Appendix 1: Post-hoc analyses on grammaticality of missing NP sentences

In claiming that Mandarin Chinese missing NP sentences – specifically, the plausible ones exemplified by (5) in the article – elicit a center-embedding illusion, we are claiming that these sentences are ill-formed and so lack a grammatical parse. As mentioned in Section 4 of the article, we constructed the Mandarin experimental materials in a way to rule out certain alternative parses. However, as an anonymous reviewer and Waltraud Paul (as associate editor of *Glossa*) independently pointed out, there might still be other grammatical parses available for these sentences, which would explain why these sentences have relatively high ratings.

We take these proposals seriously. We recognize that there are other cases where there is uncertainty over the correct parse of Chinese sentences. However, it is important to note that the mere perception of acceptability is no guarantee that a sentence does, in fact, have a well-formed parse. That is the point of studies on grammatical illusions. For example, comparative illusions like *More people have been to Russia than I have* are demonstrably ill-formed, though they can continue to be perceived as acceptable, even after many exposures (Wellwood et al. 2018).

We illustrate two putative parses below, repeating example (5) from the article as (i). The two parses are syntactically different: the first parse (ia) involves conjoining V1 and V2, as if there were a silent "and" (*bìngqiě*, to quote the anonymous reviewer) between them, so that they form a conjoined predicate of the sentence's subject "prime minister." In the second parse (ib), V1 is the main verb, while V2 is the main verb of a relative clause in which a *de* has been omitted. V2's subject is omitted as well, but interpreted as coreferring to the subject "prime minister." This relative clause is stacked with another relative clause whose main verb is V3 "hear." Despite the syntactic difference, both parses encode the same thematic relations: the prime minister meets a judge and has previously rebuked the same judge.

(i) Mandarin plausible missing NP sentence

Zǒnglǐ	zhèngzài [$_{V1}$	jiējiàn]	céng	[$_{V2}$	zébèi-guo]	gāng	[$_{V3}$	shěnlǐ]
prime.minister	PROG	meet	previously		rebuke-EXP	recently		hear

[$_{NP1}$ tānwū-àn] bùjiǔ de [$_{NP2}$ fǎguān].
corruption-case not.long.ago DE judge

a. Conjoined V1 and V2 analysis: 'The prime minister is meeting **and** has previously rebuked the judge who just heard the corruption case not long ago.'

b. Stacked relative clause analysis: 'The prime minister is meeting the judge [previously rebuked by him] [who had just heard the corruption case not long ago].' (Paul, p.c.)

We agree that the question of availability of grammatical parses is an important one. If our key sentences are, in fact, grammatical, then it undermines the point of our paper. We have run a set of post-hoc analyses on the acceptability data for Experiment 1, to ascertain the extent to which missing NP sentences could have received a grammatical parse. (We remain agnostic about what grammatical parses these sentences might have: our data only provide evidence about thematic relations, not necessarily about the structure.)

To preview our results, we found that:

1. Responses to missing NP sentences have a very different distribution compared to responses to structurally similar filler sentences that are unambiguously grammatical. Missing NP sentences are much likelier to get very low acceptability ratings.

2. There is no evidence that the distribution of responses is because there are two distinct populations of Mandarin-speaking participants, one for whom missing NP sentences are clearly acceptable, and another for whom missing NP sentences are clearly unacceptable.
3. There is often inconsistency within a participant in how missing NP sentences are rated.
4. There is substantial variability in ratings at the item level – ratings vary between sentences and within a sentence. A simulation further suggests that the between-item variability is expected given the general variability in ratings for missing NP sentences, instead of reflecting deeper differences between missing NP items (e.g. differences in grammaticality).
5. We repeat these analyses on English missing VP sentences (from Experiment 3), which are unambiguously grammatical illusions. We show that they exhibit the same variability profile as Mandarin missing NP sentences.

The variability we observed is consistent with a scenario where missing NP sentences are ungrammatical and where the illusion is probabilistic in nature. The fact that English missing VP sentences also show similar variability provides further evidence to support this conclusion.

Comparison with filler items

One way of inferring the status of missing NP sentences is by comparing them with the grammatical filler sentences in Experiment 1. These fillers were intended to be similar in structure and length. For all analyses reported below, we have excluded responses given in less than 3 seconds after the presentation of the sentences, to ensure that participants had taken the time to read and judge each sentence.

Figure 1 shows that at the item level, plausible missing NP sentences have distinctly lower ratings than grammatical fillers. (Although these missing NP sentences had relatively high ratings compared to the other conditions in Experiment 1, the mean rating for these sentences is only 3.66 out of 7.)

