THE ANATOMY OF A COMPARATIVE ILLUSION: SUPPLEMENT

ALEXIS WELLWOOD, ROUMYANA PANCHEVA, VALENTINE HACQUARD, COLIN PHILLIPS

In this supplement, we describe our original set of three experiments (Experiments 1a-c, not reported in the published paper), investigating the acceptability of so-called 'comparative illusions' (CIs) like (1), and uncontroversial nominal comparatives like (2). The materials for Experiment 1a, run with undergraduates at the University of Maryland, were used in the Amazon Mechanical Turk study reported in the published article as Experiment 1.

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

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

In our original discussion introducing our acceptability judgment studies, we combined the results of Experiments 1a-c with those of Experiment 2 (reported under that name in the published paper) in order to highlight the variability in responses across experiments that we observed with CIs. Much of the text of this supplement overlaps with that of the published article as it formed part of the original draft of that paper; in the published version, the discussion is much leaner, focusing on just Experiment 1.

1. Original data and discussion

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

First, how acceptable are CIs once they are put in a formal experimental setting? As we will see directly, the patterns of acceptability responses for sentences like (1) and (2) differed substantially: the mean acceptability ratings for CIs were much lower than were those for control sentences, and the distributions of ratings were much more variable. Importantly, however, the pattern we observed in the ratings for CIs also differed substantially from those of 'bad' filler sentences, while that for control comparatives did not differ substantially from those of 'good' filler sentences.

For now, we limit our attention to 'comparative illusions' and 'control comparatives' that differ minimally from (1) and (2): subject nominal comparatives with repeatable predicates and *than*-clause ellipsis. CIs, here and below, differ from controls only in the type of *than*-clause subject: CIs have a non-bare plural subject, while controls have a bare plural subject. Our filler sentences were designed to elicit either a low or high rating while having a similar length, degree of syntactic complexity, and, in around one-third of the cases, a similar degree of semantic complexity (i.e. comparative-type meanings) to target items, (3)-(4).

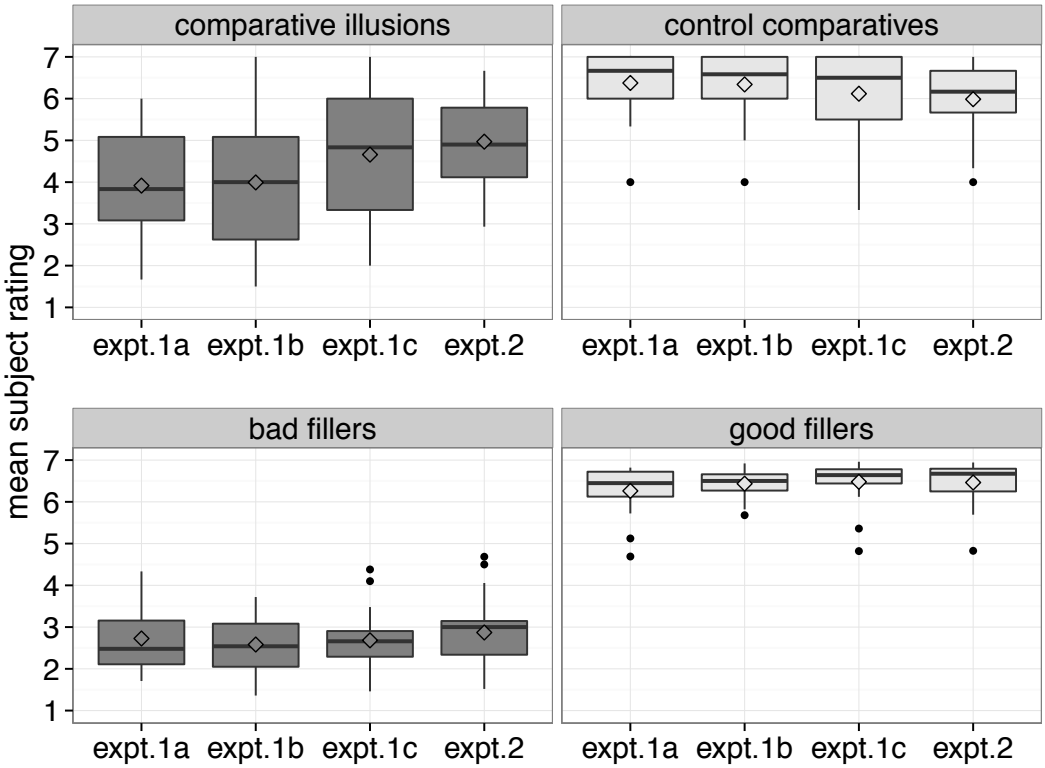
(3) **Examples of 'bad' fillers**

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.

