Journal of Semantics, 2017, 1–8
doi: 10.1093/jos/ffx001
Article

OXFORD

Post Hoc Analysis Decisions Drive the Reported Reading Time Effects in Hackl, Koster-Hale & Varvoutis (2012)

Edward Gibson
Department of Brain and Cognitive Sciences, MIT, e-mail: egibson@mit.edu

Steven T. Piantadosi
Department of Brain and Cognitive Sciences, University of Rochester

Roger Levy
Department of Brain and Cognitive Sciences, MIT

First version received 30 March 2014; Second version received 22 February 2015; Accepted 20 February 2017

Abstract

Hackl, Koster-Hale & Varvoutis (2012; hereafter HKV) provide data that suggest that in a null context, antecedent-contained deletion (ACD) relative clause structures modifying a quantified object noun phrase (NP) are easier to process than those modifying a definite object NP. HKV argue that this pattern of results supports a quantifier-raising (QR) analysis of both ACD structures and quantified NPs in object position: under the account they advocate, both ACD resolution and quantified NPs in object position require movement of the object NP to a higher syntactic position. The processing advantage for quantified object NPs in ACD is hypothesized to derive from the fact that—at the point where ACD resolution must take place—the quantified NP has already undergone QR, whereas this is not the case for definite NPs. Here, we question these conclusions. In particular, our analyses of HKV's reading time data reveal several unreported choice points, errors and concerns regarding multiple comparisons in the original HKV data analysis. Importantly, most other plausible ways to analyze these data that we describe here result in the crucial interaction being non-significant. Putting this observation together with the failure to observe the crucial interaction in Gibson & Levy (2016), we conclude that the experiments reported by HKV should not be viewed as providing evidence for the ACD quantifier-raising processing effect.

1 Introduction

In a recent attempt to find online evidence for the existence of quantifier raising (QR), Hackl, Koster-Hale & Varvoutis (2012; hereafter HKV) reported that antecedent-contained deletion (ACD) relative clauses modifying an object noun phrase (NP) whose determiner is *every* (e.g. *every doctor… was*, as in (1a)) were easier to process than those modifying an object NP whose determiner is *the* (e.g. *the doctor… was*). In contrast, in full verb controls (as in (1b)), *every* had no advantage over *the*.

(1) The understaffed general hospital was negotiating with *every/the* doctor…
  a. that the nonprofit medical organization *was*… [*was* condition]
  b. that the nonprofit medical organization *funded*… [*verb* condition]
… in order to arrange for free vaccination clinics.

HKV argued that this pattern of results supports a QR analysis of both ACD structures and quantified NPs in object position: under this account, both ACD resolution and quantified NPs in object position require movement of the object NP to a higher syntactic position. The advantage for quantified object NPs in ACD was hypothesized to arise from the fact that the quantified NP, but not the definite NP, has already undergone QR by the time ACD has to be resolved.

HKV reported results from one offline acceptability rating study and two self-paced reading studies. Gibson *et al.* (2014) showed that HKV's experimental design is confounded, with the consequence that their results could be explained by pragmatic and plausibility factors. In order to better understand the online data in HKV's self-paced reading experiments, we attempted to replicate these experiments. However, we did not find the crucial interactions in either of our experiments, as we report in Gibson & Levy (2016). We, therefore, obtained HKV's data in order to investigate how their data differed from ours.[1]

2 Data exclusion inconsistencies with what is reported in HKV Experiment 1

After having received useful comments from two reviewers of this note,[2] we found several inconsistencies with what was reported in the original HKV paper:

b. HKV misreported and misapplied the participant and item exclusions in Experiment 1.

HKV reported the following exclusion criteria in their paper for Experiment 1: 'We excluded participants who did not answer more than 75% of the total items (targets and fillers) and more than 70% of the target items correctly' (p. 168). In addition, HKV reported the following further details with their raw data: 'Data from 5 sentences (items 4d, 10d, 24f, 34d,f) were excluded due to programming or computer error. Data from items 26a, c, d and f are misaligned in the output files. To correctly align we added an empty pad

[1] HKV's data are available at http://hdl.handle.net/1721.1/76676. One additional participant data file (not in HKV's repository, but analyzed by HKV) is also available at http://tedlab.mit.edu/tedlab_website/researchpapers/30.dat.