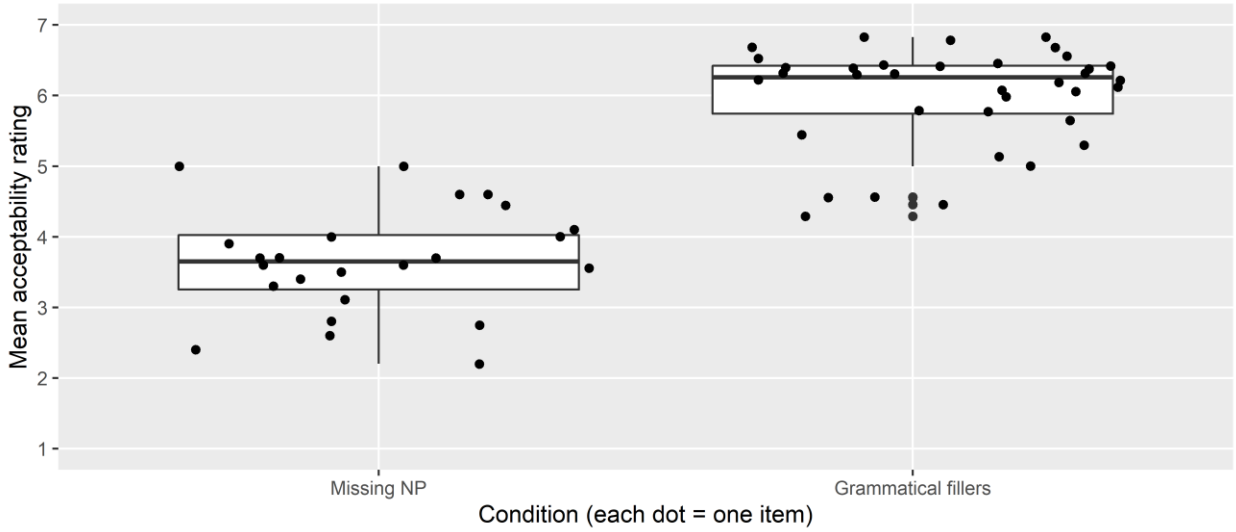

Figure 1: Distribution of ratings across items in plausible missing NP condition and grammatical fillers.

Another way of approaching this issue is by looking at the overall distribution of ratings for these conditions, aggregating across trials. For example, in the 235 responses for plausible missing NP sentences (60 participants x 4 items / participant, excluding responses given in less than 3 seconds), how many responses were a 7 (highly acceptable), or a 6, and so on?

Figure 2 shows that ratings for missing NP sentences have a relatively flat distribution. Many trials had low ratings – about 35% of all responses were a 1 or 2. This contrasts sharply with the grammatical fillers, where only 4% of responses were rated as a 1 or 2. The proportion of low ratings for missing NP sentences is surprising, if these sentences have grammatical parses.

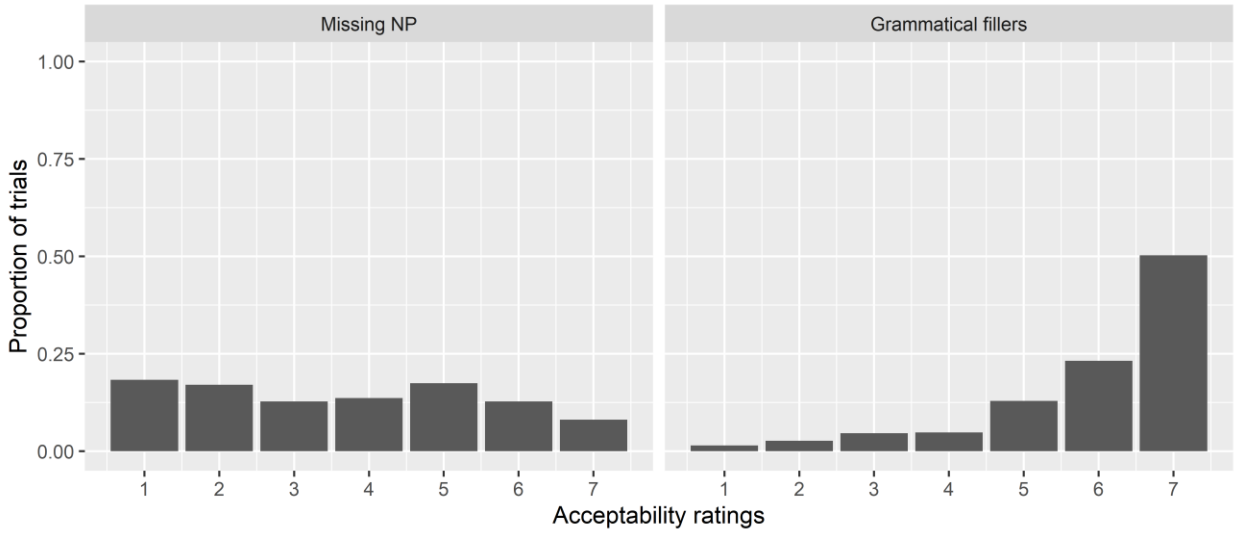

Figure 2: Distribution of ratings across trials in plausible missing NP condition and grammatical fillers.

Figures 3 and 4 present the same analyses for the English missing VP illusion data. This is a useful point of comparison, because the English examples are uncontroversially ungrammatical. Figure 3 shows a similar contrast in acceptability between the 12 illusory missing VP sentences ("Missing VP2" in Experiment 3) and grammatical fillers. Figure 4 shows that ratings for missing VP sentences across all trials have a flatter profile, like Mandarin missing NP sentences.