Date: 08 August 2016.

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



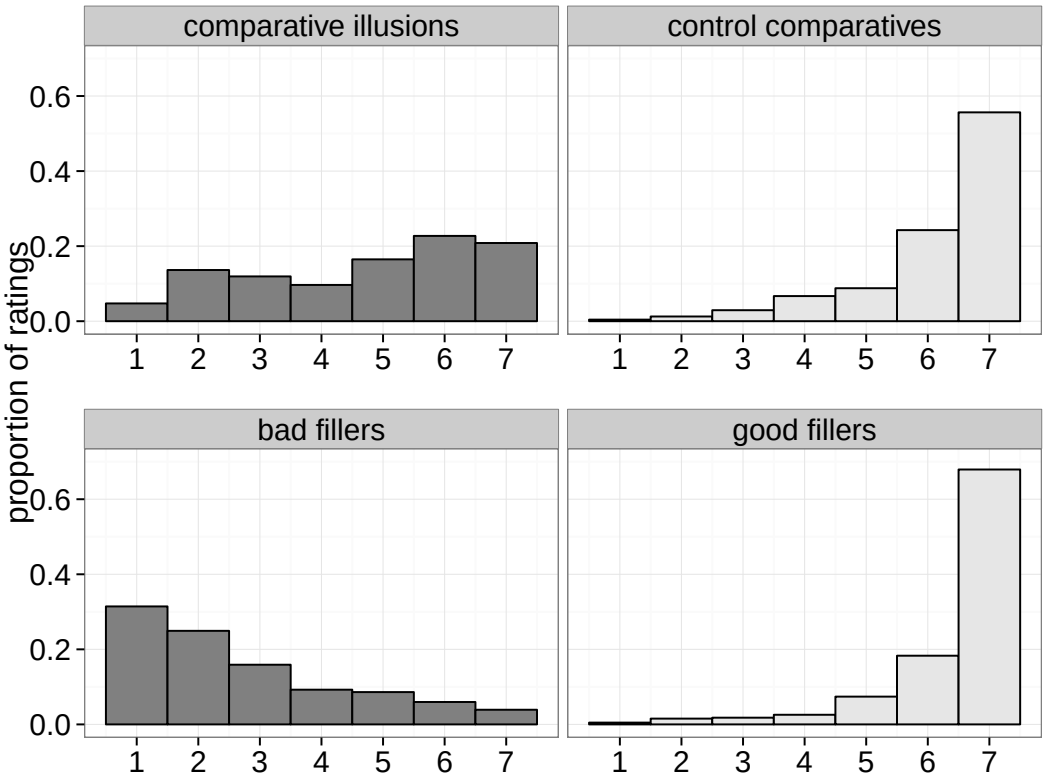
(4) **Examples of 'good' fillers**

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

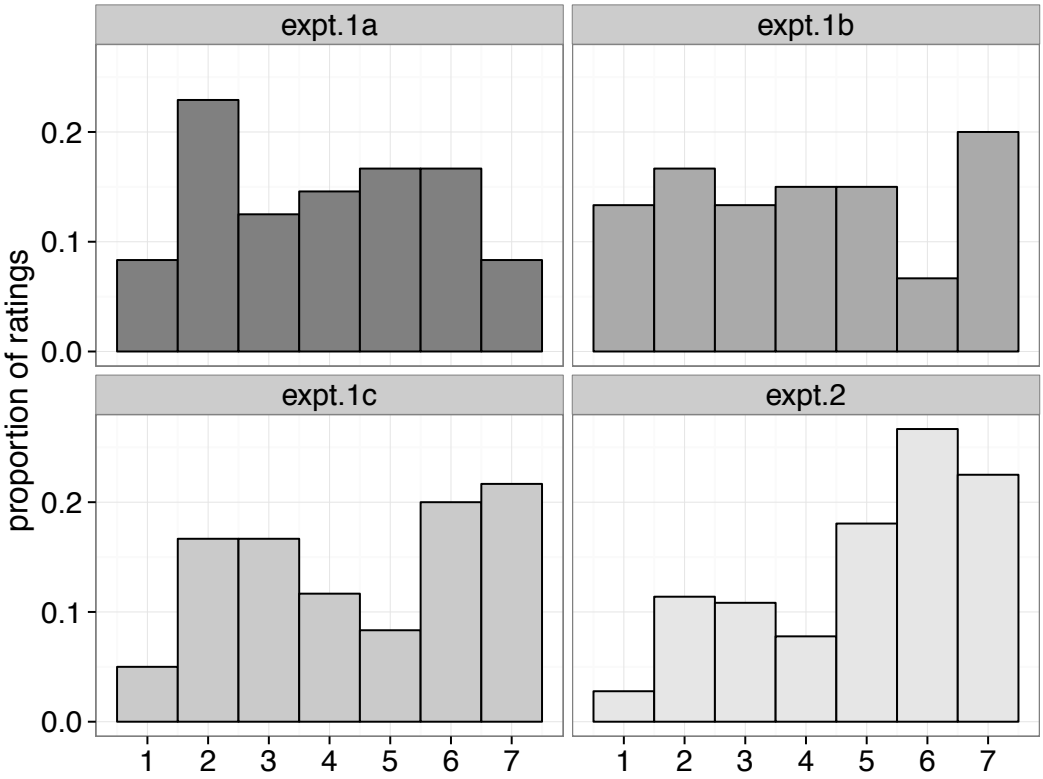
FIGURE 2. Density plots of rating distributions for classic CI-type sentences and controls (top row), and bad and good fillers (bottom row), across acceptability experiments. Each plot represents the proportion of total responses observed for each scalar value.



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

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

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



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

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

1.1. Experiment 1

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

Each hypothesis outlined in the published paper predicts that specific factors should make speakers more susceptible to CIs. For each of these hypotheses, it is necessary to identify manipulations that selectively impact CIs to the exclusion of fully grammatical controls.

Therefore, in all experiments in this series, our primary manipulation was a comparison of illusion and control conditions (the ILLUSION factor), where the illusion conditions were defined by having non-bare plural subjects in their *than*-clauses, and the control conditions were defined by having bare plural subjects in their *than*-clauses. In each study, this factor was crossed with a subset of further factors that specific hypotheses predict should impact acceptability.

The factor QUANTIFIER manipulated the comparative quantifier in the main clause subject position (*more* vs. *fewer*; (5)). This manipulation was used to test two of the four hypotheses about the source of the CI effect. The template matching hypothesis (more constrained version) relies on the