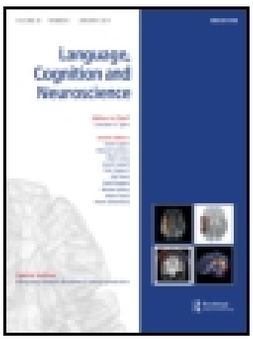

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: http://www.tandfonline.com/loi/plcp21

Why the A/AN prediction effect may be hard to replicate: a rebuttal to Delong, Urbach, and Kutas (2017)

Aine Ito, Andrea E. Martin & Mante S. Nieuwland

To cite this article: Aine Ito, Andrea E. Martin & Mante S. Nieuwland (2017): Why the A/AN prediction effect may be hard to replicate: a rebuttal to Delong, Urbach, and Kutas (2017), Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2017.1323112

To link to this article: http://dx.doi.org/10.1080/23273798.2017.1323112

Published online: 09 May 2017.

LANGUAGE, COGNITION AND NEUROSCIENCE, 2017
https://doi.org/10.1080/23273798.2017.1323112

Routledge
Taylor & Francis Group

Why the A/AN prediction effect may be hard to replicate: a rebuttal to Delong, Urbach, and Kutas (2017)

Aine Ito [a], Andrea E. Martin[b,c] and Mante S. Nieuwland[b]

[a]Faculty of Linguistics, Philology & Phonetics, Clarendon Institute, University of Oxford, Oxford, UK; [b]Max-Planck Institute for Psycholinguistics, Nijmegen, Netherlands; [c]Department of Psychology, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK

ARTICLE HISTORY Received 16 March 2017; Accepted 21 April 2017

In our recent publication "How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects" (Ito, Martin, & Nieuwland, 2016a), we report two experiments which failed to replicate existing event-related potential (ERP) evidence for prediction as reported in C. D. Martin et al. (2013), whose study resembled DeLong, Urbach, and Kutas (2005; from hereon DUK05). DeLong, Urbach, and Kutas (2017; from hereon, DUK17) recently published a commentary which depicts our publication as a case of poor scholarship that makes "no substantive contribution to the literature on what factors may matter for prediction and when." DUK17 warn that the readers of our work "will be led into serious error." In this rebuttal, we first present evidence that is inconsistent with the arguments of DUK17 regarding our own experiments, and then we briefly discuss other indications why it might be hard to observe and thus replicate the A/AN prediction effect.

To address DUK17's arguments regarding our own study, (1) we discuss why the observed null noun effect in non-native speakers in our Experiment 1 is in fact not a failure to replicate, in contrast to what DUK17 state, and we discuss how DUK17 conflate noun effects stemming from plausibility and semantic congruence in their discussion of bilingual ERP studies, (2) we report a linear mixed-effects model analysis on both of our experiments that fails to replicate the graded effect of cloze probability on article-elicited ERPs as observed in DUK05, and (3) we report the results of Bayesian analyses showing evidence in favour of the null hypothesis that article-cloze has no effect on article-elicited ERPs. Then, (4) we turn to the replicability of the landmark findings reported by DUK05, reviewing previous attempts to replicate their findings, and (5) end with a very brief discussion of the relevance of prediction effects reported in other articles.

We emphasise that we believe that prediction could play an important role in language comprehension, and we also do not see prediction as an all-or-nothing phenomenon. In concord with DUK17, we think that previously reported ERP effects elicited by articles marked for gender agreement or animacy provide evidence that favours prediction accounts over integration accounts. But the focus of our 2016 publication, and of our points in this rebuttal, is replicability of the prediction effects reported for the English indeterminate articles a/an, which abide by phonotactic but not agreement rules, and which have been argued to be evidence of the pre-activation of phonological information during reading and language processing more generally.

