Concerns with Nieuwland et al. (2017)
Katherine A. DeLong, Thomas P. Urbach & Marta Kutas
April 2017

Nieuwland et al. (2017, http://biorxiv.org/content/early/2017/02/25/111807) recently made available their (thus far to our knowledge) unpublished manuscript, describing a nine-lab replication attempt of DeLong, Urbach and Kutas (2005), hereafter DUK05. They concluded that over their nine labs, the *a*/*an* Article prediction effect reported in DUK05 did not replicate. Here, we highlight some features of their project that undermine confidence in the purported non-replication of this relatively small language ERP effect. Our central concerns involve a number of discrepant ERP and behavioral findings between the original DUK05 report and the new nine-lab study, as well as discrepancies among the nine labs themselves.

ERP data

A widely attested and robust finding (first reported in Kutas & Hillyard, 1984) is the broadly distributed pattern of high correlations between Noun N400 amplitude and Noun cloze probability (≈*r*= .9) over the scalp. DUK05 replicated these correlations with a maximal *r*-value of .84 and significant correlations at half the scalp electrodes (13 of 26). In DUK05 this graded N400 *at* the Noun was a key premise in arguing that the similarly graded N400 at the Article *before* the Noun was evidence of Noun pre-activation. On inspection of Nieuwland et al., it appears that only three of the nine labs (2=Bristol, 7=Oxford, 9=York) observe Noun correlations on par with previous research. Three other labs (3=Edinburgh, 4=Glasgow, 6=London) find weaker and/or less widely distributed Noun correlations than DUK05 (e.g., London showed weaker albeit significant correlations at just 2 channels, and Edinburgh exhibited a maximum Noun correlation *r*-value of .73). Moreover, the remaining three labs (1=Birmingham, 5=Kent, 8=Stirling) find *no* significant Noun-cloze correlations at *any* channels. In short, for whatever reasons, the majority of the labs in this cohort—six of nine—find less robust to no Noun N400-cloze correlation effects. Given the failure to find a robust N400-cloze correlation at the Noun, logically one would not necessarily expect to find an N400 effect at the preceding Article. The real question is why the majority of labs failed to find the Noun N400-cloze correlation that has been demonstrated in numerous labs over almost four decades. We return to this below.

It is also instructive to examine the labs that *did* observe robust Noun N400-cloze correlations. Two of these three labs (7=Oxford and 9=York) show Article ERPs that pattern in the expected direction (low cloze Articles with greater N400 negativity than high cloze Articles). This pattern is evident not only at single channel Cz (plotted in Nieuwland et al. Figure 2) but also in the Article whole-head ERP plots for these labs, available in Nieuwland et al.'s supplementary materials. The third lab that showed clear Noun N400-cloze correlations (2=Bristol) also stands out in Figure 1 of Nieuwland et al. as exhibiting significant Article correlations—i.e., a potential replication although the authors argue that the scalp topography of the correlations prohibits such a conclusion. Taken together, the three labs that observe the most compelling Noun N400-cloze correlations also seem to exhibit Article N400-cloze patterns most like those reported in DUK05.

Stimuli/cloze probability norms

Overall, mean cloze probability norms for both the Articles and Nouns in Nieuwland et al. are somewhat lower than those for DUK05 (by 7% and 6% respectively), and this may have contributed to the lower correlation values Nieuwland et al. report. One potential factor may be that the DUK05 stimuli were developed and cloze-normed nearly 15 years ago, in 2002, and life/society has changed considerably during that time period. We also have concerns about using stimulus materials across countries and cultures (e.g., differences between San Diego/U.S. and the U.K.), even though an attempt was made by Nieuwland et al. to

adapt a few of the stimulus materials for British participants. In addition, Nieuwland et al. report that their stimulus cloze norms were not collected at each of the individual lab locations of ERP testing, but rather in Edinburgh only. Based on our experience, gathering cloze norming data in the local population is important. We point out that the concerns raised here are not ones limited to Nieuwland et al.'s use of our stimulus materials, but rather align with our more general viewpoint (relayed repeatedly to others, as well, over the years) that replications are best conducted with materials that meet certain design criteria and which are suitable for locality and time of testing.

Question accuracy

Comprehension questions are often included in sentence reading experiments as a check that participants are on task (DUK05, mean accuracy 94%). Although overall comprehension accuracy may vary between studies with different sets of questions, it is worth noting that in Nieuwland et al., four of the nine labs (Labs 1, 3, 5, and 8) report mean comprehension question accuracies that are not just lower than DUK05 but have accuracy range minima that are alarmingly low (65%, 47%, 50% and 43%, respectively). These comprehension scores indicate that some number of Nieuwland et al.'s participants were likely not engaged in and/or were unable to perform the comprehension task, in which case their ERP data cannot be trusted as evidence for or against the replication.

Filler items

Whereas fillers comprised over half the stimuli in DUK05, filler items were not used in the attempted direct replication, presumably because description of the fillers was inadvertently omitted from the original report. This discrepancy could have been avoided if we had been apprised of the experimental procedures and aims of direct replication before the experiments were conducted. We cannot know what role, if any, fillers played in the elicitation of prediction effects in the *a*/*an* studies. Ito, Martin & Nieuwland (2016), however, offered arguments on how the filler items in their attempted replication of Martin et al. (2013) may have contributed to their *not* observing the *a*/*an* prediction effect. In our rebuttal to Ito et al. (2016)—DeLong, Urbach & Kutas (2017)—we debunked the idea that absence of fillers could explain DUK05's observed Article effects.

Number of participants

A final minor point is that we wonder why the number of participants tested in some of the nine labs differs not only from the original DUK05 study (N of 32), but why the nine labs differ from each other. Also, several of the Nieuwland et al. nine labs have odd numbers of participants, which indicates imbalanced stimulus lists.

Conclusion

The concerns raised here—relating to the interpretation of null results, stimulus norming, comprehension accuracy, use of filler items, stimulus lists, and participant N's—are general ones that apply to a wide variety of language ERP studies, including replications. To be clear: We firmly believe that replication studies are important in science. It is also important that those replications be done properly.

References

DeLong, K.A., Urbach, T.P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), pp. 1117-1121.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, 1-8. doi: http://dx.doi.org/10.1080/23273798.2017.1279339.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, 1-12. doi: http://dx.doi.org/10.1080/23273798.2016.1242761.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials reflect word expectancy and semantic association during reading. *Nature*, 307: 161-163.

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, 69(4), 574–588. doi:10.1016/j.jml.2013.08.001.

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S.V.G., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D., Rousselet, G., Ferguson, H., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D., Kohút, Z., Rueschemeyer, S., Huettig, F. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. bioRxiv, 111807; doi: https://doi.org/10.1101/111807.