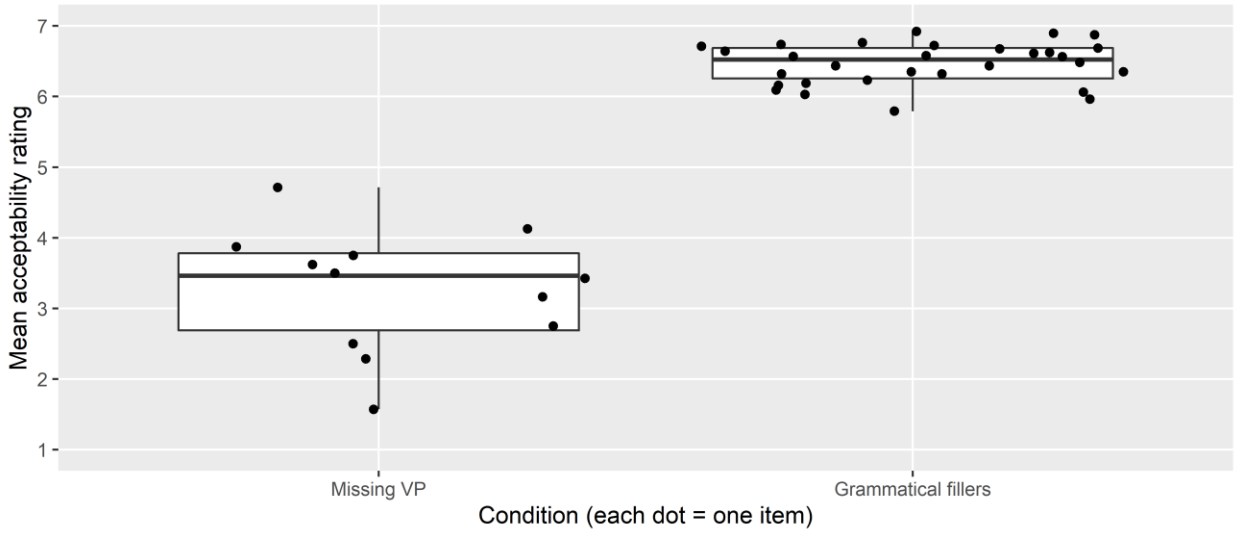

Figure 3: Distribution of ratings across items in English missing VP condition and grammatical fillers.

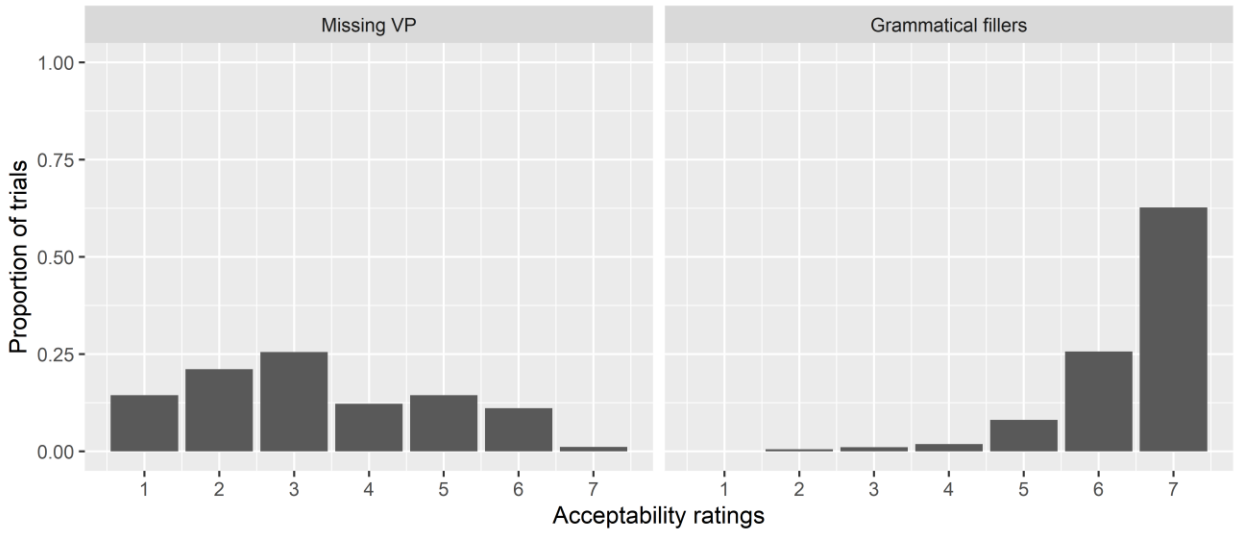

Figure 4: Distribution of ratings across trials in English missing VP condition and grammatical fillers.

For thoroughness, we also compare the Mandarin missing NP sentences with the subset of grammatical fillers in which a relative clause modifies an object, which are structurally the most similar to missing NP sentences. (The other fillers involve fronting of phrases, which produces non-canonical word order, and argument (object) ellipsis). These fillers fall into two types. In the first type (iia), the relative clause is formed by relativizing the first NP in the so-called "double subject" (or "major subject") construction in which two NPs occur in clause-initial, subject-like positions (iib). The second type of relative clauses is formed by relativizing the subject of an SVO clause (iic). We label them the "double subject relative clause (RC)" and "regular subject RC" fillers respectively.

(ii) a. Zhè wèi lǎoshī jīntiān zǎoshàng chòu-mà-le __ xiàofú hěn zāng de
this CL teacher today morning angry-scold-PFV school.uniform very dirty DE
xuéshēng yī-dùn.
student one-time
'This morning, this teacher angrily scolded the student whose school uniform was very dirty.'
("Double subject RC")

b. [$_{NP}$ Zhè gè xuéshēng] [$_{NP}$ xiàofú] hěn zāng.
this CL student school.uniform very dirty
'This student's uniform is very dirty.' ("Double subject construction")

c. Zhè míng jìzhě cǎifǎng-le __ cānyù bǐsài xǔduō-cì de xuǎnshǒu.
This CL reporter interview-PFV participate competition many-time DE athlete
'This reporter interviewed the athlete who participated in the competition many times.' ("Regular subject RC")

By focusing on these filler sentences, we make sure that the low ratings for missing NP sentences do not reflect some idiosyncratic participant bias against objects modified by relative clauses. Figure 5 shows that these grammatical fillers are generally acceptable, although we observe a slightly wider spread in ratings for the double subject RC fillers. There is still a clear contrast between missing NP sentences and these fillers.

