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 reported to be 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 results of 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. Two formal acceptability studies show that speakers are most prone to the illusion when the matrix clause supports an event comparison reading. A verbatim recall task tests and finds evidence for such construals in speakers' recollections of the sentences. We suggest that this reflects speakers' entertaining an interpretation that is initially consistent with the sentence, but failing to notice when this interpretation becomes unavailable at the *than*-clause. In particular, semantic knowledge blinds people to an illicit operator-variable configuration in the syntax. Rather than illustrating processing in the absence of grammatical analysis, comparative illusions thus underscore the importance of syntactic and semantic rules in sentence processing.

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 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 and meaningfulness that characterizes CIs has so far received little systematic investigation. 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

Date: June 2, 2018.

to preliminary results (Fults & Phillips 2004, Wellwood et al. 2009, O'Connor et al. 2012; O'Connor 2015).[1]

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 is 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.

Deriving the interpretation of (2) involves mapping the individuals satisfying the matrix and embedded predicates to degrees representing their number, and establishing whether the first number is greater than the second. In the syntactic tradition going back at least to Bresnan (1973) (see also Chomsky 1977), degrees are introduced by the MANY component of *more* (i.e., *more* is underlyingly MANY and -ER). The degree predicates are derived in tandem with a *wh*-operator that binds a variable in the abstract syntax of the *than*-clause, as in (3). This operator is akin to *how many* in (4a). It needs to combine with a bare plural NP, just like *how many* does, (4b)-(4d).

(3) ... than WH-*d* ... *d*-MANY elephants have been to Russia

(4) 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?

Semantically, this binding relation corresponds to a λ-abstraction (see especially Heim & Kratzer 1998) over degrees, (5a). A parallel degree predicate is derived in the matrix clause by quantifier raising the morpheme *-er*, delivering an LF like that in (5b). Together, these two predicates act as the restrictor and scope arguments for the degree quantifier *-er* as in (6) (Heim 2000).[2] Thus, the LF in (5b) is interpreted as a greater-than comparison between the maximal degrees that satisfy the degree descriptions in the main and *than*-clauses (i.e., $max(Q)$ and $max(P)$ in (6), respectively).

(5) a. ... than λd ... d-many elephants have been to Russia
b. -er [λd ... d-many people have been to Russia
[(than) λd ... d-many elephants have been to Russia]

(6) $[\![\text{-er}]\!] = \lambda P_{dt}.\lambda Q_{dt}.max(Q) > max(P)$,
where $max(R) = \iota d[R(d)\ \&\ \forall d'[R(d') \rightarrow d' \leq d]]$

[1] Our early results, reported in Wellwood et al. 2009, inform the present manuscript and have shaped the subsequent literature. A report on these results can be downloaded from `https://github.com/alexiswellwood/compillu`.

[2] For further details concerning the LF syntax of comparative sentences, see Heim 1985, 2000; Bhatt & Pancheva 2004, among many others. There have been several alternative characterizations of the precise semantics of *-er*, in particular the tradition following Bartsch & Vennemann 1972 and Kennedy 1999. For our purposes these differences are not important.

Grammatically, then, there are two problems with (1). For a well-formed comparative sentence, a non-overt *wh*-operator needs to appear in the *than*-clause in a position parallel to that of *-er* in the main clause; this is not possible without a bare plural. Semantically, there is no plurality of individuals that can be compared for their number (see Hackl 2001, Nakanishi 2007, Wellwood et al. 2012; Wellwood 2015 for the semantic ban on singulars in comparatives). Ignoring the syntactic rules for a moment, the interpretation we would expect for (1) would have the schematic LF in (7b), which should make as little sense as those underlying (4b)-(4d). This stands in contrast to the interpretation of (2) in (7a).

(7) max(d-many people have been to Russia) >
  a. $max(\lambda d.d$-**many elephants** have been to Russia)
  b. $max(\lambda d.d$-**many I** have been to Russia) *

The claim that CIs are ungrammatical is not incompatible with an initial perception of acceptability, as acceptability and grammaticality have often been seen to diverge (cf. Lewis & Phillips 2015). Garden path sentences (8a) and sentences with multiple center-embedding (8b) are often perceived to be unacceptable, yet nonetheless grammatical (see Bever 1970 and Lewis 1996, respectively). Conversely, in some cases ungrammatical sentences are judged acceptable, as in cases of plural attraction (8c) and NPI illusions (8d) (see Bock & Miller 1991, Clifton et al. 1999,Vasishth et al. 2008, Wagers et al. 2009, Xiang et al. 2009, Parker & Phillips 2016).

(8) 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 involve 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 (9a) (Sanford & Sturt 2002). Similarly, (9b) is said to be literally equivalent to ‘All head injuries are trivial enough to ignore’; nevertheless, speakers routinely understand (9b) as equivalent to ‘Any head injury is too important to ignore’ (Wason & Reich 1979, O’Connor 2015). In these cases, comprehenders construct and linger on a certain misinterpretation that prevents them from recognizing the error.

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

CIs appear to present a different sort of case from all of these examples, however. Sentences like (1) strike speakers as well-formed, unlike (8a) and (8b). That perception can persist, unlike the easily detectable problems with (8c) and (8d). Furthermore, one never arrives at a specific, grammatically-licensed interpretation—there doesn’t seem to be a single sort of misinterpretation that speakers eventually converge on, unlike (9a) and (9b). Rather, informal reports by colleagues, friends, and audiences at professional meetings suggest that speakers tend to believe sentences like (1) are acceptable and have a coherent interpretation, even while they struggle to articulate that interpretation.

These considerations implicate online processing in the effect. When things go right, comprehenders can infer at *more people* that they are likely to require an operator-variable configuration of a certain sort in a dependent *than*-clause. Encountering *than elephants*, the variable can be posited in the determiner position of the bare plural, and the rest of the sentence can be parsed in the normal way. Encountering *than I* does not allow for the completion of the dependency, and so the parser must wait for a suitable nominal correspondent. If the sentence continued with an expression like *than I expected*, for example, the gap could be posited as part of the elided clause, but this isn't possible in (1).

It is thus striking that CIs seem to be as acceptable as informally reported. To understand the phenomenon better, we first need to understand the conditions under which it arises, and how far it generalizes. We outline four plausible sources for the effect (§2), and test the predictions of these accounts in two formal acceptability studies (§3). We then probe how speakers recast the sentences in production, using a sentence recall study (§4). Throughout, we find evidence only for an *event comparison hypothesis* (§2.4): speakers' semantic knowledge leads them to consider an event-counting reading licensed by the normal syntactic rules in fully grammatical comparatives, but fail to notice when that interpretation is no longer available.

These results suggest that people can be 'fooled' by attractive parses, motivated in semantic analysis, that differ minimally from the problematic syntactic representations they are asked to build. The illusory effect persists when it is possible to establish a plausible 'event-counting' reading; it does not appear to depend on superficial features of the sentences that should matter on template matching accounts (§2.1; Townsend and Bever 2001), on whether ellipsis has applied (§2.2; Fults and Phillips 2004), nor on a confusion between the comparative and additive senses of *more* (§2.3). This more limited distribution reflects, we suggest, that the satisfaction of certain semantic requirements can blind comprehenders to the fact that the syntax needed to support that semantics is illegitimate.

CIs present a disconnect between apparent well-formedness and shifting-sands interpretation. Their study thus informs broad questions about the architecture of sentence processing, and the degree to which grammatical theory informs sentence processing. If CIs are tightly linked to well-motivated grammatical mechanisms, then they should be exotic, not generalizing particularly far; we discuss some initial suggestions for their crosslinguistic distribution in §5. Informal reports suggest that the effect is robust; but as we will see, the mismatch between syntax and semantics means that the effect is not fully stable, or always accessible.

2. Plausible sources for comparative illusions

We consider four hypotheses, each of which predicts specific linguistic factors to affect the acceptability of CI-type sentences in contrast to fully grammatical control sentences. We test these hypotheses in a factorial design in §3. To preview those results, we find evidence only for the event comparison hypothesis discussed in §2.4.

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 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. More generally, the observation is that each of the two clauses alone is perfectly acceptable in some contexts, and this should be enough for CIs to be acceptable.

To see how such an account would work, consider the sentences in (10). (10a) involves a comparison between individuals: the number that have been to Russia and the number that the speaker would have thought have been to Russia. (10b) involves a comparison between events: the number involving people going to Russia and that involving the speaker going 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.

(10) 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 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 (11).

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

It isn't entirely clear what a syntactic template matching account predicts for the acceptability of CI-type sentences. Based on how it is presented in print (and in personal communication with one of the authors), we can interpret the account in one of 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 relevant lexical overlap between the two clause templates against which it is compared (the more constrained theory). With respect to (ii), the constraint on template-matching should amount to there being a shared lexical item (e.g. *more*) that is grammatically relevant for each template.

On the less constrained theory, all that should matter 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, which does not seem to be correct. For example, the blends in (12c) and (13c) seem immediately unacceptable, while their constituent clauses are fine in other contexts (the (a) and (b) examples). Put differently, if the account is as free as this version of it suggests, we would expect to see a lot of odd blends even outside of CIs, while in fact these appear to be relatively rare.

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

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

The more constrained version of the theory, as in (ii) above, makes lexical overlap a specific requirement for template matching. Consider the availability of a matrix clause like in sentence (14a), but the intuitive unavailability of a *than*-clause template like in sentence (14b). Such pairs could suggest that the requisite lexical overlap is not available, and so a blend like (14c) should be judged unacceptable. On this formulation, the more constrained theory predicts that speakers should find CI-type sentences with *fewer* to be significantly less acceptable than those with *more*.

(14) a. **Fewer people have been to Russia** than I would have thought.
  b. *People have been to Russia fewer **than I have**.
  c. →? Fewer people have been to Russia than I have.

We proceed assuming the version of the template-matching hypothesis that predicts people should not judge a sentence like (14c) to be as highly acceptable as its counterpart in (1). If our 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 abstract grammar in facilitating the illusion, rather than the rudimentary grammar used in first pass parsing. In particular, it links the phenomenon with other cases in which successful applications of ellipsis ameliorate grammatical problems elsewhere.

Investigating the possibility that ellipsis facilitates the CI effect, Fults and Phillips (2004) found significant degradation in acceptability for CI-type sentences without ellipsis. (Numbers indicate means and standard errors of ratings on a 1-5 scale.)

(15) a. More people have been to Russia than I have. 3.58|.16
  b. More people have been to Russia than I have **been to Russia**. 2.92|.19

Such an account is thus plausible, as there are many reported cases in which ellipsis 'blinds' comprehenders to other illicit rule applications. Both the formal syntax (Ross 1969, Lasnik 2001, Merchant 2001, Kennedy 2003) and experimental literatures (Frazier & Clifton 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*); Richards (1997) discusses similar effects with multiple applications of *wh*-movement.

Thus, it may be that successfully resolving ellipsis in the *than*-clause of a CI plays a role in its acceptability; we call this the repair-by-ellipsis hypothesis, (16).

(16) **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 their counterparts with no ellipsis. Yet, Fults & Phillips' result supporting this hypothesis may be confounded, since identical material is preferentially deleted in the *than*-clause of a comparative in English (Bresnan 1973). Thus, simple repetition of the matrix and *than*-clause predicates in sentences like (15b) could have independently reduced participants' ratings in their experiments. Nonetheless, the hypothesis predicts that sentences like (17) with a superficially different VP between the matrix and *than*-clauses should be judged less acceptable than (1).