We also wish to state why we performed our experiments in the way that we did. The experiments were a replication attempt of the C. D. Martin et al. study, not that of DUK05. The items were presented as filler materials for another study on prediction (Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Ito, Martin, & Nieuwland, 2016b), which used isolated sentences that often contained a semantic anomaly. To have a more uniform set of items in the complete experiment, we changed the two-sentence items of the C. D. Martin et al. study into single-sentence items, and re-normed them. We first ran the experiment using an stimulus-onset asynchrony (SOA) of 500 ms because that is the SOA that C. D. Martin et al. described in their Methods section. Later, we discovered that this SOA was incorrectly reported and was, in fact, 700 ms. We therefore repeated the experiment with the longer SOA of 700 ms. This was very important, because non-native participants tend to read more slowly than native

CONTACT Aine Ito aine.ito@ling-phil.ox.ac.uk

participants (Hopp, 2009), and predictive processing could be more likely to occur, or be observable, at slower rather than at faster presentation rates. While DUK17 strongly focus on our Experiment 1, Experiment 2 was a more direct and therefore more relevant replication attempt of the Martin et al. study than Experiment 1 was, because it used the SOA that Martin et al. had actually used rather than the one Martin et al. originally reported.

1. N400 effects of noun-expectedness in non-native speakers: no failure to replicate

Our non-native participants did not show a statistically significant noun-elicited N400 effect of expectancy in Experiment 1 (500 ms SOA), but did so in Experiment 2 (700 ms SOA). DUK17 argue that, if we take the article null-effect to question the robustness of A/AN effect, we should have considered this noun null-effect from Experiment 1, likewise. DUK17 question why we did not find an N400 effect for unexpected nouns relative to expected nouns at 500 ms SOA in non-native speakers, even though N400 effects in non-native speakers have been observed using a similar SOA (Ardal, Donald, Meuter, Muldrew, & Luce, 1990; Moreno & Kutas, 2005; Weber-Fox & Neville, 1996). DUK17 thus argue that the lack of the noun effect undermines our conclusion about the article effect, and that it suggests a more general problem with our data. However, DUK17 do not acknowledge that our study differs from these studies in a critical aspect that can explain the apparent inconsistency: our comparison was between expected and unexpected words that were both *highly plausible*. All the bilingual studies cited directly by DUK17, and *all* studies mentioned in the reviews cited by DUK17, observed effects of congruity (i.e. compared semantically implausible or anomalous words to semantically plausible or non-anomalous words). Non-native speakers may be sensitive to the much stronger manipulation of semantic anomaly even when reading at 500 ms SOA, but less sensitive to subtle effects of predictability at this SOA. To our knowledge, N400 effects of expectancy in non-native speakers have not been reported in reading studies using a 500 ms SOA, only in those using a 700 ms SOA (Foucart, Martin, Moreno, & Costa, 2014; Martin et al., 2013). In our own study, we found no effects at the articles or nouns in non-native speakers with the faster 500 ms SOA, whereas we found a noun effect at the slower 700 ms SOA but no article effect. We emphasise that the noun effect we observed at the slower SOA replicates previous work, but the absence of the noun effect at the faster SOA is not a failure to replicate, because an expectancy effect in non-native speakers has not been tested at that SOA. It is still possible that our 500 ms SOA data did not have power to detect noun-N400 effects, but this would need to be supported by further evidence showing that non-native noun effects are robust at 500 ms SOA, which to our knowledge, does not exist in the first place. Given these important differences, we see no inconsistency in our argumentation[1].

If ERPs elicited by the articles are to be considered uninformative about prediction given the lack of a noun-N400 effect, then perhaps an analysis of the article results at that SOA should only include the native speakers. In our paper, we did not perform this analysis as it required following up on a non-significant group by expectancy interaction. However, given that DUK17 emphasise the marginally significant effect for the groups combined, whereas much of their discussion is actually about native speaker comprehension, we here report the analysis for native speakers alone for the interested reader. In the native group, for the three region of interest (ROIs) combined, the effect of expectancy was neither statistically nor marginally significant, $F(1, 22) = 1.5$, $p = .23$.