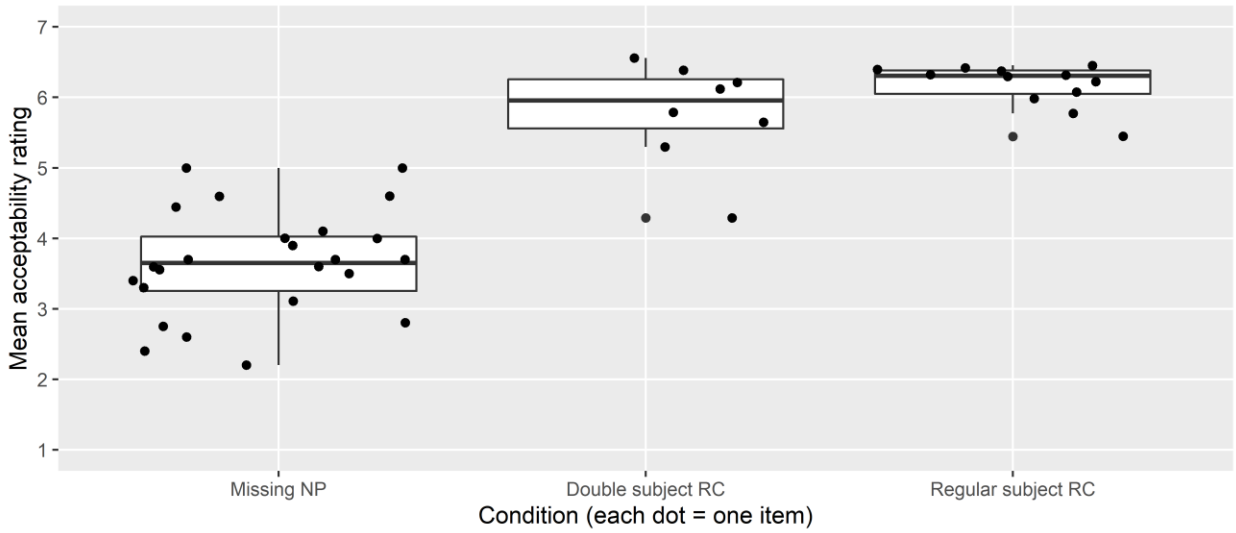

Figure 5: Distribution of ratings across items in plausible missing NP condition and selected grammatical fillers.

Aggregating across trials, double subject RC and regular subject RC conditions tend to have very high ratings. As Figure 6 shows, most trials have ratings of 6 or 7. Ratings of 1 or 2 were obtained for only 6% of trials for the double subject RC fillers and 2% of the trials for the regular subject RC fillers.

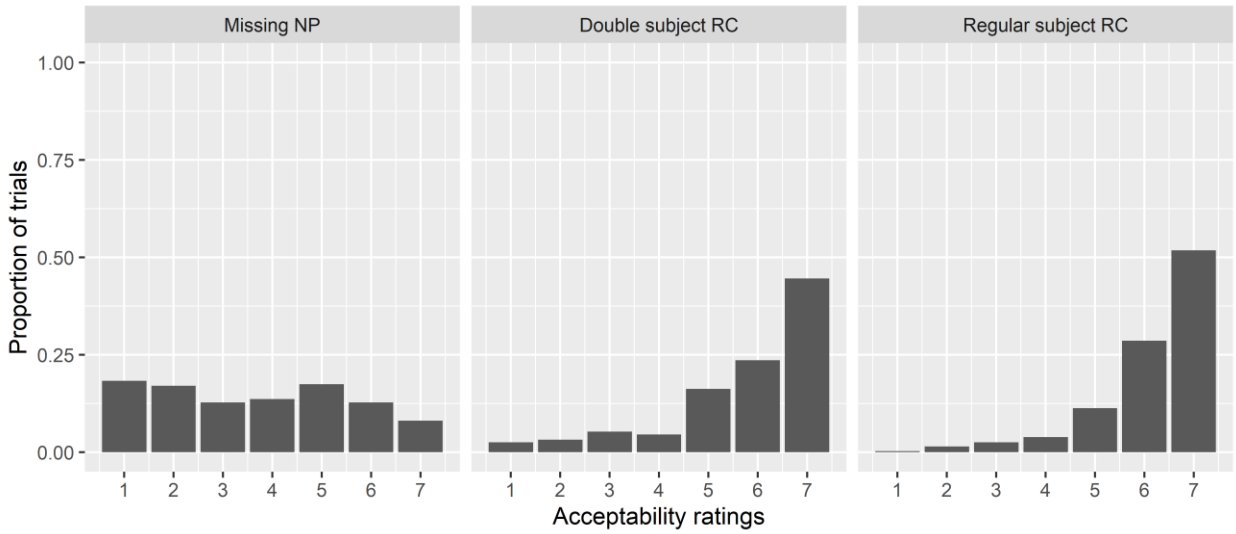

Figure 6: Distribution of ratings across trials in plausible missing NP condition and selected grammatical fillers.

Comparison across participants

One possible explanation for the relatively flat distribution of ratings for missing NP sentences is that there are two distinct populations of participants. For one population, missing NP sentences have grammatical parses (e.g. the parses in (ia) and (ib)) and so are consistently highly acceptable. For the other population, these sentences are ungrammatical, e.g. because these participants can only conjoin verbs with an overt *bìngqiě* and cannot stack relative clauses. For these participants, ratings for these sentences should be consistently low. In neither case is there an illusion.

In other words, this scenario predicts that there should be very few participants who give medium ratings and/or inconsistent ratings.

Figure 7 shows boxplots for each participant, sorting participants by the median rating they assign to missing NP sentences. While there are participants who give consistently high or low ratings, many participants are actually internally inconsistent, as reflected in the tall boxes and long whiskers, and/or give medium ratings to these sentences.