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

If participants judge sentences like (17) to be as acceptable as (1), then an explanation for the CI effect in terms of repair-by-ellipsis is less plausible.

2.3. more *ambiguity*

Applying the normal interpretive rules to CIs ultimately fails. Yet, the initial perception of acceptability could be due to speakers temporarily constructing an alternative interpretation that is coherent, and this accounts for a heightened perception of meaningfulness. 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), colleagues, friends, and audience members often suggest that there is in fact a fully coherent interpretation that could be paraphrased using either of the sentences in (18).

(18) 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* (for detailed discussion of such ambiguities, see Greenberg 2010 and Thomas 2010; cf. Grant's 2013 investigation). To see the difference between these senses, consider (19). The additive interpretation in (19a) indicates a quantity in addition to (but not necessarily greater than) a previously-mentioned quantity. The comparative interpretation in (19b) 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.

(19) 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.

Apart from the cases with an explicit *just me*, the additive interpretation seems to be generally unavailable with *than*-phrases. That is, we can't read *Mary worked on the project more than Bill did* as true if Mary's total working time was less than Bill's. If this is correct, then 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 the *than*-clause. This hypothesis is stated as in (20).

(20) **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. 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 (21a) 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, in our judgment, unacceptable, (21b).

(21) 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, *people.* This makes two predictions. First, participants should judge sentences like (21a) to be less acceptable than sentences like (1). Additionally, since *fewer* fails to support the additive interpretation, sentences like (21b) should also be judged less acceptable than (1).[3]

2.4. Event comparison

The fourth and final account that we consider links the CI effect to the semantics of comparative constructions generally, rather than to a lexical ambiguity. It ties the effect to a regular process by which a subject nominal comparative can be interpreted as a comparison between numbers of events, and a general requirement that the comparanda be non-singular. Whenever these semantic requirements are satisfied, speakers could be 'seduced' into thinking that the syntactic requirements of a sentence like (1) have also been met.

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 (22).

(22) **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 sentences like (1) as (22). But there is another possibility. It is possible to interpret numerically-quantified noun phrases as expressing counts of individuals' participations in events (Krifka 1990, Barker 1999, Schein 2017). (23a), for example, can be true even if the total number of individuals is far fewer than 5 million, so long as there are at least 5 million ridings per week. The two readings coincide in a sentence like (23b), since it is only possible for a given person to satisfy that predicate exactly once.

(23) 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 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 (24). Here, the sentence in (25) can be judged true if individuals are counted, (25a), while it can be judged false if events are counted, (25b).

(24) 10 sailboats passed through the lock 10 times each (100 passings), and 5 barges passed through the lock 50 times each (250 passings).

(25) 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

[3] 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.

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: the covert gap site inside of a *than*-clause in a nominal comparative must have an appropriate nominal host (see §1). This hypothesis is stated in (26).

(26) **Event comparison hypothesis**
CIs reflect speakers' attempts to compare numbers 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 distinct from the individual-counting interpretation. Thus, the predicate must be 'repeatable', as opposed to 'once-only' or 'non-repeatable' (cf. Nakanishi 2007, Wellwood et al. 2012, Wellwood 2015). These are exactly the conditions under which a verbal comparative is felicitous: the meaning of (27a) is clear, as Mary may be involved in however many marathons as she likes, but (27b) is odd, since individuals tend not to graduate high school multiple times.

(27) 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 (28). The event comparison hypothesis thus predicts that a CI-type sentence with a non-repeatable predicate like that in (28) should be judged less acceptable than a sentence like (1).[4]

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

This proposal is essentially semantic in nature: detecting that a predicate is repeatable could lead participants to suppose that the semantic requirements of *more* can be fulfilled, leading to a heightened sense of acceptability. An alternative possibility is that detecting the repeatability property could push participants towards syntactic reanalysis, in which *more* is categorized as an adverbial. Importantly, this syntactic alternative could predict a difference between *more* and *fewer* (e.g., **People have been to Russia fewer than I have*). In what follows, we focus on the semantic version of the event comparison hypothesis; in §4, we pit the semantic and syntactic versions against each other.

3. Acceptability judgment studies

In two acceptability judgment studies with 88 unique participants, we investigated the robustness of the CI effect, and which properties are essential to it. The focus of the studies was on specific manipulations of CI-type sentences used to test different hypotheses about the cause of the CI effect. However, an important preliminary is to establish how robust the illusions are in a more carefully controlled setting. The overall picture is that responses to CI-type sentences are much more variable than other categories of sentence—sometimes receiving judgments of high acceptability, low acceptability, and everything in between.

[4] An anonymous reviewer questions whether sentences with the same predicate as in (1)—specifically, *to have gone to Russia*—in fact has an event-counting reading. We think a good diagnostic is whether the predicate sounds felicitous with a *N times* adverbial: for example, *I have been to Russia three times* is intuitively acceptable and meaningful, while *I graduated high school three times* is odd.

We can address the question of robustness by comparing the overall patterns of acceptability for our test sentences that were maximally similar to (1) and (2). 'Maximally similar', for these purposes, means subject nominal comparatives with repeatable predicates and *than*-clauses with non-bare plural subjects and VP ellipsis ('comparative illusions') versus the same but with bare plural subjects ('control comparatives'). We can also contrast these patterns with what we observed for our filler sentences, which were designed to elicit either a low or high rating while having a similar length, degree of syntactic complexity, and, around one-third of the time, similar semantic complexity (i.e. comparative-type meanings), (29)-(30).

(29) **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.

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

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 overall picture can be seen in Figures 1 and 2. The average acceptability ratings for CI-type sentences minimally different from (1) were much lower than for the controls, and the distribution of ratings was much more variable: the mean ratings were around 2 points lower for CIs than for controls, but almost 1 point higher for CIs than for 'bad' fillers (Figure 1). The averaged responses spanned almost the entire range for CIs, while they spanned only around 4 points for control sentences and 'bad' fillers. Participants' responses were highly consistent, except for CIs (Figure 2): the controls and 'good' fillers clearly tended toward the high end and 'bad' fillers to the low end, while the ratings for CIs were fairly evenly spread along the scale.

This initial survey of the data suggests an answer to our preliminary question: sentences maximally similar to (1) showed a high degree of variability in responses, in that they were judged completely acceptable nearly as often as they were judged completely unacceptable. (The same pattern was consistently observed in our initial experiments, not reported here.[5]) Thus, anecdotal reports of the robustness of the CI effect do not translate into stable patterns of high acceptability in a lab setting. Crucially, though, the variability that we observe in participants' responses is limited to CIs; as we discuss below, it is possible that this pattern reflects whether a given rating arises following a speaker's initial versus considered judgment.

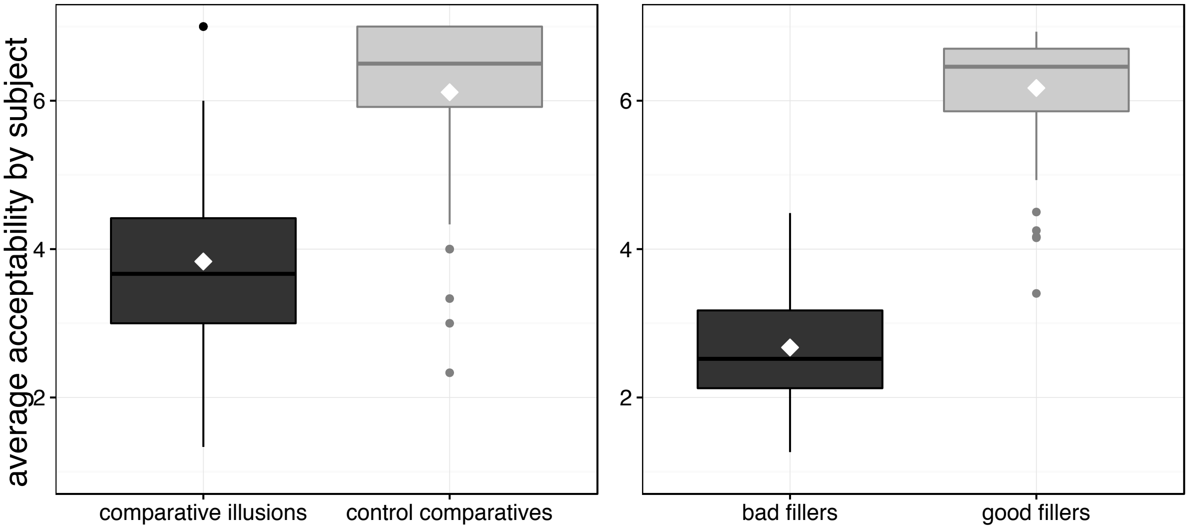
For now, though, we turn to our second question: what factors make participants more likely to assign a higher rating? As we will see, only one factor consistently had such an effect: CI-type sentences with repeatable predicates were consistently rated higher than were those with non-repeatable predicates,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

3.1. Experiment 1

Experiment 1 was an acceptability judgment task with responses recorded on a 7 point scale, where 1 was 'unacceptable' and 7 was 'acceptable'. Participants were given a couple of examples of 'acceptable' and 'unacceptable' sentences to get them started; the exact instructions and examples issued to participants for our acceptability experiments can be

[5] Discussion of those results can be downloaded from `https://github.com/alexiswellwood/compillu`.

Figure 1. Boxplots of mean participant ratings for classic CI-type sentences and controls (left), and fillers (right) in Experiment 1, on a 1-7 scale. 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; and filled circles indicate values outside of these ranges (i.e. outliers).



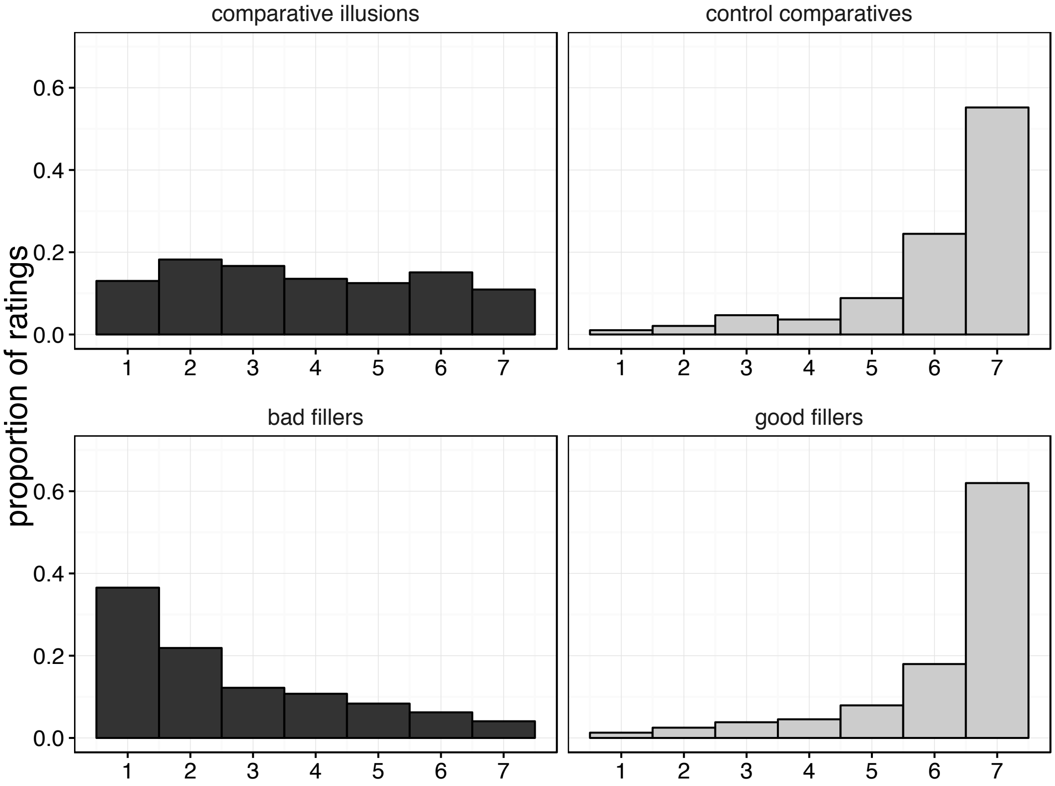
found in Appendix A. We recruited 64 participants on Amazon's Mechanical Turk, all native speakers of American English as determined by self-reporting, who received $6.65 for 40 minutes of participation.