[2] Although we did not have access to HKV's R analysis scripts, reviewer 2 (R2) did. We make some conclusions based on his/her observations of these scripts.

between words labeled 4 and 5 in the data files. Data from 5 additional items (24d, 30c,d, 33c, 39d) were excluded from items analysis because of accuracy rates below 75%.'[3]

There were several inconsistencies between the reported item and participant exclusions in Experiment 1 and what was actually done by HKV. First, the term 'target items' was not defined as the items in the Experiment, but rather as items in either the experiment in question or a second experiment that was run concurrently, but not reported in HKV. More critically, there were inconsistencies in item exclusions, such that it appears that HKV actually excluded all items below 25% accuracy (or 75% *error*) after removing the low-accuracy participants (although even this criterion does not recover the exclusions reported by HKV). In the analysis that we report, we use the following participant exclusions: 7, 17, 18, 21, 27, 45; and the following items: 4, 10, 24, 30, 33, 34, 39, 44. These are the exclusions that R2 reports. We find that these exclusions result in analyses that are most consistent with HKV's reported statistical patterns, although it remains unclear how exactly HKV decided to exclude this particular set of participants and items.

When we use these data exclusions, and we trim our data in the order that R2 suggests that HKV did,[4] our analyses are roughly consistent with the means and analyses that HKV report in their paper. The critical analysis for Experiment 1 is the 2 × 2 {the, every} × {verb, was} interaction at the third word following the disambiguating verb. This analysis [$F1(1, 43) = 6.73$, $P = 0.013$; $F2(1, 51) = 5.13$, $P = 0.028$] is similar to the F-tests reported in the HKV paper: $F1(1, 43) = 7.987$; $P < 0.017$) (p. 169); $F2(1, 51) = 4.40$; $P < 0.05$ (p. 189).[5]

We compared these analyses with those with more principled participant and item exclusions (First, it should be noted that if we examine the raw data, with no trimming whatsoever, the effects that HKV report are not significant). If we examine the comprehension question accuracies across HKV's participants in Experiment 1, we find a mean accuracy of 86%, including two poor-performing participants at 52.3% (participant 27) and 53.9% (participant 21). All other participants answered more than 73% of the questions correctly, with no further poor performing outliers. We, therefore, omit these two participants from our analyses. One item (item 39) was not run in one condition in HKV's experiment, so we also omit this item altogether.[6] Comprehension questions for four items—33, 44, 48 and

[3] These item numbers are the numbers in HKV's data analysis files, which are different from the numbers associated with the items in HKV's appendices. All item numbers that we use are the ones in HKV's analyses.

[4] According to R2, HKV trimmed their data in the following order: first, they computed residual RTs for each subject, using all items. Then they removed RTs that were two standard deviations above or below condition/region means by subject. Then they remove outlier subjects and items. This seems like an odd order of trimming: in particular, if certain items are to be omitted because they are outliers in some way, then it makes most sense to trim these items *first*, before computing the residual RTs. If one performs the trimming in this order, then we observe similar ANOVA results, but with a slightly higher P-value for the $F1$: $F1(1, 43) = 5.21$, $P = 0.0275$; $F2(1, 51) = 5.33$, $P = 0.0251$. This is relevant because HKV multiply the P-values by 3 in order to control for the number of regions that they analyzed. These analyses would no longer count as below 0.05 under such a residual RT computation, even ignoring all the other issues observed here.

[5] Following HKV, we analyze their data using participant and item ANOVAs. Using linear mixed effect regressions, we obtain similar patterns with respect to which analyses are significant v. not, at the $P < 0.05$ level. So, for brevity, we present only the ANOVA analyses.

[6] In a document that is supplied with the raw data, HKV state that 'Data from 5 sentences (items 4d, 10d, 24f, 34d,f) were excluded due to programming or computer error." But all the data for these

53—were answered at an average rate of below 52%, well below the next lowest accuracy item, at 66%. We, therefore, omit these items from our analyses. With these exclusions, the *F*-tests at the third word following the disambiguating verb for correctly answered trials are *F*1(1, 46) = 4.44, *P* = 0.041; *F*2(1, 53) = 2.71, *P* = 0.105, which do not satisfy the Bonferroni corrections that HKV apply to correct for the number of regions that they analyzed (section 3).

b. HKV's Experiment 1 included two further conditions whose results are not reported in the paper. Namely, in addition to the {the, every} × {verb, was} conditions below (from pp. 154–155) there was an extra level for the determiner factor (determiner 'a'), as follows:

The understaffed general hospital was negotiating with:
a. the doctor that the nonprofit medical organization funded...
b. the doctor that the nonprofit medical organization was...
c. every doctor that the nonprofit medical organization funded...
d. every doctor that the nonprofit medical organization was...
Extra 'a' unreported conditions:
e. a doctor that the nonprofit medical organization funded...
f. a doctor that the nonprofit medical organization was...

