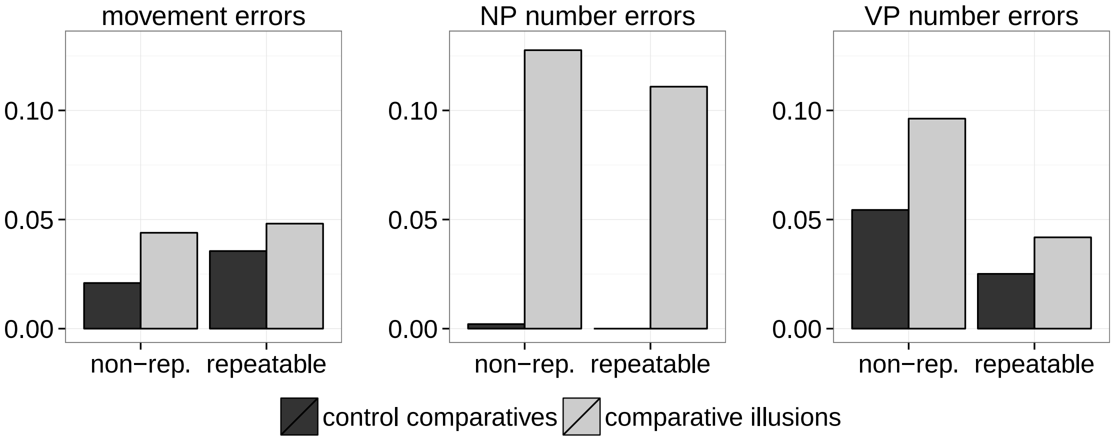
With respect to modifier deletion errors, coded only within the Object items, there were no effects: illusion ($\beta = .22, \mathrm{SE} = .3, \chi^2 = .52, p = .47$), predicate type ($\beta = -.23, \mathrm{SE} = .44, \chi^2(1) = .26, p = .61$), illusion × predicate type ($\beta = -.08, \mathrm{SE} = .63, \chi^2(1) < .1, p = .89$). That is, the critical adjective was deleted at the same rate across the board.

4.1.7. Discussion

Experiment 3 investigated CI-type sentences in production. We found that participants failed to recall a sentence with a comparative form more on the illusion conditions than on the control conditions (illusion 43/276, control 28/276). Given our assumption (following Potter & Lombardi 1990) that the task demands in this experiment require regeneration of a form to go with a stored meaning, this could suggest that it was more difficult to store a meaning for CI-type sentences. We did not find that the nonrepeatable illusion conditions were more difficult to recall than the repeatable illusion conditions, however.

On successful recall trials, we found that participants made repeatability and number changes substantially more on the illusion conditions than on the control conditions. Participants were also more likely to change the repeatability of the predicate in the nonrepeatable conditions than in the repeatable conditions (Aspect items). This pattern is consistent with the semantic version of the event comparison hypothesis, and provides a clear link between the acceptability and recall data: the less acceptable the comparative sentence, the more modifications required in order to successfully store a meaning for that sentence. Importantly, these modifications were essentially semantic in nature; we failed to find an error pattern with moving *more* that corresponded to the acceptability pattern (Figure 14).

FIGURE 14. Proportion of errors in Experiment 3. The denominator used for each error type equaled the number of trials for which it was coded: the full data set for movement and NP number errors (478 trials), and the ASPECT trials for VP number errors (239 trials).



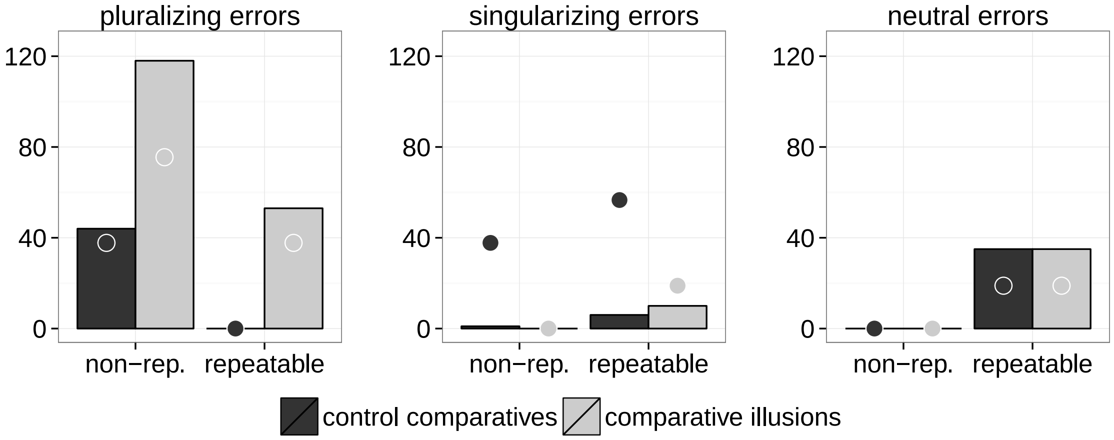
Results of our test within the OBJECT items probing for modifier deletion as a function of PREDICATE TYPE were not conclusive. In the nonrepeatable trials, the adjective was a semantically complex expression like *first*, whereas in repeatable trials it was a relatively simple adjective like *tasty*. This relatively lower rate of deletion of the complex adjective could reflect a tension between wanting to retain a highly semantically informative expression, with the fact that its retention would deliver a nonrepeatable event description. Regardless, we found that the modifier was deleted at approximately the same rate across conditions.