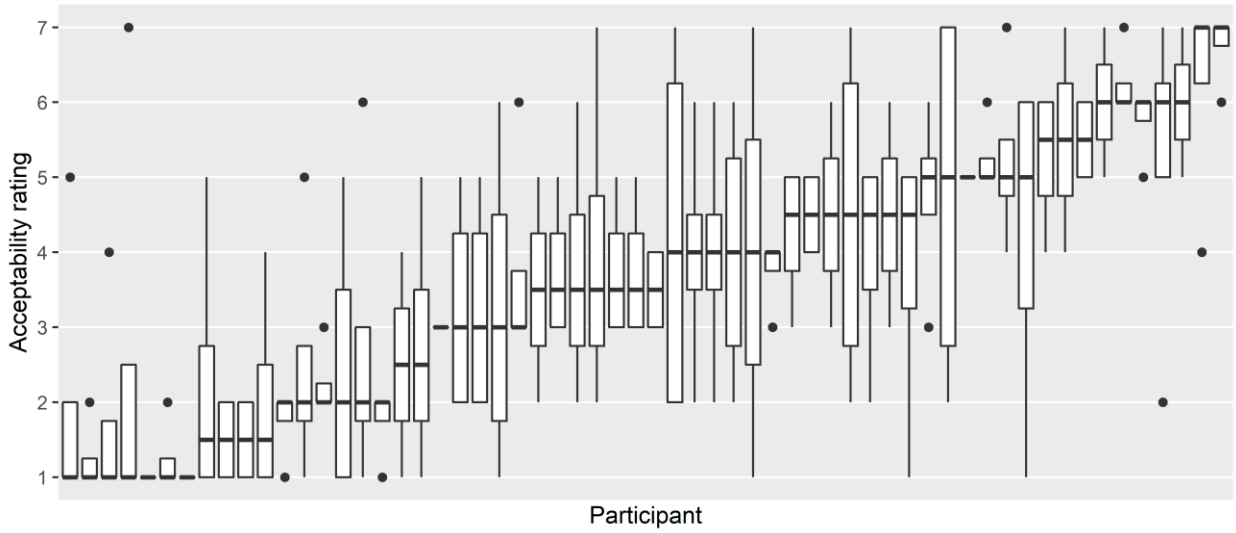

Figure 7: Distribution of ratings for plausible missing NP sentences, by participant.
(bold line in the middle of boxplot = median rating)

Figure 8 repeats the participant analysis for English missing VP sentences. Again, visually, there is no evidence suggesting that one group of participants find missing VP sentences to be uniformly acceptable and another group that find them to be uniformly unacceptable. As in the case of Chinese, many participants give inconsistent ratings and/or medium ratings to missing VP sentences.

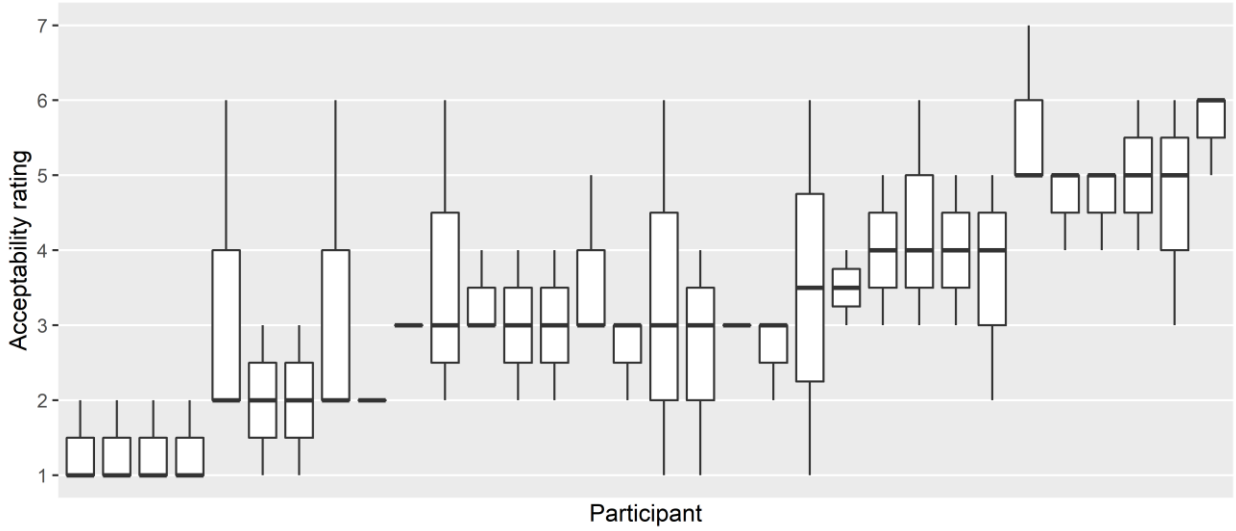

Figure 8: Distribution of ratings for English missing VP sentences, by participant.
(bold line in the middle of boxplot = median rating)

Comparison across missing NP items

In this section, we examine variability at the item level. Figure 9 shows that the 24 missing NP items vary in their mean acceptability ratings. If missing NP items have grammatical parses, this item-wise variability, as well as the participant-level variability seen in Figure 7, could be taken to mean that the grammatical parse is unevenly accessible to speakers. Perhaps, as suggested by the anonymous reviewer mentioned in introduction, the conjoining of verbs (or stacking of relative clauses) is only plausible in some but not all missing NP sentences, for instance, because of the lexical semantics of the verbs involved.