2. A double failure to replicate the pre-activation gradient

In Ito, Martin, et al. (2016a), we replicated the analysis protocol reported by Martin et al. (2013), which included categorising articles into high- and low-cloze articles and comparing the ERPs elicited by these categories. This analysis differs from DUK05, who argued for probabilistic pre-activation based on the correlation in which pre-nominal articles elicited gradually smaller N400s as cloze probability increased. DUK17 argue that our categorical analysis was not sensitive enough to pick up the graded relationship between cloze and ERP activity.

To address this issue, we performed a linear mixed-effects model analysis (Baayen, Davidson, & Bates, 2008) on single-trial data with cloze probability as a continuous variable (see also Ito, Martin, et al., 2016b; Nieuwland, 2016). The full R code and the data sets we used for this analysis are available on the Open Science Framework at osf.io/ttgj2. This analysis models variance at the level of the subject and the item and therefore yields a better estimate of a graded underlying effect than a factorial ANOVA or the correlation approach used by DUK05 (see also Section 4). We ran a model that included random intercepts by subject and by item and random slopes for cloze probability by participants and by items (R lmer-syntax: N400 ~ cloze + (cloze | subject) + (cloze | item)). The variable cloze probability was centred. This model was first run on a

dataset including data from both SOAs (500 and 700 ms) and all three ROIs (frontal, central, and parietal), and then for separate SOAs and ROIs. The data only included native speakers for a more direct comparison with DUK05. This model did not reveal a significant effect of cloze, $\beta = .28$, $SE = .23$, $t = 1.2$, and inclusion of cloze did not improve the model fit compared to the model without cloze, $\chi^2(1) = 1.5$, $p = .2$. The summary of the model run on subsets of data for each SOA and ROI is presented in Table 1 and Figure 1. While there is a small graded trend at the 500 ms data, there is virtually no evidence of such a trend at the 700 ms SOA where prediction is more likely to occur (Ito, Corley, et al., 2016). None of these models found a statistically significant effect of cloze. In both experiments, we failed to replicate the statistically significant pre-activation gradient at the articles.

3. Support for the null hypothesis: a Bayesian analysis

Like most research in psycholinguistics, our studies involved null hypothesis significance testing. However, statistical non-significance cannot demonstrate that the null hypothesis is true, in fact, *p*-values from canonical ANOVAs, *t*-tests, or correlations can never allow acceptance of the null hypothesis (e.g. Masson, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007). To overcome this hurdle in interpretation, we performed Bayesian analyses to quantify the obtained evidence in support of the null hypothesis that cloze had no effect on article-elicited ERPs.

Our first analysis computed the Bayes factor and Bayesian posterior probabilities for the null hypothesis following Masson (2011). This Bayesian approach evaluates evidence for the null hypothesis compared to the alternative hypothesis that there is an effect of cloze on article-elicited ERPs. We first ran a one-way ANOVA testing effects of expectedness (expected *vs.* unexpected) on the dataset that contained both SOAs and all three ROIs. The ANOVA did not show a significant effect of expectedness, $F(1, 45) = .1$, $p = .7$. The evidence for the null hypothesis was positive, $p(H_0|D) = .87$, according to Raftery's (1995) classification.

Table 1. Summary of linear mixed-effects models and model comparisons for each SOA and each ROI.