Because these extra conditions are not documented by HKV anywhere, we do not know what their hypotheses or intentions were with respect to them. However, we can still analyze them as HKV analyzed other similar conditions in their paper. When the indefinite condition is included, the 3 × 2, {the, every, a} × {verb, was}, interaction is not significant by either participants or items for the particular set of participant and item exclusions that most closely match HKV's (*F*1(2, 86) = 3.04, *P* = 0.053; *F*2(2, 102) = 2.01, *P* = 0.139). There is also no significant interaction when using the more principled participant and item exclusions discussed above (*F*1(2, 93) = 2.49, *P* = 0.089; *F*2(2, 107) = 1.48, *P* = 0.233).

c. HKV also excluded incorrect trials from their analyses,[7] but they did not document this in the paper or supplementary information. Whereas it is not uncommon to analyze correct trials only,[8] this decision is relevant here because the reported analyses with HKV's participant and item exclusions are not significant at the standard threshold when all trials are included: (*F*1(1, 43) = 3.90, *P* = 0.055; *F*2(1, 51) = 3.52, *P* = 0.066). Furthermore, there is no significant interaction using the more principled participant and item exclusions discussed above (*F*1(1, 46) = 1.86, *P* = 0.179; *F*2(1, 54) = 1.66, *P* = 0.203), and there is no significant 3 × 2 interaction when the extra (indefinite determiner) conditions described above are included [HKV's participant and item exclusions: *F*1(2, 93) = 1.59, *P* = 0.210; *F*2(2,

conditions appear to be present in their raw data files. In particular, the number of trials in these item–condition pairs matches the number of trials that were presented for other conditions in the same item. We, therefore, include these data in our analyses.

[7] R2 pointed this out to us.

[8] Although excluding trials with incorrectly answered questions may seem preferable so as to reduce noise from trials in which participants were reading inattentively, under many circumstances these exclusions can distort the interpretation of RTs in sentence reading. For example, if less-attentive reading of the critical region is likely to lead to incorrect question answers in one condition but not in another, then excluding incorrectly answered trials will inflate critical region RTs in the former condition even if the critical region is of equal processing difficulty and has the same overall distribution of RTs across conditions.

102) = 1.97, $P = 0.145$; our more principled participant and item exclusions: $F1(2, 93) = 1.23$, $P = 0.296$; $F2(2, 116) = 1.06$, $P = 0.351$].

3 *Post hoc* choices in selecting the regions to analyze

More importantly, setting the issues with the data exclusion aside, the regions that HKV analyzed appear to have been chosen *post hoc*. The analyses are, therefore, liable to have inflated false positive rates. In particular, HKV picked the analysis region in Experiment 1 based on visually inspection, selecting a region with the biggest difference in the predicted direction. For example, on p. 169 they said: ‘On both P3 and P4, visual inspection reveals that rRTs for sentences that have both a definite article and an ACD site (the-was) are higher than those for the other three conditions.’ And in Experiment 2 (p. 178), they said the following: ‘Visual inspection of the region we are interested in, the region that starts with the V/Aux site, reveals a complex crossing pattern over the first two words of that region followed by a prominent separation of rRTs that persists over the next two words before rRTs collapse again to comparable levels across all conditions.’

Using visual inspection to select the region with a large predicted difference and then statistically analyzing that difference can dramatically increase the false positive rate [e.g. see Kriegeskorteet al.(2009) and Simmons et al. (2011)]. For example, imagine that one was testing whether a coin was fair by flipping the coin 1000 times. If one found the longest sequence of consecutive heads and analyzed only trials in and around that sequence, one would find a ‘significant’ difference from 50% chance. However, this would not of course mean that the coin is unfair.