ambiguity of *more* as both a determiner and an adverbial. The additive *more* hypothesis relies on the ambiguity of *more* as a lexical item with either a comparative or an additive semantics. Since *fewer* cannot function as an adverbial, and it lacks an additive semantics, both hypotheses predict that CI-type sentences with *fewer* should fail to elicit the CI-effect. (The examples illustrating the factors are simplified for presentation purposes; these are not actual experimental items.)

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

The factor ELLIPSIS manipulated whether VP ellipsis had applied in the *than*-clause (ellipsis vs. no ellipsis; (6)). This manipulation was used to test the repair-by-ellipsis hypothesis. That account holds that the acceptability of CIs depends on VP ellipsis, and thus predicts that CI-type sentences with an unelided VP should not elicit the CI effect. However, previous research supporting this prediction (Fults and Phillips 2004) failed to take into account the grammatical preference for deletion in comparatives (Bresnan 1973). In our design, the VP in the 'no ellipsis' conditions differed just enough from the matrix clause VP to circumvent this preference.

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

The factor PREDICATE TYPE manipulated whether the VP in the comparative was repeatable for a given agent (repeatable vs. nonrepeatable; (7)). This manipulation was used to test the event comparison hypothesis, which holds that the CI effect is due to a persistent event-comparison reading; this interpretation is only grammatically licensed in the matrix clause if it contains a repeatable predicate like *eat pizza*. This account thus predicts that CI-type sentences with nonrepeatable predicates like *graduate high school* should fail to elicit the CI effect.

(7) PREDICATE TYPE
More girls **ate pizza/graduated high school** than the boy did.