Do these results reflect participants' attempts to assign a meaning to an unacceptable sentence so that it can be recalled, and is 'plurality' the right dimension for fixing on that meaning? We think there are reasons to believe that the answer to both of these questions is 'yes', although our comments will be mainly speculative.

The first reason is that errors that made a representation 'more plural' were much more frequent in our data than were errors that made a representation 'more singular', or which were neutral with respect to number (i.e. deleting an adjective like *tasty* in the OBJECT conditions), and these occurred at a higher rate than would be expected purely on the basis of how often the opportunity to make the error was presented. This is shown in Figure 15, with the counts across conditions of non-movement errors divided into each of these categories, and plotted along with hypothetical counts if the errors were distributed solely based on how many opportunities there were for making that error.

The second reason concerns how often participants rendered a CI-type sentence fully grammatical, i.e. as a subject nominal comparative with a bare plural in its *than*-clause. If the repeatable illusion conditions already provide the elements that are needed for the event-counting reading, then participants needn't resort to such a modification of a target sentence as often as they might on the nonrepeatable illusion conditions. We found that, out of our 179 illusion trials in which the target sentence's *than*-clause contained a definite

Figure 15. Counts of observed errors in Experiment 3, categorized according to whether they were 'pluralizing', 'singularizing', or 'neutral' with respect to number (bars; 302 observations). The observed counts are compared to hypothetical counts (dots), calculated as the number of observed errors distributed as a proportion of the number of opportunities available to make each kind of error.



description, 48/87 (55.2%) of the nonrepeatable trials were rendered grammatical and 42/92 (45.7%) of the repeatable trials were rendered grammatical.

Overall, these results suggest that participants attempt to assign a meaning to CI-type sentences. The types of meanings that they assign are ones that render the sentence more compatible with the logical semantics of the comparative quantifier. That this process is essentially semantic, and not syntactic in nature, is supported by the fact that, very often, our participants left *more* in its subject syntactic position.

5. General discussion

This paper represents the first systematic attempt to understand the source of the illusory effect of 'Escher sentences'. Such sentences have been reported to be remarkably acceptable to speakers of English, despite having no coherent sense, and no grammatical analysis according to contemporary syntactic and semantic theories of comparatives. Early consideration of the phenomenon lead researchers to suggest that they reflect a sort of 'shallow' processing—speakers fail to notice the anomaly, because they aren't really attending to the grammar of the sentence anyhow. In contrast, our results suggest that fine-grained semantic properties play a role in determining how acceptable CI-type sentences are to speakers.

We first set out to address the questions: how robust are the illusions? And, how well does their reported acceptability stand up in a formal experimental context? The results of our acceptability studies (Experiments 1a-c, and Experiment 2) suggest that participants notice the anomaly only probabilistically, assigning it a higher rating on a trial in which they fail to notice the mismatch between form and meaning, and a lower rating when they do notice. Generally, this happened on a trial-by-trial basis, except under conditions where the highly repetitive nature of the stimuli likely drew participants' explicit attention—for example, the repetition of forms like *than I have* or *than I did* in Experiment 1b—leading to an increase in lower ratings in the second half of the experiment.

Next, we asked how far the effect generalizes beyond the canonical example in (1). Informants and linguists alike have informally suggested various accounts of what drives the illusion, making different predictions as to how far it should generalize. For instance, it might be due to a shallow processing mechanism: a cursory analysis finds that the CI matches familiar clausal templates, and so the sentence is judged acceptable (§2.1). Instead, it might be due to some form of repair-by-ellipsis: the syntactic problem with CIs is somehow eliminated from detection, roughly analogous to other familiar examples (§2.2). Or perhaps speakers do interpret the sentence, just assigning it a meaning that its form can't support: either they misanalyze the comparative quantifier *more* as the homophonous additive meaning (§2.3), or they persist in an interpretation in terms of a comparison of numbers of events (§2.4).