Each hypothesis outlined in §2 predicts that specific factors should make speakers more susceptible to CIs. Specifically, they predict that different manipulations will impact CIs to the exclusion of fully grammatical controls. Therefore, our primary manipulation was a comparison of illusion and control conditions (the factor COMPARATIVE), with the 'illusion' conditions defined as those with non-bare plural subjects in their *than*-clauses, and the control conditions were defined as those with bare plural subjects in their *than*-clauses. This factor was crossed with a subset of further factors that specific hypotheses predict should selectively impact the acceptability of CI-type sentences.

The within-items factor QUANTIFIER manipulated the comparative quantifier in the main clause subject position (*more* vs. *fewer*; (31)). This manipulation was used to test two of the four hypotheses about the source of the CI effect. The template matching hypothesis relies on the fact that *more* can function as both a determiner and an adverbial. The additive *more* hypothesis relies on the ambiguity of *more* as having 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.[6]

[6] The examples illustrating the factors are simplified for presentation purposes; these are not actual experimental items. Two of our experimental items are given in Figure 3 and in Appendix B. The complete set of experimental sentences can be viewed at `https://github.com/alexiswellwood/compillu`.

Figure 2. Histograms of ratings for classic CI-type sentences and controls (top row), and bad and good fillers (bottom row), in Experiment 1. Each plot represents the proportion of total responses observed for each scalar value.



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

The within-items factor ELLIPSIS manipulated whether VP ellipsis had applied in the *than*-clause (ellipsis vs. no ellipsis; (32)). 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, VPs in the 'no ellipsis' conditions differed just enough from the matrix clause VP to potentially circumvent this preference.

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

The between-items factor REPEATABILITY manipulated whether the VP in the comparative was repeatable for a given agent (repeatable vs. non-repeatable; (33)). 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 non-repeatable predicates like *graduate high school* should fail to elicit the CI effect.

Figure 3. Schemata for repeatable and non-repeatable items in Experiment 1, representing 16 unique conditions. Factors represented are repeatability (between items—repeatable, non-repeatable), 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 contained 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 non-repeatable item

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

(33) repeatability
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', (34)), and was counterbalanced within the illusion conditions. This manipulation tested the additive *more* hypothesis, which requires the possibility of an inclusion relation in order to license a 'not just me' interpretation. The example in (34) involves matching/mismatching gender; variants on this included (mis)matching nationality, profession, age, etc. The additive *more* hypothesis predicts that 'inclusion not possible' trials should fail to elicit the CI effect.

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

All factors apart from subject inclusion were fully crossed, for a total of 16 unique conditions; subject inclusion was counterbalanced within the illusion conditions. Two sample items of the 'inclusion not possible' type used in Experiment 1 are given in Figure 3 (see Appendix B for a tabular version of this diagram).[7] The top diagram represents an item with a repeatable VP, and the bottom an item with a non-repeatable VP. Any path through the figure from left to right corresponds to one condition, and the 8 possible paths through each diagram together correspond to the 16 experimental conditions.

Table 1 summarizes the predictions of the four accounts relative to these factors, with '>' representing the directionality of the prediction: 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 thus predicted

[7] A typical 'inclusion possible' trial would have an expression like *Canadians* in the matrix subject position, and a name in the *than*-clause subject position.

to be interactions, affecting the illusion conditions but not the control conditions. As noted above, SUBJECT INCLUSION was tested only within the illusion conditions.

TABLE 1. Predicted interactions by hypothesis and factor. Each hypothesis (apart from additive *more*) predicts an interaction between the factor COMPARATIVE and the factor listed at the top of each column. '>' indicates the predicted direction of the interaction: the illusion conditions should be more acceptable on the left-hand factor level than on the right-hand level, as compared to the control conditions. The additive *more* hypothesis makes a prediction only within the illusion conditions: those where subject inclusion is possible should be judged more acceptable than those where it is not possible. A '-' indicates that the hypothesis makes no predictions for that factor.

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

Our 48 experimental items were distributed across 8 lists, and combined with 144 filler sentences for a 1:3 ratio of experimental to filler sentences, creating 8 questionnaires.[8] 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 (29) and (30) 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 (29a) and (30a) for examples), which were included to help mask the experimental items.

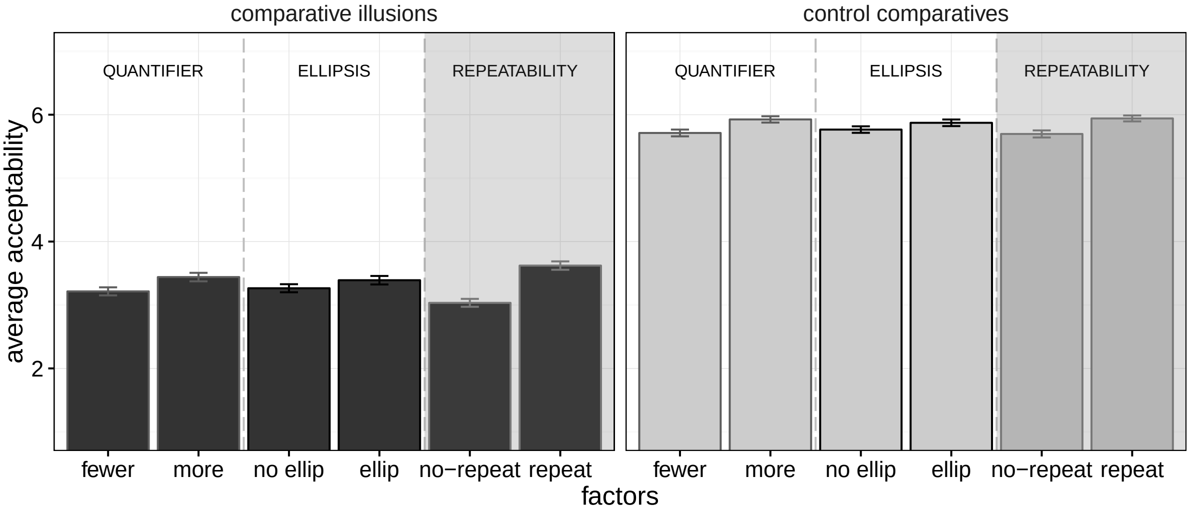
In this experiment, CI-type sentences featured singular proper names, 3rd person pronouns, and definite descriptions as *than*-clause subjects. We made this choice because 3rd person expressions provided a minimal contrast with the matrix bare plural NP, and because it would not be possible to use first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

Results Our participants rated the control conditions more highly than the illusion conditions, as can be seen in Figure 4. The means for the illusion conditions should be interpreted with caution here, since they reflect a mixture of high and low ratings, as well as ratings for many sentences that our experimental manipulations predicted should lead to lower scores. Overall, *more* was better than *fewer* (the factor QUANTIFIER), ellipsis was better than no ellipsis (the factor ELLIPSIS), and repeatable predicates were better than non-repeatable (the factor REPEATABILITY). CI-type sentences where inclusion was possible did not differ from those where inclusion was not possible (discussed below). Only (non-)repeatability impacted the CI-type sentences more than the control sentences, as can be clearly seen in Figure 4. These results provide support only for the event comparison hypothesis.

For the statistical analyses here and below, unless otherwise noted, we report the results of linear mixed effects regressions (LMERs) with maximal random effects structure, including

[8] The experimental items can be viewed at `https://github.com/alexiswellwood/compillu`.

Figure 4. Barplots of mean participant ratings by factor in Experiment 1, on a 1-7 scale. Only the factor repeatability showed the predicted interaction. Error bars represent standard error.



random intercepts and slopes by participant and item (Barr et al. 2013), with all factors entered into the model at the same time. Our χ^2 and p values for main and interaction effects were assessed via likelihood ratio tests of the model m that includes the relevant fixed effect, and a model variant m' that does not include that effect. All analyses were conducted using R's lme4 package (Bates et al. 2014).

Participants judged the illusion conditions to be less acceptable than the control conditions overall, as reflected in a much lower mean rating for CI-type sentences (illusion 3.33, control 5.82), $\beta = 2.49, \mathrm{SE} = .18, \chi^2(1) = 91.7, p < .001$. It is worth repeating, though, that the overall mean for the illusion conditions masks a substantial amount of variability.

Participants judged comparatives with *fewer* less acceptable than comparatives with *more* overall (*fewer* 4.46, *more* 4.68), $\beta = .21, \mathrm{SE} = .05, \chi^2(1) = 13.7, p < .001$. We did not necessarily predict this result, but it is plausibly related to the fact that *fewer* has a negative component to its meaning (e.g., Deschamps et al. 2015). Importantly, replacement by *fewer* failed to disproportionately impact the illusion conditions (illusions: *more* 3.44, *fewer* 3.21, controls: *more* 5.93, *fewer* 5.71), $\chi^2(1) < .1, p = .86$. This lack of interaction fails to support the syntactic template matching hypothesis or the additive *more* hypothesis, both of which predict the illusory effect to depend (at least in part) on the quantifier.

Participants judged comparatives without ellipsis somewhat less acceptable than those with ellipsis overall (no ellipsis 4.51, ellipsis 4.63), $\beta = -.12, \mathrm{SE} = .06, \chi^2(1) = 3.34, p = .07$. This effect could have been due to the additional length of the sentences. Significantly, the effect was constant within both the illusion and control conditions (illusions: ellipsis 3.39, no ellipsis 3.26; controls: ellipsis 5.87, no ellipsis 5.77), $\chi^2(1) < .1, p = .9$. These results fail to support the repair-by-ellipsis hypothesis, which predicted that CI-type sentences, but not controls, would receive higher ratings with *than*-clause ellipsis.