The factor SUBJECT INCLUSION manipulated whether the denotation of the *than*-clause subject could be included in the denotation of the subject NP of the matrix clause ('inclusion possible' vs. 'inclusion not possible'; (8)). This manipulation was used to test the additive *more* hypothesis, which requires the possibility of an inclusion relation in order to license a 'not just me' interpretation of the comparative. In our design, 'inclusion possible' trials usually involved gender-matching the NPs the matrix and *than*-clause subjects, and 'inclusion not possible' involved mismatching these NPs. The additive *more* hypothesis predicts that 'inclusion not possible' trials should fail to elicit the CI effect.

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

Sample repeatable item

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

Sample nonrepeatable item

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

(8) SUBJECT INCLUSION
More **boys** called to complain than **he/she** did.

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

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

Sets of items were distributed across 8 lists in a Latin Square design, and were then combined with filler sentences for roughly a 1/3 ratio of experimental to filler sentences, creating 8 questionnaires. The number of items differed across experiments in the series; specific details are given below. Fillers were designed to approximate the complexity of the experimental items, and were evenly split between those that should elicit lower and higher ratings (see discussion of (3) and (4) 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 (3a) and (4a) for examples), which were included to help mask the experimental items.

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

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

1.1.1. Experiment 1a

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

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

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

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

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

There was a reliable main effect of QUANTIFIER, with the *more* conditions rated more highly than the *fewer* conditions (*more* 4.61, *fewer* 4.38; $\beta = .24, \text{SE} = .1, \chi^2(1) = 5.77, p < .02$). Importantly, there was no interaction of ILLUSION and QUANTIFIER ($\beta < .01, \text{SE} = .22, \chi^2(1) < .01, p = .98$). Replacing *more* with *fewer* was associated with a small decline in ratings in the illusion conditions (*more* 3.22, *fewer* 2.99) and in the control conditions (*more*

5.99, *fewer* 5.76). These results are not consistent with the syntactic template matching hypothesis or the additive *more* hypothesis, both of which predicted a substantial difference between the comparative quantifiers specifically within the illusion conditions.

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

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

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

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

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

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

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

Second, the experiment was 192 trials long, which may have provided participants ample time to notice the anomaly. If this explanation were correct, however, we might expect that

the average ratings for CI-type sentences would decline more over time than the average ratings for control sentences. This possibility was not supported by a linear regression analysis: average ratings within the illusion and control conditions remained stable between the first and second halves of the experiment (ORDER × ILLUSION $\beta = .001, \text{SE} = .002, \chi^2(1) = .67, p = .41$).

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

1.1.2. Experiment 1b

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

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

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

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

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

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

5.86, no ellipsis 5.62). These results are consistent with Experiment 1a in failing to support the repair by ellipsis hypothesis.

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

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

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

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

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

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

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

1.1.3. Experiment 1c

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

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

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

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

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

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

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

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

possible 3.46, not possible 4.18). These results are in the same unexpected direction as in Experiment 1a, and fail to support the additive *more* hypothesis.

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

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

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

However, the higher overall means are also consistent with another possibility. This experiment included plural subjects of the *than*-clause in CI-type sentences, which could themselves allow for satisfaction of the event-counting reading. That is, even with a nonrepeatable predicate like *graduate high school*, a plurality of events can be inferred if the subject is plural: one event for each member of the plural subject; cf. (9). Any non-bare plural subject NPs in the *than*-clause of a subject comparative is equally ungrammatical according to syntactic-semantic theory (see the published paper for discussion and citations); yet, if such phrases are plural, they can nonetheless support the inference of a plurality of events.

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

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

1.2. Discussion of Experiments 1a-c

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

The event comparison hypothesis predicted an interaction between the factors ILLUSION and PREDICATE TYPE. If the CI-effect requires that the predicate be 'repeatable' for a given agent, then CI-type sentences with such predicates should be judged more acceptable than those with nonrepeatable predicates. The only consistently reliable interaction effect that we

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

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

found in Experiments 1a-c was due to the difference between repeatable and nonrepeatable predicates, supporting the event comparison hypothesis.

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

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

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

CI-type sentences were not rated as highly as their fully grammatical and interpretable counterparts, but they were most acceptable across Experiments 1a-c when the understood predicate was repeatable. The mean ratings for CIs was overall highest in Experiment 1c,

when a variety of subject-types (singular and plural) appeared in the *than*-clause. One possible explanation for this is that variety decreased the salience of the CIs, and thus made participants less likely to notice the anomaly on a given trial.

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

References

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

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