Our four acceptability experiments tested these hypotheses (§3), and found evidence only for the event comparison hypothesis. The acceptability of CI-type sentences was only positively impacted when its predicate could be interpreted as repeatable (i.e. as involving the kinds of events that a single individual can participate in multiple times) or when an otherwise-ungrammatical subject type is plural (and thus provide multiple events via multiple agents). This interpretation is grammatically legitimate in the matrix clause, but it is not supported by the syntax of the *than*-clause. Thus, we find that the source of the illusion lies in a failure to notice that the event comparison reading, however tempting, is not a literal interpretation of the sentence.

Our sentence recall task probed the production of CIs, in a novel application of the sentence recall task to anomalous sentences (§4). Examining the patterns of changes made between target and recall, we found evidence that adjudicated between the semantic and syntactic versions of the event comparison hypothesis. Speakers' attempts to rescue the CI from uninterpretability tended to involve 'pluralizing' the representation, rendering the sentence more consistent with the logical semantic requirements of the comparative. In contrast, speakers only rarely displaced the comparative quantifier from a determiner to an adverbial or direct object position; suggesting that, in general, our participants were highly faithful to the syntax of the construction, while nevertheless persisting in an event comparison interpretation.

The illusion involves entertaining the event interpretation early on during the main clause, potentially as early as the matrix subject is encountered. Ultimately, this interpretation is so tempting that comprehenders are blinded to the fact that the syntax doesn't literally support it: it is not syntactically possible to posit either a silent determiner *how many* or adverbial *how much* in the *than*-clause. Because of the syntax-semantics mismatch, the event comparison interpretation is not fully stable, or always accessible. This work doesn't address the question of precisely at what point the event-counting reading becomes attractive, however.

A result reported by Fults & Phillips 2004 potentially speaks to this question. Their Experiment 1 involved contrasting CIs with extraposed *than*-clauses (as in the classic example in (1)) and unextraposed (e.g. *More people than I have have been to Russia*), finding substantial degradation when extraposition had not occurred (extraposed 3.58/5, unextraposed 2.87/5). It is possible that only the individual-counting reading is entertained at the point of the matrix subject, and the event-counting reading becomes attractive only once a repeatable predicate is encountered. That is, if individual comparison is immediately ruled out at *than*, the event comparison reading doesn't have a chance.

This study is situated within the context of discussion about the need for superficial interpretive mechanisms that are in some sense distinct from the process of analysis that formal semanticists generally worry about. In contrast to 'good enough' approaches (see Ferreira & Patson 2007 for an overview), the acceptability facts that we uncovered reveal that speakers interpret CI-type sentences deeply: they are judged acceptable only to the extent that they satisfy the semantic requirements of the comparative construction. Moreover, the results of our sentence recall experiment show that this process does not require fixing the structure to suit the meaning; rather, our participants were syntactically faithful, even while modifying elements of the sentence to better support that meaning.

References

Barker, Chris. 1999. Individuation and quantification. Linguistic Inquiry 30(4). 683–691.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68. 255–278.

Bates, Douglas, Martin Maechler, Benjamin M. Bolker & Steven Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. http://CRAN.R-project.org/package=lme4.

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. In J. R. Hayes (ed.), Cognition and the development of language, 279–362. Wiley.

Bock, Kathryn & Carol A. Miller. 1991. Broken agreement. Cognitive Psychology 23. 45–93.

Bresnan, Joan. 1973. Syntax of the comparative clause construction in English. Linguistic Inquiry 4(3). 275–343.

Chomsky, Noam. 1977. Conditions on transformations. In Essays on form and interpretation, 81–162. New York, New York: Elsevier North-Holland, Inc.

Christianson, K., A. Hollingworth, J. Halliwell & F. Ferreira. 2001. Thematic roles assigned along the garden path linger. Cognitive Psychology 42. 368–407.

Clifton, Jr., Charles, Lyn Frazier & Patricia Deevy. 1999. Feature manipulation in sentence comprehension. Rivista di Linguistica 11. 11–39.

Ferreira, F., V. Ferraro & K. Bailey. 2002. Good enough representations in language comprehension. Current Directions in Psychological Science 11(1). 11–15.

Ferreira, F. & N.D. Patson. 2007. The 'good enough' approach to language comprehension. Language and Linguistics Compass 1(1-2). 71–83.

Forster, K. I. & J. C. Forster. 2003. DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers 35. 116–124.