Participants judged comparatives with non-repeatable predicates less acceptable than those with repeatable predicates overall (repeatable 4.78, non-repeatable 4.37), $\beta = .41, \mathrm{SE} =$

$.07, \chi^2(1) = 25, p < .001$. Crucially, though, this effect was magnified in the illusion conditions as compared to controls (illusions: non-repeatable 3.03, repeatable 3.62; controls: non-repeatable 5.7, repeatable 5.94), $\beta = -.34, \mathrm{SE} = .16, \chi^2(1) = 4.3, p = .04$. This interaction supports the event comparison hypothesis, which predicted the acceptability of CI-type sentences to depend on the availability of a repeated-events interpretation.

Finally, the possibility of a 'just me' interpretation failed to impact participants' judgments of acceptability. Ratings in the illusion conditions that failed to support an additive interpretation were in fact higher than ratings for the illusion conditions that did support an additive interpretation (inclusion not possible 3.44, inclusion possible 3.23), though this effect was not borne out statistically, $\chi^2(1) = 1.9, p = .17$. Nonetheless, this pattern of results fails to support the additive *more* hypothesis, which predicted a substantial difference in the opposite direction.

Discussion Our primary interest in Experiment 1 was testing what could be responsible for the CI-effect: the perception that sentences like (1) are acceptable and meaningful, but ultimately have no coherent sense. We tested four factors that were predicted to selectively affect the acceptability of CIs than of fully grammatical controls, in light of the four hypotheses presented in §2. Each hypothesis predicted that specific factors should impact the acceptability of CI-type sentences over and above any effects on control sentences.

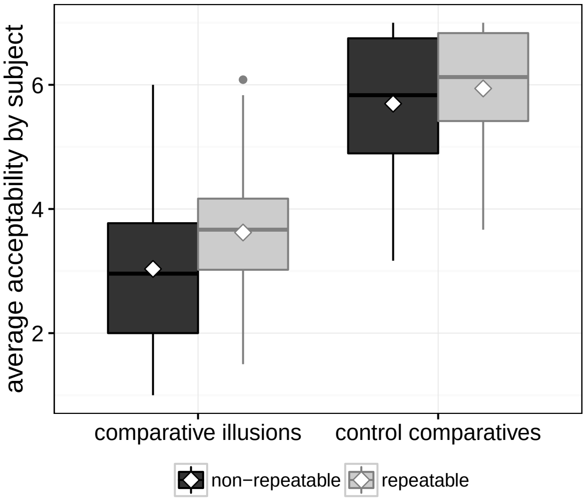
Our results provide support only for the event comparison hypothesis. This hypothesis predicted an interaction between the factors COMPARATIVE and REPEATABILITY. If the CI-effect requires that the predicate be 'repeatable' for a given agent, then CI-type sentences with such predicates should be judged more acceptable than those with non-repeatable predicates. This was the only reliable interaction effect that we observed in Experiment 1 (Figure 4). In fact, we tested the same manipulations as in the present experiment in three preliminary studies, and found consistent support only for this effect (these results can be viewed in a report downloadable from `https://github.com/alexiswellwood/compillu`).

The syntactic template matching hypothesis predicted an interaction between the factors COMPARATIVE 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*, likely due to a difference in negativity.

The repair-by-ellipsis hypothesis predicted an interaction between the factors COMPARATIVE 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, the

FIGURE 5. Boxplots of mean participant ratings by repeatability in Experiment 1, on a 1-7 scale. 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; filled circles represent outlying values.



trend we found was in the opposite direction. This hypothesis also predicted an interaction between the factors COMPARATIVE and QUANTIFIER, since *fewer* lacks the requisite additive semantics; yet, this effect was not observed.

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 nearly as likely to receive a high rating as a low rating, as can be seen in Figure 5. The range of participant responses was particularly wide for non-repeatable illusions. One possibility is that this reflects a probabilistic process: if participants fail to notice the anomaly on a given trial, they will rate it higher; if they notice it, they will rate it lower. Importantly, however, CI-type sentences were most acceptable when their predicate was repeatable for a given agent, as reflected both in the higher mean and the narrower range of ratings.

Nevertheless, it is possible that one feature of Experiment 1 could have artificially decreased the ratings for CI-type sentences. First, the *than*-clause subjects in our illusion conditions did not feature first person pronouns, differently from the classic illusion in (1). We included this feature of the design, in part, in order to manipulate the factor SUBJECT INCLUSION. Yet, third person pronouns and definite descriptions require discourse antecedents in normal usage, which were inaccessible in our experiment. We manipulate the features of the *than*-clause subject and test their effects on acceptability in Experiment 2.

By manipulating those features, we can again draw on the semantics to suggest another clear prediction. It is quite possible that the repeatability of the predicate is not the only

Figure 6. Schema for items in Experiment 2, representing 12 unique conditions. Factors represented are repeatability (repeatable, non-repeatable), and subject type. The bare plural *boys* marks the control condition.

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

route to a plurality of events, but a plural subject could do the same. Even with a non-repeatable predicate, a plural subject can indicate a plurality of events: in *the girls graduated high school*, there is a single graduation event for each girl. Thus, a plural subject in the *than*-clause could, on its own, heighten the overall acceptability of a CI under the event comparison hypothesis. Another possibility, of course, is that plural subjects lend themselves more easily to a misparse of the sentence as fully grammatical (i.e., if one dropped the *the* from *the girls* in *More boys ate pizza than the girls did*). We discuss the latter possibility more when we turn to production, in §4.

3.2. Experiment 2

We tested how features of the subject NP in the *than*-clause impact the acceptability of CI-type sentences. The event comparison hypothesis explains the effect of the factor repeatability in terms of speakers entertaining CI-type sentences as a comparison between pluralities of events. A plural *than*-clause subject may independently lend plausibility to this interpretation. Experiment 2 thus tested the prediction that the number of the subject directly impacts the acceptability of CI-type sentences, as opposed to other nominal features.

This study had a 12 condition, 2 x 6 design manipulating the factors repeatability and subject type. The factor repeatability 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). We did not manipulate these dimensions factorially due to conditions that were either impossible or, intuitively, better omitted.[9] The bare plural subject condition was included as a fully acceptable control. A guide to the conditions is given in Figure 6 (see Appendix C for a tabular version of this guide).

In light of Experiment 1, we expected effects of the factors subject type, such that the illusion conditions (i.e. those with non-bare plural subjects) would be rated lower than the control conditions, and repeatability, such that the non-repeatable conditions would be rated lower than the repeatable conditions. We also expected an interaction between the factors subject type and repeatability: the effect of (non-)repeatability should be greater in the illusion conditions than in the control conditions, in line with the event comparison hypothesis.

[9] The combination of 1st person and definite description (singular or plural) is not possible, and use of 3rd person plural pronouns sounds intuitively contradictory (e.g., *More girls ate pizza than they did*), which could decrease acceptability independently of the CI phenomenon, which is our main concern.

We planned comparisons between subsets of the illusion conditions to test which properties of the *than*-clause subject impacted 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 affect the acceptability of the illusion conditions: those 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, since both a repeatable VP or a plural *than*-clause subject can indicate a plurality of events. Such a pattern would reflect 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 non-repeatable predicate and a singular subject.

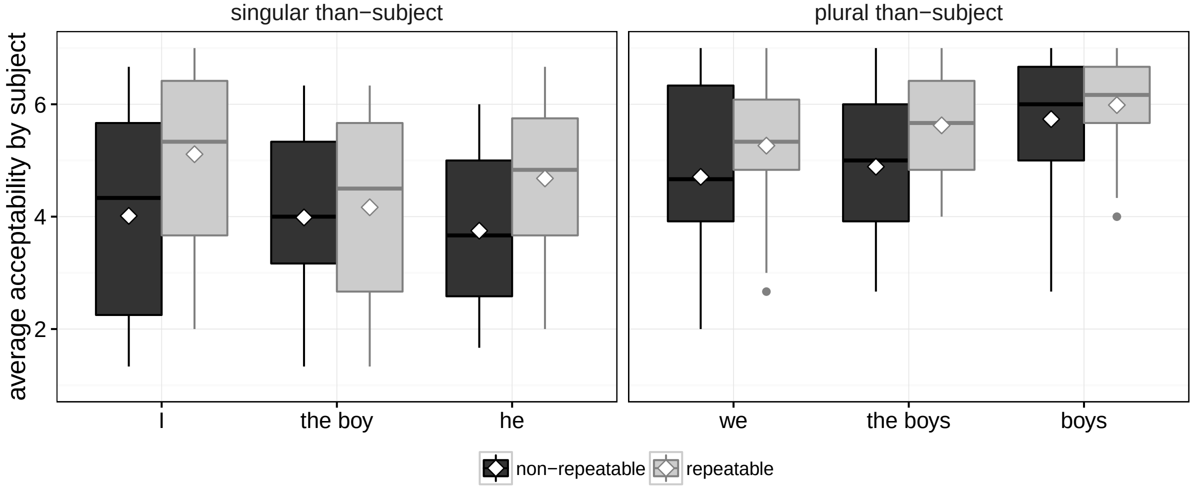
We distributed 36 sets of items across 6 lists in a Latin Square fashion. These were combined with 108 filler sentences to create 6 questionnaires. Fillers were designed to be evenly split between sentences that should elicit a low rating and those that should elicit a high rating. Acceptability judgments were recorded on a 7 point scale where 1 is 'unacceptable' and 7 is 'acceptable'; the instructions were the same as those for Experiment 1, and are provided in Appendix A. 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 Experiment 2 replicated the major effects of Experiment 1 (Figure 7). Overall, the control conditions were rated more highly than the illusion conditions, as can be seen by comparing the last column of Figure 7 to all of the others. Moreover, for all *than*-clause subject types, the comparatives with repeatable predicates were rated more highly than those with non-repeatable predicates (light versus dark grey boxes). The CI-type sentences (i.e., those with non-bare plural subjects) with plural subjects were rated more highly than those with singular subjects. These results confirm and extend the findings of Experiment 1, supporting the event comparison hypothesis.

First, with respect to the major effects that replicate those of the previous experiment, that participants judged comparatives without bare plural subjects less acceptable than comparatives with bare plural subjects (illusions 4.62, control 5.86) was revealed in a strong main effect of SUBJECT TYPE, $\beta = 1.24, \mathrm{SE} = .19, \chi^2(1) = 24.29, p < .0001$. They judged comparatives with non-repeatable predicates less acceptable than those with repeatable predicates as well (non-repeatable 4.51, repeatable 5.14), $\beta = .48, \mathrm{SE} = .14, \chi^2(1) = 9.58, p < .01$.

Furthermore, the manipulation REPEATABILITY had a disproportionate impact on the illusion conditions as compared to the control conditions: participants rated the non-repeatable illusion conditions substantially lower than the repeatable illusion conditions (non-repeatable 4.27, repeatable 4.97), an effect that was weaker for the control conditions (repeatable 5.99, non-repeatable 5.74), $\beta = -.45, \mathrm{SE} = .25, \chi^2(1) = 3.15, p = .08$. Thus Experiment 2 confirms the results of Experiment 1.

Figure 7. Boxplots of mean participant ratings by subject plurality and repeatability in Experiment 2, on a 1-7 scale. All columns except that labeled with *boys* represent ratings for CI-type sentences. 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; filled circles represent outlying values.