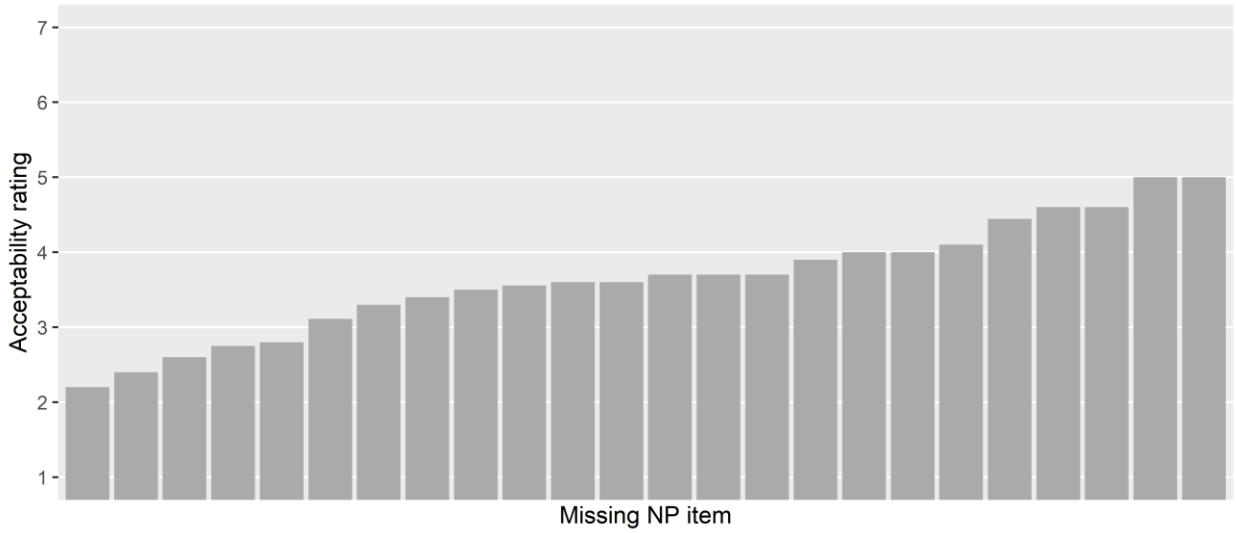

Figure 9: Mean acceptability ratings for plausible missing NP sentences, by item.

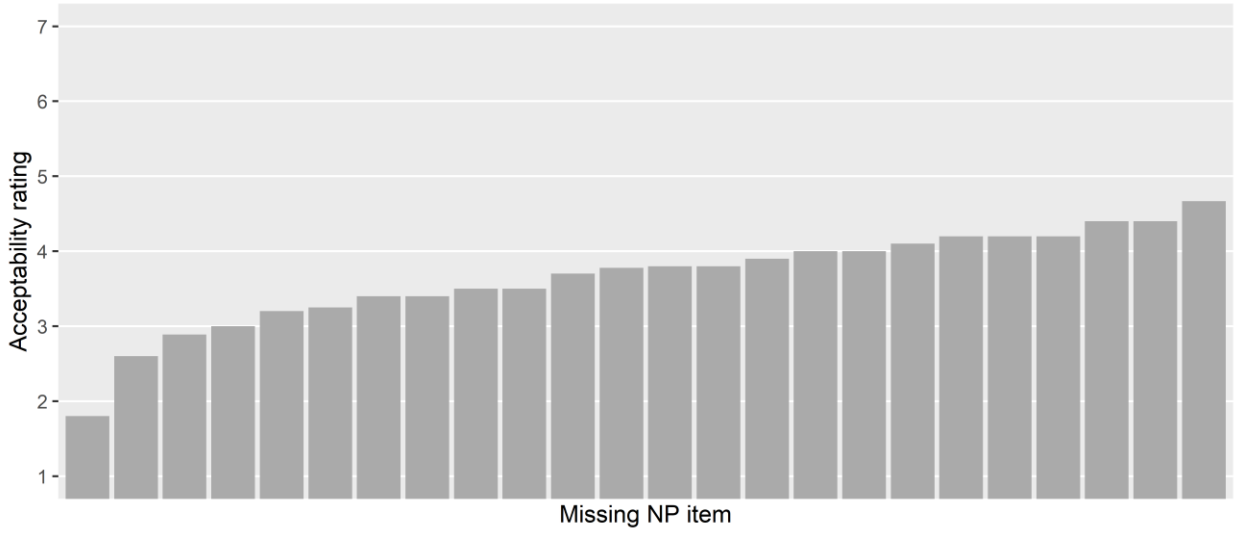

Figure 10: Simulated mean acceptability ratings for plausible missing NP sentences, by item.

If this view is right, we might expect the item-wise variability to be greater than chance. We operationalize the notion of "greater than chance" in the following manner. We create a second dataset of acceptability ratings by randomly shuffling around (i.e. sampling without replacement) the missing NP ratings in our original dataset, matching the number of ratings per item in the original dataset: for instance,

if item #1 had 10 ratings in the original dataset, item #1 would also have 10 (randomly-assigned) ratings in the new dataset.

Figure 10 shows mean acceptability for this new randomized dataset; the item-wise variability here illustrates what we should observe under chance alone. The item-wise variability seems comparable to that in Figure 9, i.e. the item-wise variability in our actual data (Figure 9) is probably no greater than chance.

Returning to our original dataset, we further observe that ratings can vary substantially within a given missing NP item, as the boxplots in Figure 11 show. This within-item variation suggests that these missing NP items are unlikely to have a grammatical parse: if they did, one might expect these items to have consistently high acceptability ratings.