Frazier, Lyn & Charles Clifton, Jr. 2011. Quantifiers undone: reversing predictable speech errors in comprehension. Language 87(1). 158–171.

Fults, Scott & Colin Phillips. 2004. The source of syntactic illusions. CUNY 2004 poster.

Grant, Margaret Ann. 2013. The Parsing and Interpretation of Comparatives: More than Meets the Eye. Amherst, MA: University of Massachusetts-Amherst dissertation.

Greenberg, Yael. 2010. Additivity in the domain of eventualities (or: Oliver Twist's *more*). In Martin Prinzhorn, Viola Schmitt & Sarah Zobel (eds.), Proceedings of Sinn und Bedeutung 14, 151–167. Vienna.

Hackl, Martin. 2001. Comparative quantifiers and plural predication. In K. Megerdoomian & Leora Anne Bar-el (eds.), Proceedings of WCCFL XX, 234–247. Somerville, Massachusetts: Cascadilla Press.

Heim, Irene. 1985. Notes on comparatives and related matters. Unpublished manuscript, University of Texas, Austin.

Kennedy, Christopher. 2003. Ellipsis and syntactic representation. In Kerstin Schwabe & Susanne Winkler (eds.), The interfaces: Deriving and interpreting omitted structures (Linguistics Aktuell 61), 29–54. John Benjamins.

Krifka, Manfred. 1990. Four thousand ships passed through the lock: object-induced measure functions on events. Linguistics and Philosophy 13. 487–520.

Lasnik, Howard. 2001. When can you save a structure by destroying it? In Minjoo Kim & Uri Strauss (eds.), Proceedings of the North East Linguistic Society 31, 301–320. Georgetown

University: GLSA.

Lewis, Richard L. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research* 25(1). 93–115.

Lewis, Shevaun & Colin Phillips. 2015. Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research* 44(1). 27–46.

Merchant, Jason. 2001. *The syntax of silence: sluicing, islands, and the theory of ellipsis*. Oxford: Oxford University Press.

Montalbetti, Mario. 1984. *After binding*. Cambridge: Massachusetts Institute of Technology dissertation.

Nakanishi, Kimiko. 2007. Measurement in the nominal and verbal domains. *Linguistics and Philosophy* 30. 235–276.

O'Connor, Ellen, Roumyana Pancheva & Elsi Kaiser. 2012. Evidence for online repair of Escher sentences. In Emmanuel Chemla, Vincent Homer & G. Winterstein (eds.), *Proceedings of Sinn und Bedeutung 17*, 363–380. Paris: ENS.

Parker, Dan & Colin Phillips. 2015. Negative polarity illusions and the format of hierarchical encodings in memory. College Park: University of Maryland, m.s.

Potter, Mary C. & Linda Lombardi. 1990. Regeneration in the short-term recall of sentences. *Journal of Memory and Language* 29. 633–654.

Richards, Norvin W., III. 1997. *What moves where when in which language?* Cambridge, Massachusetts: Massachusetts Institute of Technology dissertation.

Ross, John Robert. 1969. Guess who? In Robert I. Binnick, Alice Davison, Georgia M. Green & Jerry L. Morgan (eds.), *Papers from the Annual Meeting of the Chicago Linguistic Society*, 252–286. Chicago, Illinois: Chicago Linguistic Society.

Sanford, A. & P. Sturt. 2002. Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Science* 6. 382–386.

Schein, Barry. forthcoming. Conjunction reduction redux. Manuscript, USC.

Thomas, Guillaume. 2010. Incremental *more*. In Nan Li & David Lutz (eds.), *Proceedings of Semantics and Linguistic Theory 20*, 233–250. Ithaca, NY: CLC publications, Cornell University.

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

Vasishth, Shravan, Sven Brüssow, Richard L. Lewis & Heiner Drenhaus. 2008. Processing polarity: how the ungrammatical intrudes on the grammatical. *Cognitive Science* 32(685-712).

Wason, P. & S.S. Reich. 1979. A verbal illusion. *Quarterly Journal of Experimental Psychology* 31. 591–597.

Wellwood, Alexis. 2015. On the semantics of comparison across categories. *Linguistics and Philosophy* 38(1). 67–101.

Wellwood, Alexis, Valentine Hacquard & Roumyana Pancheva. 2012. Measuring and comparing individuals and events. *Journal of Semantics* 29(2). 207–228.

Wellwood, Alexis, Roumyana Pancheva, Valentine Hacquard, Scott Fults & Colin Phillips. 2009. The role of event comparison in comparative illusions. CUNY 2009 poster.

Xiang, Ming, Brian Dillon & Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language* 108. 40–55.