SOA (ms)	ROI	β	SE	t	Model comparison
500	Frontal	.44	.36	1.2	$\chi^2(1) = 1.5$, $p = .2$
	Central	.41	.36	1.1	$\chi^2(1) = 1.3$, $p = .3$
	Parietal	.35	.35	1.0	$\chi^2(1) = 1.0$, $p = .3$
700	Frontal	.12	.36	.3	$\chi^2(1) = .1$, $p = .7$
	Central	.17	.35	.5	$\chi^2(1) = .2$, $p = .6$
	Parietal	.20	.33	.6	$\chi^2(1) = .4$, $p = .5$

We additionally performed a Bayes factor replication test (Verhagen & Wagenmakers, 2014) to compare our results more directly to those of Martin et al. We should emphasise, however, that this comparison must be interpreted with great caution because the electrode reference differed between studies (as did many other variables, as we described in our previous paper). This analysis computes the replication Bayes factor BF_{r0} to evaluate evidence for the null hypothesis H_0 compared to the alternative replication-hypothesis H_r that cloze impacts article-ERPs with the strength of effect reported by Martin et al. (Boekel et al., 2015; Jeffreys, 1961). Based on the ANOVA *F*-values reported by Martin et al., we computed *t*-values for frontal and central ROIs in native speakers, where the article expectedness effect was significant, and used these *t*-values for the tests. Our data included native speakers from both SOA groups. The results from the Bayes factor replication tests are in Figure 2. The test for the frontal ROI and for the central ROI yielded moderate to strong evidence for the null hypothesis, $B_{r0} = .13$ and $B_{r0} = .07$, respectively, which means that the replication data are about (1/.13 ≈) 8

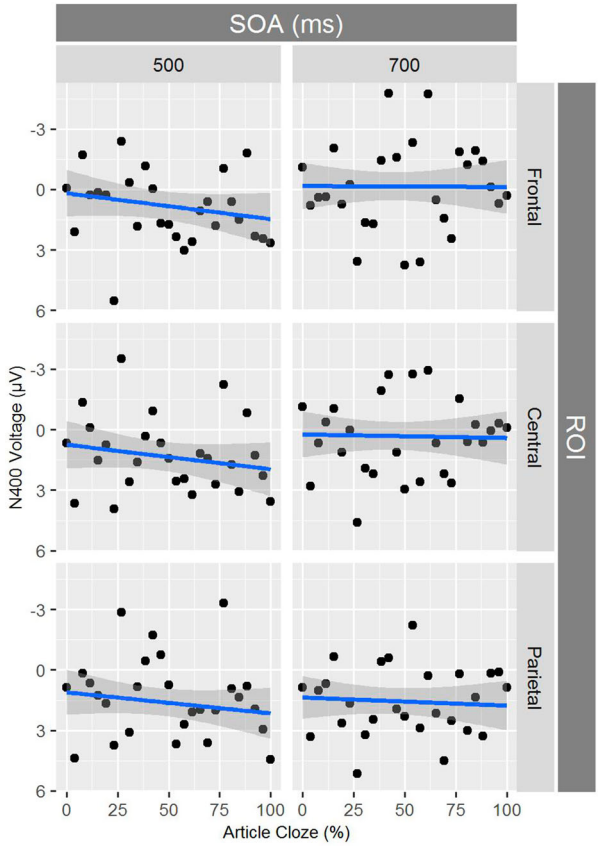

Figure 1. N400 voltage (mean values in the 250–400 ms time period) at three ROIs as a function of article-cloze probability. Each dot represents the average N400 voltage observed at one level of cloze value. The line represents the model fit, and the error bands represent 95% confidence intervals.