Turning to our comparisons between subsets of the illusion conditions, we found that participants rated the illusion conditions with 1st person singular pronouns as more acceptable than those of 3rd person singular pronouns (Person comparison: *I* 4.56, *he* 4.22), suggesting that Person affects how acceptable CI-sentences are perceived to be, $\beta = -.35, \mathrm{SE} = .18, \chi^2(1) = 3.41, p = .06$. As noted above, this could be due to an independent effect of the lack of discourse antecedents for third person pronouns. This possibility could be tested in a future study that manipulates acceptability as a function of antecedent accessibility.

The fact that our participants did not differentiate between singular definite descriptions and singular 3rd person pronouns, which share the requirement for a discourse antecedent, suggests the same conclusion (Sort comparison: *the boy* 4.08, *he* 4.22), $\chi^2(1) = .52, p = .47$. The lack of a difference here also suggests that Sort (pronominal or definite) itself does not significantly impact the acceptability of CI-type sentences.

With respect to Number, the statistical analysis suggests that participants did not differentiate the *I* and *we* conditions (*I* 4.56, *we* 4.99), $\chi^2(1) = 2.59, p = .11$, contrary to our expectation. However, participants did differentiate the *the boy* and *the boys* conditions (*the boy* 4.08, *the boys* 5.26), $\beta = 1.17, \mathrm{SE} = .26, \chi^2(1) = 14.83, p < .001$, suggesting that Number (singular or plural) can impact the acceptability of CIs, in the direction predicted by the event comparison hypothesis.

Probing these results further, we conducted a linear regression just within the illusion conditions and found a main effect of REPEATABILITY (repeatable 4.97, non-repeatable 4.27), $\beta = .69, \mathrm{SE} = .14, \chi^2(1) = 17.67, p < .001$, and of Number (plural 5.12, singular 4.28),

$\beta = .84, \text{SE} = .2, \chi^2(1) = 13.1, p < .001$. Furthermore, these effects were additive: there was no interaction between Number and REPEATABILITY, $\chi^2(1) = .18, p = .67$.

Discussion Experiment 2 investigated which features of the subject NP in the *than*-clause would impact the acceptability of CI-type sentences. Plurality of the subject NP significantly affected the ratings: CI-type sentences with plural subjects were rated more highly than were those with non-plural subjects. Furthermore, this experiment also confirmed the major effect in Experiment 1: the repeatable illusion conditions were consistently rated more highly than were the non-repeatable illusion conditions.

The most highly-rated illusion conditions combined plural subjects and repeatable predicates (mean: 5.44); next highest were the conditions with plural subjects and non-repeatable 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 and non-repeatable predicates (3.92). In fact, CI-type sentences with repeatable predicates and plural subjects reached nearly the level of acceptability of controls.

These results 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 was (and thus, the more suggestive of an event-counting reading), the more likely participants were to judge the sentence as acceptable. Importantly, this was 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 these results is: why did even the most obstinately 'singular' sentences (i.e. those with singular *than*-clause subjects and non-repeatable VPs—those least likely to suggest an event comparison interpretation) still receive a fairly high average rating (i.e., 3.92/7)? If a higher acceptability rating for a CI-type sentence depends on its supporting an event-counting reading, then whenever these sentences fail to provide such support, we might expect to observe much lower acceptability.

We think the answer to this question requires considering how we manipulated repeatability in our items. That is, we know of events like high school graduations that they are once-only per individual, but this requirement is not enforced grammatically in English. Consider (35), which transparently bears the unlikely interpretation (i.e., unlikely given what we know about the distribution of such events in the population). Hence, it is possible that speakers could allow for a predicate like *graduate high school* to support an event-counting reading, which sometimes lead them to consider even our most 'singular' items as reasonably acceptable.

(35) Mary graduated high school three times.

3.3. 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). Experiment 2 showed that another dimension in the subject noun phrase of the *than*-clause—plurality—similarly positively impacted their acceptability. These effects are predicted by the event comparison hypothesis, on the assumption that plural subjects license plural event reading of their associated verb phrases; 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, 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 (i.e., 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 for nominal comparatives. Nonetheless, our acceptability judgment data suggest that an interpretation in terms of a comparison of events is, at least temporarily, entertained.

We observed high variability in the rating scores that our experimental participants assigned to CI-type sentences, but not to control sentences. We speculated that this could reflect a probabilistic process by which participants are sometimes 'fooled' into thinking that the CI-type sentence is acceptable, just in case they an event-counting reading is maintained. Such readings are possible only when the verb phrase of the comparative is repeatable, or when its *than*-clause subject is plural. If participants failed to notice that the reading isn't syntactically licensed, they might assign it a lower rating.

While we find these data compelling, acceptability judgment studies are a fairly indirect method of assessing interpretation. For instance, it could be that repeatable predicates or plural *than*-clause subjects just make CI-type sentences 'sound' better, without playing any important role in how participants are interpreting them. Perhaps our participants aren't doing any interesting interpretation over such strings at all. Such tasks cannot definitively tell us whether processing CIs importantly involves use of the interpretive cues that we have found to impact the judgments.

If people 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

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 tell us not only what participants do when they are asked to produce anomalous CI-type sentences, but it could also provide clues as to how the sentences are interpreted. We built upon a paradigm developed by Potter & Lombardi (1990), asking: how good are participants at recalling CI-type sentences? And, to the extent that they are reasonably successful, is there evidence for the event-counting reading in the forms that they recall?

Producing a sentence involves (at least) mapping a 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 the type of recall: short-term recall is fairly high-fidelity with respect to the form of the sentence, while long-term recall often returns the 'gist', or a suitable paraphrase of the sentence's meaning. This contrast has usually been taken as evidence for two distinct production processes: short-term, verbatim recall, depends on a stored surface representation of the form; while long-term recall depends on the normal processes involved in language production.

Potter and Lombardi (1990) hypothesized, in contrast, that a single set of mechanisms is used for language perception and production: the pathway 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 recall, other words might be more active than those in the target.

Potter and Lombardi 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 sometimes containing a 'lure' (a near-synonym to a word in the target sentence), then judge whether a subsequent word had appeared in the list. Immediately afterward, participants recalled the initial sentence aloud; they 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 for recall, since there would then be no explanation for how a new word came to be incorporated into such a representation. (They found the same effect even when the list-probe task occurred prior to the presentation of the target sentence, suggesting that the effect wasn't merely due to a 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 when it was possible to maintain an event-counting interpretation. Such a meaning is consistent with the 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 more difficult.

Our extension of Potter and Lombardi's methodology represents, to our knowledge, the first time that a sentence recall task has been used to probe the production of syntactically 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 doesn't literally have one. In light of this novel application, this initial study will be more exploratory than those we have so far presented.

4.1. Experiment 3

We investigated whether event comparison is relevant to how speakers interpret CI-type sentences, by investigating how they are produced in a verbatim sentence recall task. Building on the results we reported in §3, we hypothesized two ways such a reading would be supported: by a repeatable VP in the matrix clause, or by a plural *than*-clause subject.

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 possible 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 non-repeatable VPs, and singular subject NPs in the *than*-clause) sometimes received higher ratings. Since the grammatical properties of non-repeatable 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 COMPARATIVE), more errors on the non-repeatable conditions than on the repeatable conditions (a main effect of REPEATABILITY), and overall the most errors on the non-repeatable illusion conditions. Furthermore, we expect participants to pluralize the *than*-clause subject more than, say, they would singularize the bare plural in the corresponding control sentence.

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 §2.4. If this alternative is correct, then we should find that participants displace the comparative quantifier during recall of the illusion targets more than in the control targets.

4.1.1. Design

As in the acceptability experiments, we manipulated the factors COMPARATIVE (illusion, control) and REPEATABILITY (repeatable, non-repeatable), both within items. 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 non-repeatable 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*, (36).[10]

(36) ASPECT contrast

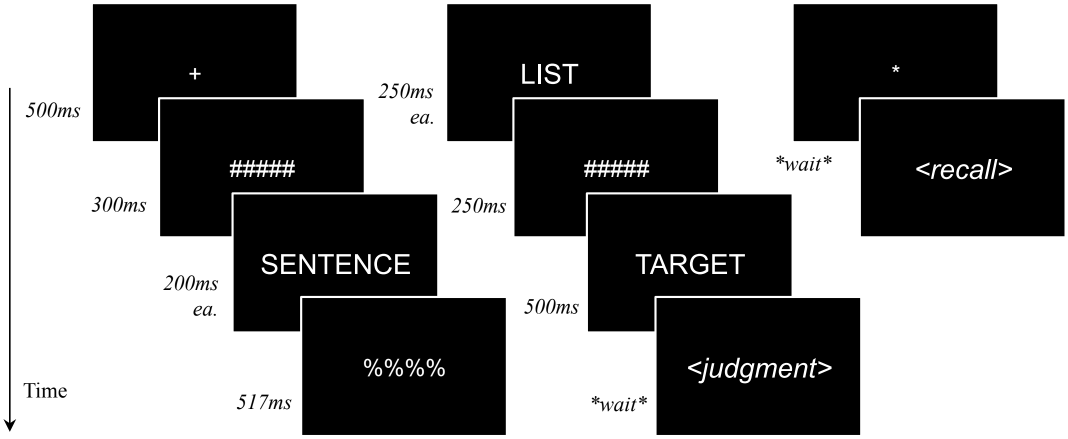
a. Mary {**started, finished**} reading the book. [non-repeatable VP]

b. Mary {**continued, was**} reading the book. [repeatable VP]

OBJECT items were classified as non-repeatable if they had an ordinal modifier, and repeatable otherwise, (37).

[10] These are simplified examples intended to illustrate the relevant contrasts. A pair of actual items from the experiment is given in Figure 9 below, and the full set may be viewed at `https://github.com/alexiswellwood/compillu`.

Figure 8. Schematic presentation of procedure from Experiment 3.



(37) Object contrast

a. Mary ate her {**first, last**} cupcake. [non-repeatable VP]

b. Mary ate a {**tasty, strawberry**} cupcake. [repeatable VP]

4.1.2. Procedure

Our procedure follows that laid out in Potter & Lombardi (1990), 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. There were no blank screens between the words. 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 to verbally recall the sentence in the Recall phase. The experiment was preceded by 6 practice trials to insure familiarity with the procedure. The complete instructions issued to participants can be found in Appendix D.

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. (The expanded tabular form appears in Appendix E.) In all of our items, *more* was preceded by an unrelated adverbial phrase; we included this aspect of the design in order 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 supports a better environment for testing the syntactic version of the event comparison hypothesis. All of our illusion conditions had singular *than*-clause subjects.

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

Aspect item

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

Object item

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

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.[11] Fillers contained no ungrammatical or anomalous sentences, 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, and the experiment was implemented in DMDX (Forster & Forster 2003).

4.1.4. Error coding

The recall data were first transcribed from audio format to text format. Next, they were coded for overall failure of recall, as well as two broad categories of errors: movement and non-movement. To illustrate these types of errors, 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. The data were transcribed and coded by the first author. We discuss the types of errors that occurred but which were not coded at the end of this section.

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.*

[11] The complete set of experimental sentences can be viewed at `https://github.com/alexiswellwood/compillu`.

Movement errors A response was classified as a movement error if the nominal determiner *more* was recalled in an adverbial or direct object position, (38). 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 that *more* would be moved to an adverbial position at a higher rate in the repeatable illusion conditions than in the non-repeatable 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 non-repeatable illusion conditions, since in this case there is more motivation to correct the representation. (Note that, in some cases, a single response included *more* displaced to both positions; these errors were coded as a single movement error.)