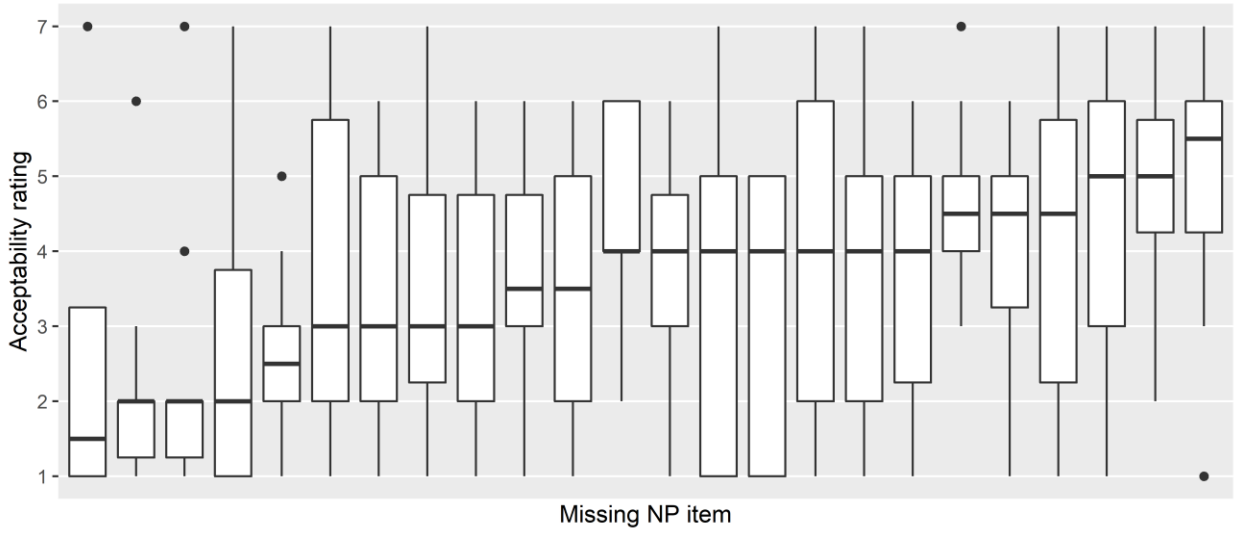

Figure 11: Distribution of ratings for (actual) plausible missing NP sentences, by item.
(bold line in middle of boxplot = median rating)

Repeating the above analyses for the 12 English missing VP items reveals similar distributional patterns. We show actual and simulated mean acceptability ratings at the item level in Figures 12 and 13. Visually, the item-wise variability in these figures resemble that observed for Mandarin. Figure 14, an analysis of actual ratings, shows that acceptability ratings often vary within a given missing VP item.

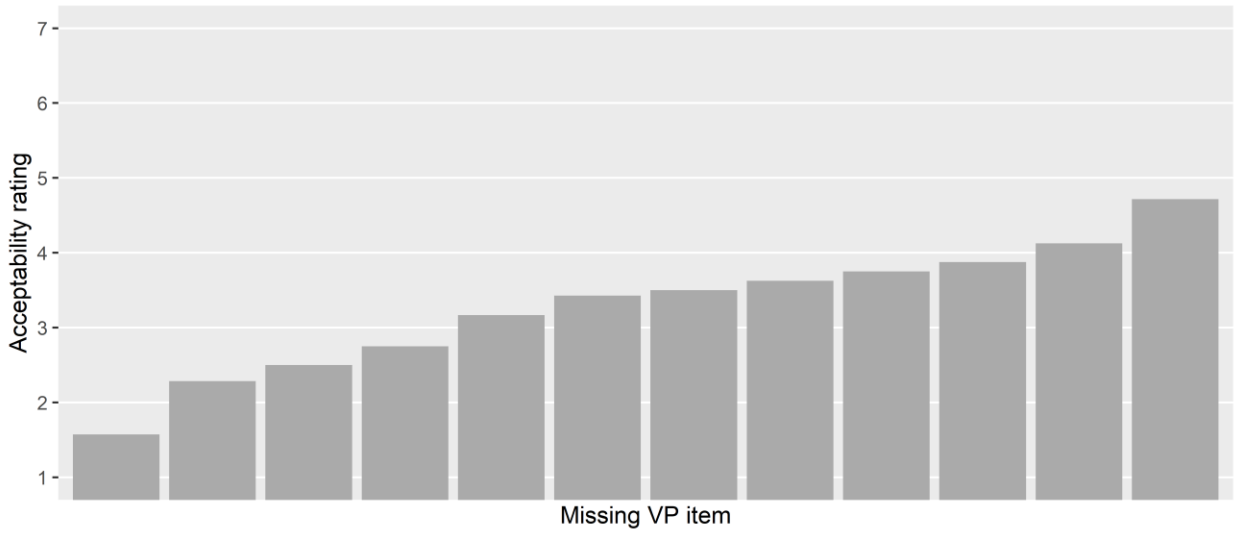

Figure 12: Mean acceptability ratings for English missing VP sentences, by item.