For Experiment 1, HKV note that they need to do a Bonferroni correction on the *P*-value, because they did not predict the effect at the region that they analyzed:

> p. 169: ‘A repeated measures ANOVA reveals a Determiner by Ellipsis interaction on P3, which remains significant after correcting for multiple comparisons, ($F(1, 43) = 7.987$; $P < 0.017$).’ Footnote 30: ‘We apply a Bonferroni correction of three comparisons given that the earliest point at which an effect of ellipsis can be expected is on P1, which is the first word that provides conclusive evidence for the presence of an ACD site in our items.’

HKV argue that a Bonferroni correction of 3 is needed. However, they analyzed each of four words (not three): the four regions after the V/AUX position (ignoring the verb region itself, where they argue that they did not expect to find the predicted effects, which is questionable). So they should be minimally dividing their *P*-values by 4, not 3. This is the *minimal* Bonferroni correction that should be applied. In fact, in addition to analyzing each of four one-word regions following the verb, HKV also analyzed two-word regions, which potentially increases the number of comparisons further. Had they applied this minimal correction, neither the *F*1 nor *F*2 ANOVA *P*-values would remain significant, even adopting their particular data analysis procedures, as described above. Of course, these *P*-values are even higher (i) when data exclusion is more principled; (ii) when the three-way interaction is investigated; and (iii) when incorrect trials are not omitted.

In any case, we can visualize HKV’s data from Experiment 1 to see if the distribution is consistent with their theory’s predictions. HKV’s QR hypothesis predicts that the interaction effect should be shifted positive away from zero. We can investigate this by plotting the average size of the interaction effect for each participant, and examining this distribution. The interaction effect was calculated by computing the difference of the differences

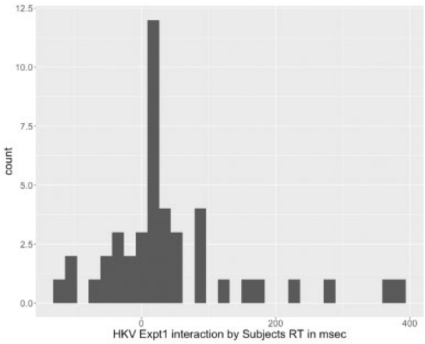

Figure 1 Histogram of the interaction effect for each participant in HKV's Experiment 1, using HKV's data trimming procedures (e.g. with data from trials answered correctly only).

associated with each quantifier (every, the) for each participant. This histogram is presented in Figure 1.

We can see from the histogram in Figure 1 that the RTs are centered right around 0 ms, with two outlier RTs, corresponding to subjects 26 and 32. This distribution looks strikingly similar to the results from Gibson & Levy's (2016) attempted replication, except for the outliers. Even removing only the most extreme of these (subject 32) causes the effect to be non-significant. Thus, if HKV intended that the QR hypothesis would apply to most participants, and not to a small subset of participants who would drive the effect, then the distribution of RTs does not seem consistent with this hypothesis, even using their own trimming procedures for their own data.

4 Similar inconsistencies in Experiment 2

For Experiment 2, HKV analyze a different region than the region they analyze in Experiment 1, focusing on the region two words after the disambiguating word instead of three words after. Experiment 1 could, and should, have been used as a way to 'lock in' the region where an effect is expected and to avoid having to examine multiple regions. By observing the effect in a different region, HKV inflate the expected false positive rate. If we analyze the same word position in Experiment 2 as HKV analyzed in Experiment 1, we see that the crucial 2 × 2 interaction between determiner {the, every} and verb {verb, did} is not significant ($F1(1, 47) = 1.22$; $P = 0.276$; $F2(1, 59) = 0.337$, $P = 0.564$).

Instead, HKV focus on the region two words after the disambiguating word: 'Specifically, a repeated measures ANOVA with Determiner (the, every) and Ellipsis (No-Ellipsis, Small-Ellipsis) reveals a significant interaction such that rRTs for *the* increase across the two Ellipsis levels, while the ones for *every* do not, ($F1(1, 47) = 4.520$; $P < 0.05$)' (p. 180).

We successfully replicated this analysis: $F1(1, 47) = 4.53$, $P = 0.039$; $F2(1, 59) = 4.83$, $P = 0.032$. However, these results do not survive Bonferroni correction for the number of regions analyzed. In particular, the minimal Bonferroni correction is 3, one for each word position following the relative clause verb. And, as with Experiment 1, the results of Experiment 2 are not robust to the inclusion of incorrect trials ($F1(1, 47) = 2.4$; $P = 0.13$; $F2(1, 59) = 3.48$, $P = 0.07$).