(38) **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. [direct object 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, (39). Within the control conditions, this involves rendering the plural NP subject of the *than*-clause as singular, (40). 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.)

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

(40) **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 non-repeatable) than the target sentence, and was only coded for the Aspect items. Within the non-repeatable 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, (41). Within the repeatable conditions, it modifies the VP so that it is non-repeatable: this involved recalling a VP with a continuative or copular verb with an initiative or terminative verb, (42). The semantic event comparison hypothesis predicts more NP number errors on non-repeatable trials than on repeatable trials.

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

(42) **VP number error** [repeatable → non-repeatable]
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 non-repeatable conditions, deletion of this adjective critically renders the VP potentially repeatable, (43). For the repeatable conditions, deletion of the adjective has no effect on the repeatability of the predicate, (44). The semantic event comparison hypothesis predicts more modifier deletion errors on non-repeatable trials than on repeatable trials.

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

(44) **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 did*).

4.1.5. Error coding examples

As an example of how the errors were coded, consider one of our target CI-type ASPECT items in (45). This item instantiates a CI-type sentence because the *than*-clause subject is a singular definite description, and it is labeled a non-repeatable item because one can only begin talking (in the relevant context) once.

(45) At the party more seniors began talking with the professors than the junior did.

In (46), three examples of responses to the target in (45) are given. Each instantiates distinct but overlapping error patterns. Each was coded as involving a VP number error because the *began*-complex was missing (error direction: singular → plural). Additionally, each was coded as involving an NP number error, because the *than*-clause subject was pluralized (direction: singular → plural). Additionally, since in (a) *more* appears in an adverbial position, and in (b) it appears in the direct object position, these were counted as Movement errors.

(46) a. At the party the seniors talked to the professors more than the juniors did.
b. The seniors talked about more professors than the juniors did.

c. At the party more juniors were talking at the professors than the seniors.

Thus, for any given response, participants may have incurred multiple errors, each coded for and localized in different parts of the response.

4.1.6. 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.

4.1.7. Results

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 non-repeatable illusion conditions were more difficult to recall than the repeatable illusion conditions, however.

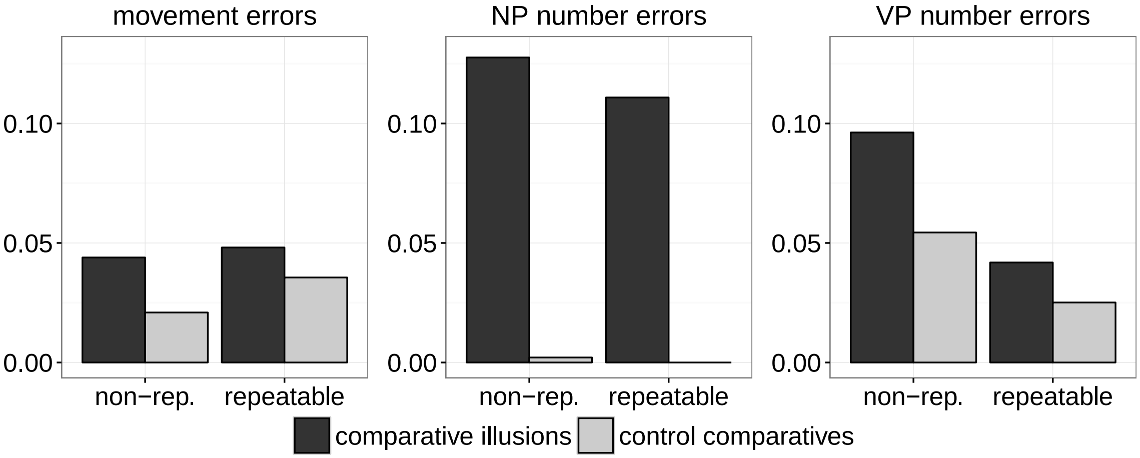
On successful recall trials, we found that participants made repeatability and number changes substantially more on the illusion conditions than on the control conditions. Participants were also more likely to change the repeatability of the predicate in the non-repeatable conditions than in the repeatable 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 10).

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

With respect to the statistical analyses, 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 in this section are the result of two types of analysis, either logistic or linear mixed effects regressions, with maximal random effects terms where possible (Barr

Figure 10. 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).



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, and indicate explicitly when we present the results of a linear analysis. As above, χ^2 and p values are assessed via model comparisons.

First, we examined the rate of global failure in recall. We found that illusion targets were more difficult to recall than control targets (see Figure 11), $\beta = .68, \text{SE} = .31, \chi^2(1) = 5.12, p = 0.024$. By this measure it did not appear that recalling non-repeatable targets was more difficult than recalling repeatable targets, $\chi^2(1) = .7, p = .4$, and there was no interaction between COMPARATIVE and REPEATABILITY, $\chi^2(1) = 2.18, p = .14$.

Next, we examined error rates in the distractor task (Figure 11). We did not find that participants made greater errors in this task for the illusion conditions as opposed to the control conditions. Excluding those trials on which participants failed at the point of recall, we found no difference in distractor error by the factor COMPARATIVE, $\chi^2(1) = .32, p = .57$, or by REPEATABILITY, $\chi^2(1) = .15, p = .7$; nor was there any interaction between these factors, $\chi^2(1) = .13, p = .72$.[12]

Turning to the counts of errors within successful recall trials, we found more errors for illusion targets than for control targets, manifesting as a main effect of the factor COMPARATIVE: a total of 260 errors were identified on 232 illusion recall targets, but a total of 113 errors on 246 control recall targets (Figure 12), linear: $\beta = .69, \text{SE} = .10, \chi^2(1) = 25.32, p < .001$. On this measure, we found no effect of the factor REPEATABILITY, linear: $\chi^2(1) = .38, p = .5$, and no interaction between the factors COMPARATIVE and REPEATABILITY, linear: $\chi^2(1) = .8, p = .4$.

Among the total errors observed, we found that participants made marginally more movement errors for illusion targets than for control targets (illusion 44, control 27; Figure 13,

[12] 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 recall failure and distractor task errors in Experiment 3.

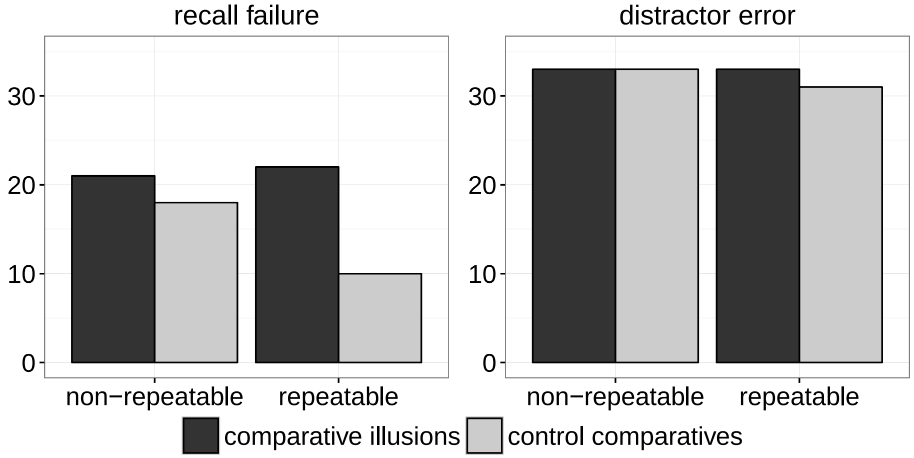
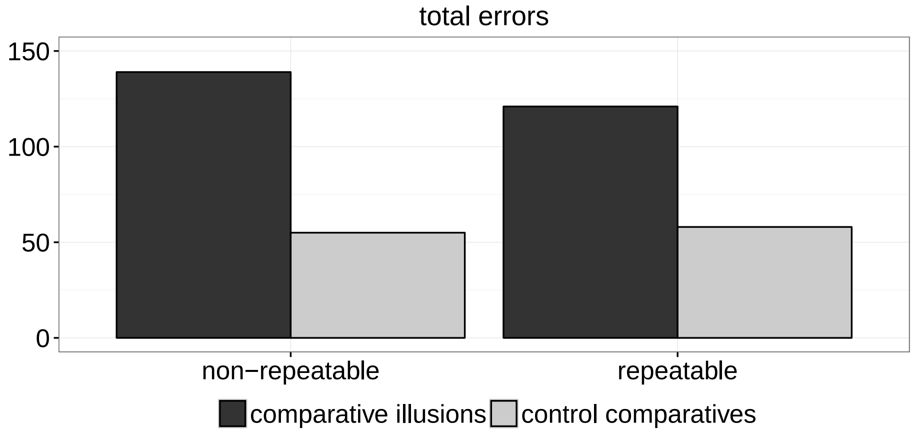


Figure 12. Counts of errors logged on successful recall trials by condition in Experiment 3.



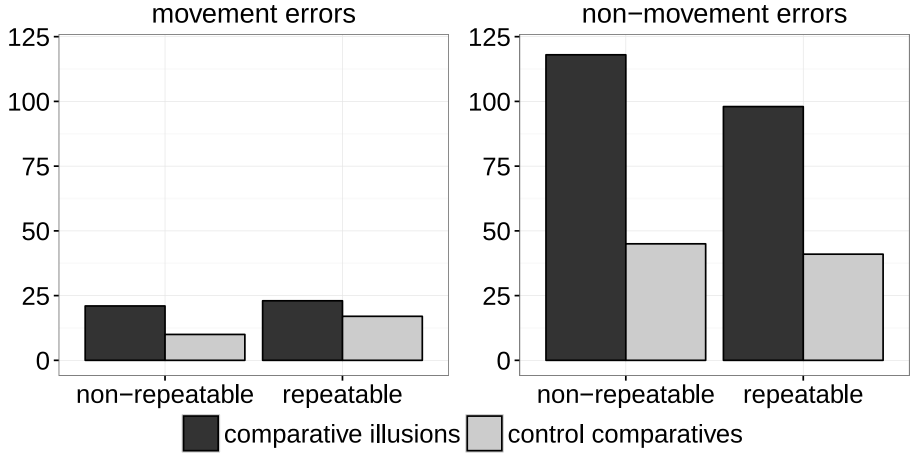
first panel), $\beta = .9, \mathrm{SE} = .48, \chi^2(1) = 3.43, p = .064$. There was no effect of the factor REPEATABILITY for this type of error, $\chi^2(1) = .5, p = .48$, and no interaction between REPEATABILITY and COMPARATIVE, $\chi^2(1) < .1, p = .8$.

We observed a similar pattern of results for non-movement errors, except there were many more errors of this type for the illusion targets than for the control targets (illusion 216, control 86; Figure 13, second panel), linear: $\beta = .61, \mathrm{SE} = .097, \chi^2(1) = 24.04, p < .001$. There were, however, no significant effects of REPEATABILITY at this level of categorization, linear: $\chi^2(1) = 1.5, p = .2$, and no interaction between the factors REPEATABILITY and COMPARATIVE, linear: $\chi^2(1) = .87, p = .35$.