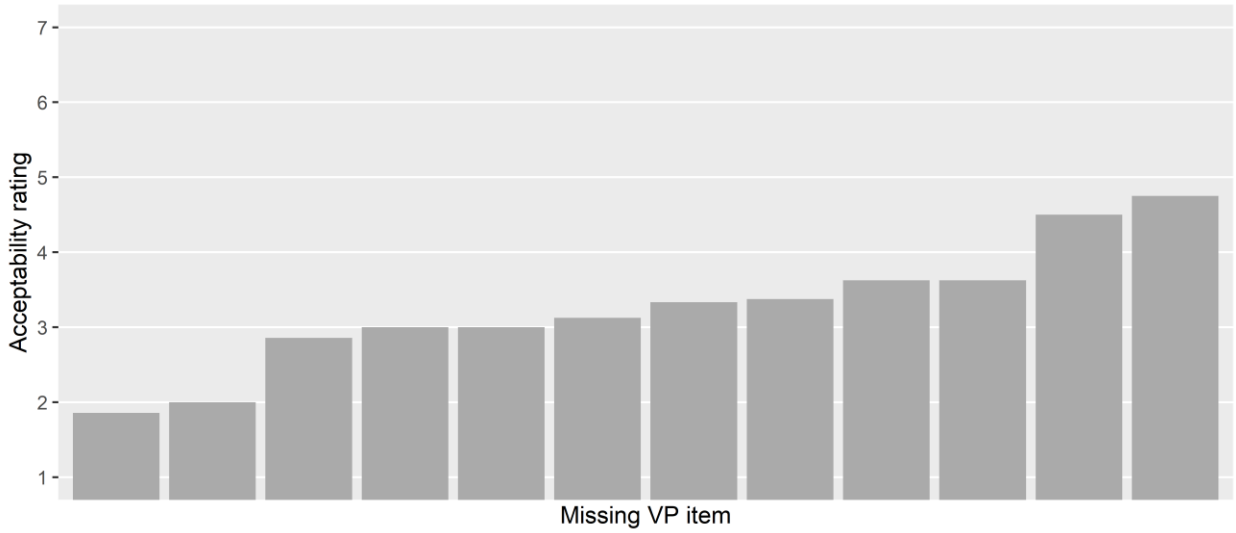

Figure 13: Simulated mean acceptability ratings for English missing VP sentences, by item

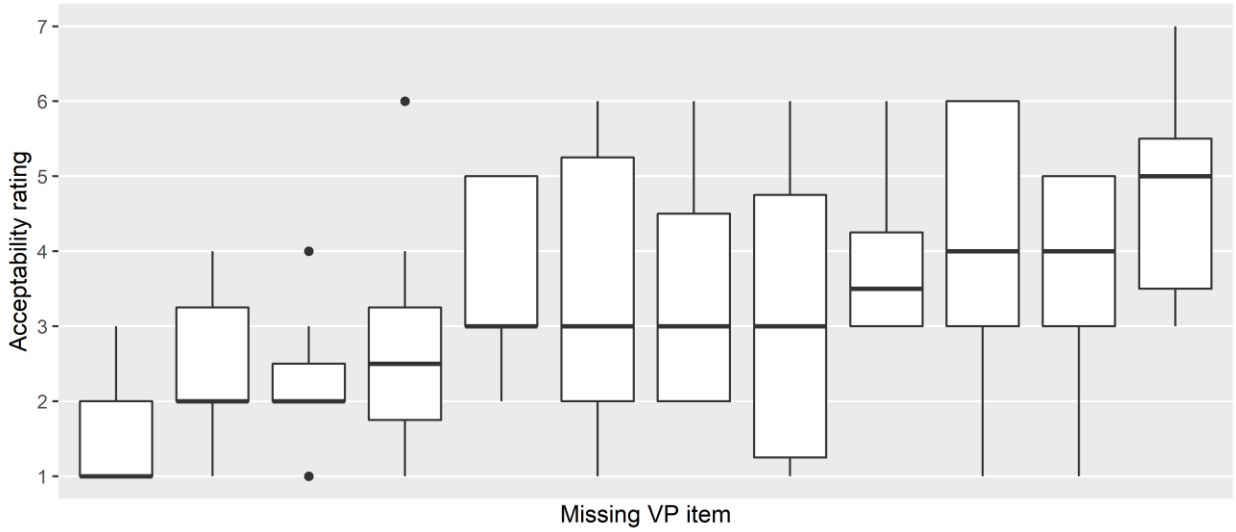

Figure 14: Ratings for (actual) plausible missing VP sentences, by item.
(bold line in middle of boxplot = median rating)

Conclusion

To summarize, we saw that missing NP sentences have very different acceptability profiles from structurally similar grammatical fillers. The acceptability profile cannot be attributed to the idea that there are two distinct populations of Mandarin speakers, one that finds missing NP sentences highly acceptable (grammatical) and one that finds them highly unacceptable (ungrammatical). In fact, participants are often inconsistent in their ratings for missing NP sentences.

We also observed item-wise variability for missing NP sentences – some missing NP sentences seem more acceptable than others. However, a simulation suggests that such a spread is no more than is expected by chance, given the overall variability in ratings. In addition, we saw within-item variability, which suggests that many missing NP items are unlikely to have a grammatical parse.

Importantly, we showed that the same generalizations also hold for missing VP sentences in English, which are uncontroversially ungrammatical.

These generalizations can be easily made sense of if missing NP sentences are ungrammatical but illusory. To the extent that the illusion is probabilistic in nature, it explains why speakers do not always give consistent ratings and why these items do not receive uniformly high ratings: in at least some trials, speakers can detect the ill-formedness of these sentences, and so assign them low ratings.

Treating missing NP sentences as ungrammatical and illusory would also provide a straightforward explanation for why missing NP sentences have acceptability profiles (at the item and participant levels) that strongly resemble missing VP sentences, whose illusory status is not in doubt. We further note that similar variability in ratings have been observed for other grammatical illusions, such as the comparative illusion, like *More people have been to Russia than I have* (Wellwood et al. 2018, data from preliminary experiments, available at https://github.com/alexiswellwood/compillu/blob/master/wphp-JoS-report-prelim-expts.pdf).

In short, Mandarin Chinese missing NP sentences have very different acceptability profiles when compared to grammatical fillers, whether at an item (sentence) level or at the trial level. Looking within missing NP sentences, we see variability in ratings at the participant or item level. Similar conclusions can be drawn for English missing VP sentences, which are widely agreed to be ungrammatical. These results overall suggest that Mandarin native speakers do not readily assign a grammatical parse (whether that involves conjunction or stacking) to missing NP sentences.

References

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