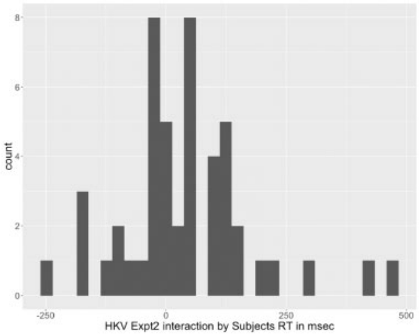

Figure 2 Histogram of the determiner {the, every} × verb {verb, did} interaction effect for each participant in HKV's Experiment 2, using HKV's data trimming procedures (e.g. with data from trials answered correctly only).

As for Experiment 1, we can visualize HKV's data in Experiment 2. Here, we focus on the crucial 2 × 2 interaction between determiner {the, every} and verb {verb, did} by plotting the average size of the interaction effect for each participant, and examining this distribution. The interaction effect was calculated by computing the difference of the differences associated with each quantifier (every, the) for each participant. This histogram is presented in Figure 2. As in Experiment 1, HKV's QR hypothesis predicts that the interaction effect should be shifted positive away from zero.

As for Experiment 1, we can see from the histogram in Figure 2 that the RTs are centered around 0 ms, with two outlier RTs, corresponding to subjects 3 and 10. Removing either one of causes the effect to be non-significant. Thus, similar to the results from Experiment 1, if HKV intended that the QR hypothesis would apply to most participants, and not to a small subset of participants who would drive the effect, then the distribution of RTs does not seem consistent with this hypothesis, even using their own trimming procedures for their own data.

Conclusions

In general, it is probably impossible to do away with *all* the choice points in a self-paced reading (or any other) experiment. For example, it may be difficult to *a priori* predict *exactly* where an effect will emerge in the RT record. In HKV's design, the effect might have predicted to occur exactly at the relative clause verb, or a word or two later, or perhaps even further downstream. In this particular pair of experiments, however, there were many choice points, including (i) which word to analyze: the relative clause verb, or one, two or three words following this verb; (ii) whether or not to report analyses of additional conditions that were run; (iii) which data from which participants and items to include; and (iv) whether to analyze all trials, or just those whose questions were answered correctly by participants. Perhaps most importantly, there is no analysis that produces a significant interaction in both experiments if one analyzes the same region across experiments: HKV analyze the word three words after the relative clause verb in Experiment 1, but they analyze the word two words after the relative clause verb in Experiment 2. Furthermore, analyses with all trials—not just trials with correct answers to comprehension questions—or analyses with the additional conditions that were run in Experiment 1—result in the crucial interaction being non-significant. Putting these observations together with the failure to observe the crucial interaction in Gibson & Levy (2016), we conclude that the experiments reported by HKV should not be viewed as providing evidence for the ACD QR processing effect.

Download notes

An analysis file for the analyses reported here is available at https://osf.io/z5s2w/.

Acknowledgements

We thank Leon Bergen, Ev Fedorenko, Shravan Vasishth and three anonymous reviewers from the *Journal of Semantics* for comments on earlier drafts of this work. We especially thank Kyle Mahowald for help with many of these analyses. We also owe a large thank you to one reviewer, who worked out important differences between HKV's stated analyses and what they actually did. R.L. was supported by a fellowship at the Center for Advanced Study in the Behavioral Sciences.

References

Gibson, E., P. Jacobson, P. Graff, K. Mahowald, E. Fedorenko, & S. T. Piantadosi (2014), 'A pragmatic account of complexity in definite antecedent-contained-deletion relative clauses'. *Journal of Semantics* 32:579–618.

Gibson, E., & R. Levy (2016), An attempted replication of Hackl, Koster-Hale, Varvoutis (2012). Cornell ArXiv: arXiv:1605.00178 [q-bio.NC]; R-code and data available at https://osf.io/t6anw.

Hackl, Martin, Jorie Koster-Hale, & Jason Varvoutis (2012), 'Quantification and ACD: evidence from real time sentence processing'. *Journal of Semantics* 29:145–206.

Kriegeskorte, N., W. K. Simmons, P. S. Bellgowan, & C. I. Baker, (2009), 'Circular analysis in systems neuroscience: the dangers of double dipping'. *Nature Neuroscience* 12:535–540.

Simmons, J. P., L. D. Nelson, & U. Simonsohn (2011), 'False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant'. *Psychological Science* 22:1359–1366.