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

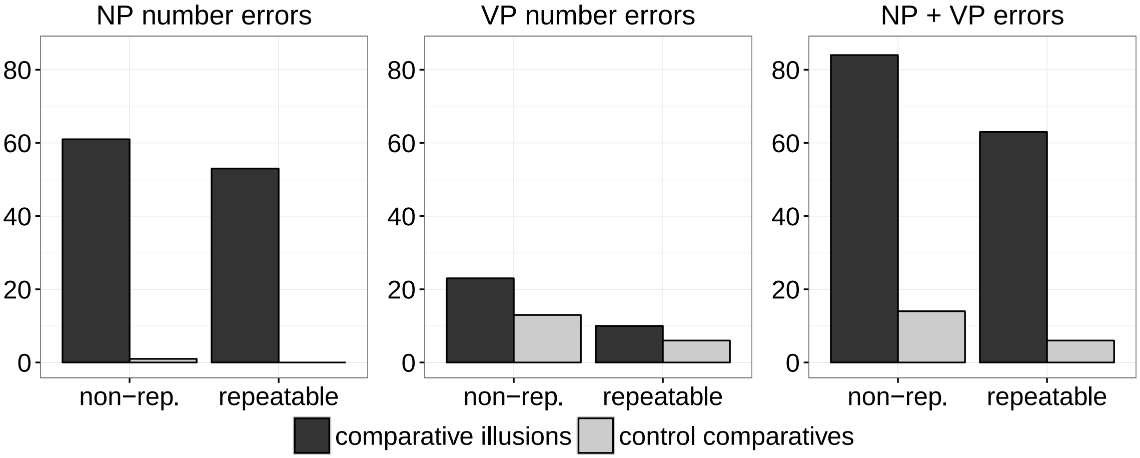
With respect to *than*-clause subject number errors (NP number; Figure 14, first panel), we found that participants made many errors on illusion targets (non-repeatable 61, repeatable 53), and virtually none on control targets (non-repeatable 1, repeatable 0), $\beta = 8.78, \mathrm{SE} = 1.9, \chi^2(1) = 231.6, p < .001$. Recall that errors for illusion targets are singular → plural, and plural → singular on control targets. The paucity of errors in the control conditions violated

Figure 13. Counts of movement and non-movement errors by condition in Experiment 3.



the assumptions of the logistic regression, and we could not analyze the effect of the factor REPEATABILITY on the full dataset. Comparing just within the illusion conditions, we found no effect of this factor on NP number errors, $\chi^2(1) = 1.7, p = .2$.

Figure 14. 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 non-repeatable, and non-repeatable targets repeatable.



Turning to repeatability errors (VP number; Figure 14, second panel), coded only within the ASPECT items, we found that participants made more errors on illusion targets than on control targets (illusion 33, bare plural 19), $\beta = 1.1, \text{SE} = .5, \chi^2(1) = 4.7, p = .03$. We also found that participants made more errors on non-repeatable targets than on repeatable targets (non-repeatable 36, repeatable 16), $\beta = 1.53, \text{SE} = .68, \chi^2(1) = 5.0, p = .03$). There was no interaction between the factors COMPARATIVE and REPEATABILITY, $\chi^2(1) < .1, p = .9$).

Inspecting Figure 14, 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 studies: the most errors were observed in the non-repeatable illusion condition, followed by the repeatable illusion condition; and there was a marginal difference between the non-repeatable and repeatable control conditions. Summing these two error types, this difference is reflected in a main effect of COMPARATIVE, linear: $\beta = .58, \mathrm{SE} = .09, \chi^2(1) = 25.2, p < .001$, and of REPEATABILITY, linear: $\beta = .12, \mathrm{SE} = .06, \chi^2(1) = 3.56, p = .059$. There was no interaction, however, linear: $\chi^2(1) = 1.6, p = .2$.

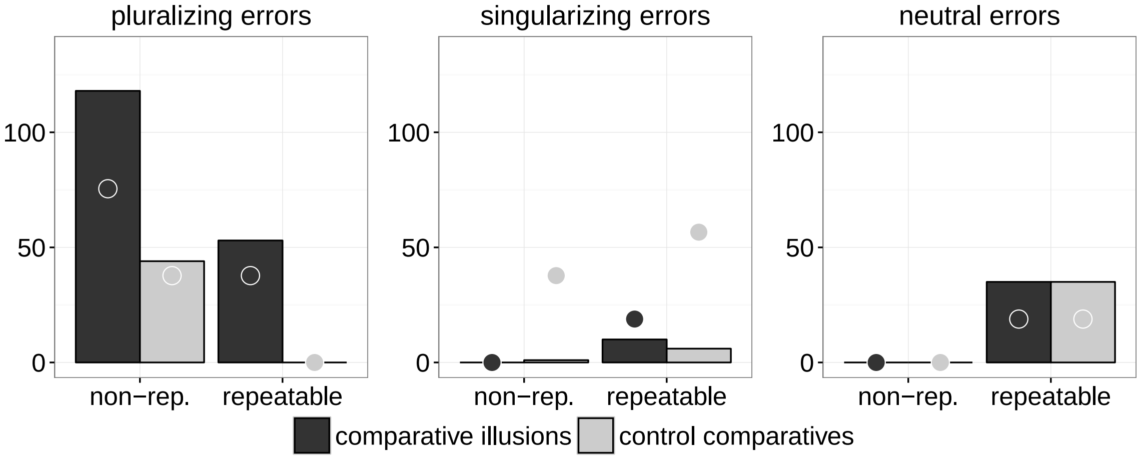
With respect to modifier deletion errors, coded only within the OBJECT items, there were no effects: the critical adjective was deleted at the same rate across the board. Participants did not drop the crucial modifier more on the illusion than on the control targets, $\chi^2 = .52, p = .47$, nor on the non-repeatable than on the repeatable targets, $\chi^2(1) = .26, p = .61$. Similarly, there was no interaction, $\chi^2(1) < .1, p = .89$.

4.1.8. Discussion

Experiment 3 investigated CI-type sentences in production. Overall, we observed many more errors made for CIs than for control sentences. Do these results reflect participants' attempts to assign a meaning to an unacceptable sentence so that it can be recalled, and is 'event comparison' the most relevant predictor of speakers' fixing on that meaning? We think there are reasons to believe that the answer to both of these questions is 'yes'.

The first reason is that errors suggestive of an event comparison reading were much more frequent in our data than were errors that would fail to support that reading, or which were neutral in this respect (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 presented itself. 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.

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 to make each kind of 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 non-repeatable illusion conditions. We found that, out of our 179 illusion trials in which the target sentence's *than*-clause contained a definite description, 48/87 (55.2%) of the non-repeatable trials were rendered grammatical and 42/92 (45.7%) of the repeatable trials were. This suggests that a non-trivial proportion of the higher acceptability judgments for CIs in 3 could have reflected fully grammatical repair.

Why did we observe so many more NP number errors than VP number errors (Figure 14), in light of the role of event comparison? This could simply be due to the fact that the relevant NP error involves a functional item (the morpheme *-s*), while the VP number error involves changing lexical items (e.g., changing *began* to *continue*). It may be more costly to override verbal lexical information in contrast to mere functional information. If so, then we may expect to observe more VP number errors if this CI recall were tested in a language with explicit perfective (singular) and imperfective (non-singular) aspect.

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 semantics of the comparative quantifier: a repeatable VP and a plural *than*-clause subject both support the possibility of an event counting reading, which is (sometimes) licensed in fully grammatical comparatives with bare plurals. 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 presents the first systematic attempt to understand the source of the illusory effect of CIs, so-called 'Escher sentences'. Such sentences have been informally reported to be remarkably acceptable to speakers of English, despite having no coherent sense, and no grammatical analysis according to the theory of comparatives discussed in §1. Early consideration of the phenomenon motivated researchers to suggest that they reflect a sort of 'shallow' processing—speakers fail to notice the anomaly, because they aren't really attending to rich grammatical detail. In contrast, our results suggest that fine-grained semantic properties play a role in determining how acceptable speakers find CI-type sentences.

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 1 and 2, as well as three preliminary experiments) suggest that, in general, the acceptability of CI-type sentences is highly variable both within and across studies. We speculated that this could be due to the likelihood of participants noticing the mismatch between form and meaning—if people notice the anomaly, they may be more likely to assign it a lower rating. Across all studies, the illusion sentences were more likely to be rated more acceptable whenever the *than*-clause could be construed as providing a plurality of events.

Next, we asked how far the effect generalizes beyond the canonical example in (1). Native speaker consultants and linguists alike have informally suggested various accounts of what drives the illusion, each of which makes different predictions about how far it should generalize. For instance, it might be due to a processing mechanism that finds the CI matches familiar clausal templates (§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 doesn't support: either they misanalyze the quantifier *more* in its homophonous additive sense (§2.3), or they persist in an interpretation in terms of a comparison of numbers of events (§2.4).

Our two 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 is plural (thus providing multiple events, via a multiplicity of agents). This interpretation is grammatically legitimate in the matrix clause, but is not supported by the syntax of the *than*-clause. Thus, we find that the explanation for the illusion lies, in part, in a failure to notice that the event comparison reading, however tempting, is not an interpretation that the sentence ultimately allows.

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 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 faithful to the syntax of the construction, while nevertheless persisting in an event comparison interpretation.

What do these results lead us to expect from a processing perspective? When comprehenders encounter a comparative clause like *More people*, they posit a nominal comparative construction. Upon encountering *have been to Russia*, they implicitly recognize that the sentence is compatible with two construals: a comparison of a number of individuals, or of a number of events. Encountering *than*, they posit the operator corresponding to *how many*, and wait to associate it with a suitable variable. The syntactic position of the variable is uncertain at this point, but its category is grammatically fixed—it should associate with a nominal. If the sentence continues as in (2), they encounter a suitable nominal immediately; if it continues as (1), what happens next depends on the predicate: in the non-repeatable case participants readily recognize that the syntactic constraint cannot be fulfilled, but they are less likely to notice this in the repeatable case. Why?

Since our participants did not reanalyze *more* as an adverbial in the recall experiment, we think that the illusion arises when the *than*-clause operator is mistakenly classified based on its semantic rather than syntactic properties. That is, people attempt to analyze the operator as involving an illicit adverbial attachment, which is more likely to be successful (i.e., pass undetected) whenever a plurality of events is possible. If so, we might expect a processing cost for comparative illusions that are independent of the predicate, or indeed of the ratings ultimately assigned. O'Connor's (2015) results are suggestive here: she found that participants' reading times were slower following the ellipsis site for CI-type sentences as compared to controls, and that the reaction times correlated with acceptability ratings for controls but not for illusions. If shallow processing was behind high acceptability ratings for CIs, we would expect to see no slowdown when participants eventually gave a high rating.

Such an account is compatible with an asymmetry we observe for Bulgarian, where only the overt occurrence of the counterpart of this operator—*kolkoto*—leads to categorical unacceptability for CI-type sentences; that is, the continuation in (47a) is perfectly acceptable while that in (47b) is judged bad, and (48) is judged illusory in the same manner as its English correspondent. Thus, the presence of an overt form of the operator could lead to a stronger, less violable error signal that it has no suitable syntactic correspondent.

(47) 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.'

(48) Poveče hora sa bili v Rusija ot men.
more people are been in Russia than me
'More people have been to Russia than me.'

While the details of this sort of account await future research, our suggestion is that we should not observe CI-type effects in languages where the head of the operator-variable dependency is overt.[13]

The illusion involves entertaining an event-counting interpretation early on in the matrix clause, potentially as early as the matrix verb phrase is encountered. Ultimately, this interpretation is so tempting that comprehenders are often blinded to the fact that the syntax doesn't literally support it; it is grammatically illicit to posit either a determiner or adverbial position for the covert operator in the *than*-clause of a sentence like (1). Yet, one thing that speakers know is that, in case the matrix clause supports a plurality of events, (1) nonetheless meets the semantic requirements of comparative *more*. Because of the mismatch between syntax and semantics, though, this interpretation is not always stable.

An important question raised by an anonymous reviewer is why, particularly as regards the additivity hypothesis, we should see a mismatch between the conscious justifications that speakers give for the acceptability and interpretability of (1), and the interpretations that our experiments find evidence for. Either the "additive" paraphrase is genuinely the source of the illusion for some speakers, but we simply failed to detect it in our experiments, or speakers' conscious reports do not accurately reflect why the sentences sound so natural to them during incremental interpretation. We think it likely that the additive paraphrase is simply easier for speakers to articulate than the event comparison paraphrase, as few non-semanticists will possess the tools needed to explain the difference between comparisons of individuals and comparisons of events.

[13] More generally, we may expect different patterns of CI-type effects depending on fine-grained syntactic details of a language; cf. Christensen 2016 for an investigation of CI-type sentences in Danish.

This study suggests that CIs thus do not illustrate processing in the absence of detailed analysis, but underscore the importance of such analysis in sentence processing. It is thus informative for discussions about the need for interpretive mechanisms that are in some sense distinct from the process of analysis that formal syntacticians and semanticists generally worry about. In contrast to ‘good enough’ approaches, for example (e.g., Christianson et al. 2001; see Ferreira & Patson 2007 for an overview), these theories make direct predictions about the scope and limits of CI-type phenomena. While our account has features reminiscent of such approaches, it depends for its explanation on detailed proposals about the syntax-semantics of comparatives crosslinguistically. These proposals severely delimit the space of things that can go wrong, and allow us to make specific, testable predictions about how far CI-type effects should generalize, both in English and in other languages.

Acknowledgments This work was supported by a UMD Baggett Fellowship and SSHRC award 752-2010-0499 to AW, a Mellon Foundation Fellowship to RP, NSF BCS-0848554 to CP, and NSF IGERT award SBE-0801465 to the University of Maryland. Thanks go to the audience at the CUNY 2009 Human Sentence Processing Conference in Davis, to the Cognitive Neuroscience of Language group at UMD, as well as Masaya Yoshida, Scott Fults, Alexander Williams, Barry Schein, Norbert Hornstein, Jeffrey Lidz, Paul Pietroski, Brian Dillon, Ellen Lau, Alan Bale, Benjamin Spector, Shevaun Lewis, Ewan Dunbar, Matt Goldrick, Klinton Bicknell, Ellen O’Connor, Alex Fine, Roger Levy, Hal Tily, aand Michaël Gagnon for discussion, as well as Dave Spencer and Nathan Ycas for their help running the experiments.

Appendix A. Instructions for acceptability tasks

The following instructions were provided to participants in both Experiments 1 and 2. The scalar values indicated following the example sentences were circled on a sample 1-7 scale.

> Welcome to the experiment!
>
> In this experiment, you will read many sentences. For each sentence, please rate the sentence based on whether you think it is an acceptable sentence (6 or 7) or an impossible/unacceptable sentence (1 or 2). Some sentences may not sound completely impossible, while also not being completely acceptable—in these cases, use the more intermediate ratings (3-5).
>
> Note, however, that you are not being asked to judge whether the sentence is plausible or not (i.e. it would require 'too strange' a context to make the sentence plausible); rather, you are only being asked to judge whether the sentence sounds like possible English or not. For example, (a) below describes a likely scenario, but most English speakers find it unacceptable (in contrast to (b)). Sentence (c) describes an unlikely scenario, yet given the proper situation, you could write/speak (c) without any problem.
>
> a. The children decorated the sparkling ornaments onto the tree. [2/7]
> b. The children decorated the tree with sparkling ornaments. [7/7]
> c. The purple elephant played chess with the balding porcupines. [7/7]
>
> You are also not being asked to judge whether the sentence is acceptable according to grammatical rules you may have learned in school—only whether the sentence sounds natural and good. For example, people often say that it's 'bad' to end a sentence with a preposition like *with*, however most English speakers find (d) below to be a perfectly fine sentence (in contrast to (e)).
>
> d. I know who Julie saw Mary with. [7/7]
> e. I know who Julie saw Mary and. [2/7]
>
> As you work through the sentences on the following pages, please keep in mind that each sentence is different, and you may feel very differently towards two sentences which at first seem superficially similar. In that respect, judge each sentence individually, and not in comparison with other sentences you have read.

Appendix B. Experiment 1 conditions

Tabular version of the item schematic from Experiment 1 (Figure 3). The factor REPEATABILITY was manipulated between items, the other factors were manipulated within items. The factor SUBJECT INCLUSION was counterbalanced across the illusion conditions; these sample items represents 'inclusion not possible' items.

Sentence	COMP.	QUANT.	ELLIP.	REPEAT.
More girls ate pizza than the boy did.	illusion	more	ellipsis	repeatable
More girls ate pizza than the boy ate yogurt.	illusion	more	no ellip.	repeatable
More girls ate pizza than boys did.	control	more	ellipsis	repeatable
More girls ate pizza than boys ate yogurt.	control	more	no ellip.	repeatable
Fewer girls ate pizza than the boy did.	illusion	fewer	ellipsis	repeatable
Fewer girls ate pizza than the boy ate yogurt.	illusion	fewer	no ellip.	repeatable
Fewer girls ate pizza than boys did.	control	fewer	ellipsis	repeatable
Fewer girls ate pizza than boys ate yogurt.	control	fewer	no ellip.	repeatable
More girls graduated H.S. than the boy did.	illusion	more	ellipsis	nonrep.
More girls graduated H.S. than the boy ate yogurt.	illusion	more	no ellip.	nonrep.
More girls graduated H.S. than boys did.	control	more	ellipsis	nonrep.
More girls graduated H.S. than boys ate yogurt.	control	more	no ellip.	nonrep.
Fewer girls graduated H.S. than the boy did.	illusion	fewer	ellipsis	nonrep.
Fewer girls graduated H.S. than the boy ate yogurt.	illusion	fewer	no ellip.	nonrep.
Fewer girls graduated H.S. than boys did.	control	fewer	ellipsis	nonrep.
Fewer girls graduated H.S. than boys ate yogurt.	control	fewer	no ellip.	nonrep.

Appendix C. Experiment 2 conditions

Tabular version of the item schematic from Experiment 2 (Figure 6). The factor REPEATABILITY was manipulated between items, and SUBJECT TYPE was manipulated within items.

Sentence	Person	Sort	Number	REPEAT.
More girls ate pizza than I did.	1st	pronoun	singular	repeatable
More girls ate pizza than we did.	1st	pronoun	plural	repeatable
More girls ate pizza than the boy did.	3rd	definite	singular	repeatable
More girls ate pizza than the boys did.	3rd	definite	plural	repeatable
More girls ate pizza than he did.	3rd	pronoun	singular	repeatable
More girls ate pizza than boys did.	**control**	**NP**	plural	repeatable
More girls graduated H.S. than I did.	1st	pronoun	singular	nonrep.
More girls graduated H.S. than we did.	1st	pronoun	plural	nonrep.
More girls graduated H.S. than the boy did.	3rd	definite	singular	nonrep.
More girls graduated H.S. than the boys did.	3rd	definite	plural	nonrep.
More girls graduated H.S. than he did.	3rd	pronoun	singular	nonrep.
More girls graduated H.S. than boys did.	**control**	**NP**	plural	nonrep.

Appendix D. Instructions for sentence recall task

In the recall task, participants were told by the experimenter that the task is a memory task. We're interested in how well they can recall sentences aloud, after an intermediate task designed to make this more difficult. ollowing this verbal instruction, participants read the following instructions on the screen. They then had 6 practice trials while the experimenter remained to answer any questions they had.

> At the start of each trial, you will see a cross +, followed by ##### and a SENTENCE. At the end of the sentence, you will see %%%%%, followed by a LIST of five words, and #####.
>
> You will see a CAPITALIZED word. Press J if this word was in the LIST, press F if this word was not in the LIST.
>
> Immediately afterwards, recall the SENTENCE aloud. When you are finished speaking, press ENTER.
>
> Then you may take a brief break, or go immediately to the next trial.
>
> This will be clearer in a moment. Press SPACEBAR for some practice.

Appendix E. Sample items from Experiment 3

Tabular version of the item schematic from Experiment 3 (Figure 9). There were two types of items (aspect and item), and the factors comparative and repeatability were manipulated within item. '...' abbreviates an initial adverbial phrase. Here 'W&P' abbreviates *War & Peace*; none of our items contained abbreviations.

Sentence	item type	comp	repeat.
... more young people continued reading W&P than the old man did.	aspect	illusion	repeatable
... more young people continued reading W&P than old men did.	aspect	control	repeatable
... more young people began reading W&P than the old man did.	aspect	illusion	nonrep.
... more young people began reading W&P than old men did.	aspect	control	nonrep.
... more girls wrote a charming haiku than the boy did.	object	illusion	repeatable
... more girls wrote a charming haiku than boys did.	object	control	repeatable
... more girls wrote their first haiku than the boy did.	object	illusion	nonrep.
... more girls wrote their first haiku than boys did.	object	control	nonrep.

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.

Bartsch, Renate & Theo Vennemann. 1972. *Semantic structures: A study in the relation between semantics and syntax*. Frankfurt am Main: Athenaum.

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.

Bhatt, Rajesh & Roumyana Pancheva. 2004. Late merger of degree clauses. *Linguistic Inquiry* 35(1). 1–46.

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.

Christensen, Ken Ramshoj. 2016. The dead ends of language: The (mis)interpretation of a grammatical illusion. In *Let us have articles betwixt us – Papers in historical and comparative linguistics in honour of johanna l. wood*, 129–160. Department of English, School of Communication & Culture, Aarhus University.

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.

Deschamps, Isabelle, Galit Agmon, Yonatan Lewenstein & Yosef Grodzinsky. 2015. The processing of polar quantifiers, and numerosity perception. *Cognition* 143. 115–128.

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.

Heim, Irene. 2000. Degree operators and scope. In Brendan Jackson & Tanya Matthews (eds.), *Proceedings of SALT X*, 40–64. Cornell University, Ithaca, NY: CLC Publications.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in generative grammar*. Malden, MA: Blackwell.

Kennedy, Chris. 1999. Gradable adjectives denote measure functions, not partial functions. *Studies in the Linguistic Sciences* 29(1). 65–80.

Kennedy, Chris. 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. 2015. *Comparative illusions at the syntax-semantics interface.* Los Angeles, CA: University of Southern California dissertation.

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. 2016. Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition* 157. 321–339.

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. 2017. *'and': Conjunction Reduction Redux.* Cambridge, MA: MIT Press.

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).

Wagers, Matthew W., Ellen F. Lau & Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61(2). 206–237.

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.