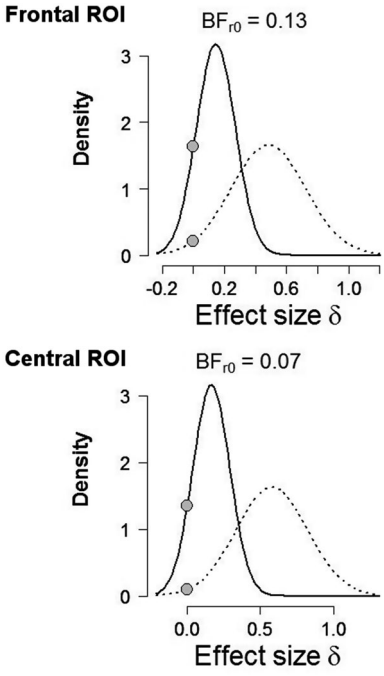

Figure 2. Results of the Bayes factor replication tests for the frontal ROI (top) and central ROI (bottom). The dotted lines represent the posterior from native speakers' data of the Martin et al.'s (2013) study, which was used as a prior for the effect size in our replication tests. The solid lines represent the posterior distributions after the data from the replication attempt are taken into account. The density (*y*-axis) of the prior distribution at $\delta = 0$ reflects the believability of H_o without the replication data, and the density of the posterior distribution at $\delta = 0$ reflects the believability of H_o after seeing the replication data. The grey dots indicate the ordinates of this prior and posterior at the sceptic's null hypothesis that the effect size is zero. The ratio of these two ordinates gives the result of the replication test, the BF_{r0}.

times and (1/.07 ≈) 14 times, respectively, more likely to have occurred under H_0 than under H_r. The meta-analysis Bayes factor with pooled data from both Martin et al.'s and our studies yielded $B_{10} = .16$ and $B_{10} = .20$ for the frontal ROI and for the central ROI, respectively, indicating that the combined data were about (1/.16 ≈) 6 times and (1/.20=) 5 times more likely under the null hypothesis. In other words, these results further confirm our failure to replicate the article-N400 effect, as we obtained moderate to strong evidence in support of the null hypothesis.

Finally, we also performed a Bayesian mixed-effect model analysis using the brms package (Bürkner, in press). Because Martin et al. did not use cloze as a continuous predictor and because they used a different electrode reference than ours, we based our prior on the output from a recent multi-lab replication attempt (Nieuwland et al., 2017; for discussion, see next section), which for a pre-registered analysis was an effect size of 0.296 μV for a 100% cloze difference (other aspects of the prior were the same as in Nieuwland et al.). We here report the estimate of the effect size (*b*), the 95% Credible Interval (i.e. the interval of which we can be 95% confident contains the true effect), and the posterior probability of the estimated effect is positive or negative as a percentage. For simplicity, we only report the results after collapsing the data over the three ROIs and ignoring the SOA, but the data patterns are roughly similar for separate ROIs. This yielded $b = .27$, CrI [−0.17, 0.72], positive 88.6%, negative 11.4%. Like in Nieuwland et al., we also reanalysed the data using a 0.1 Hz filter to address slow signal drift while presumably not impacting N400 activity. Using the new prior from Nieuwland et al., this analysis yielded $b = .12$, CrI [−0.36, 0.59], positive 69.9%, negative 30.1%, thus weakening the observed pattern. Like in Nieuwland et al., a similar pattern of cloze was already observed in the −200 to −100 ms pre-article time window ($b = .13$, CrI [−0.18, 0.43], positive 79.2%, negative 20.8%), shedding doubt on the conclusion that the previous patterns are elicited by the articles themselves. In all the analyses, zero was within the credible interval. Based on these analyses, we reach a similar conclusion as Nieuwland et al., namely that the evidence for the A/AN prediction effect is not convincing.

In sum, the previous three sections offer additional, statistical evidence against the A/AN prediction effect reported by Martin et al. (2013), and against the pre-activation gradient reported by DUK05. Beyond our own findings, however, there are also various indications that such an effect might be hard to observe, as we shall see in the next section.

4. On the replicability of the Delong et al. (2005) pre-activation gradient

To our knowledge, there is currently no published replication of the pre-activation gradient reported by DUK05. There have been, however, various attempts to replicate that effect, both conceptual and direct, including those cited by DUK17. For example, in the DeLong thesis Chapter 4, which included the same materials of DUK05 along with other filler sentences, the correlation analysis did not yield statistically significant effects, and only an ANOVA showed a significant main effect of article expectedness. In DUK05, however, the same ANOVA did not yield a significant article effect. This was reported in the DeLong thesis Chapter 2 (the chapter-equivalent to DUK05), but was not reported in DUK05. It is unclear which analysis was planned or performed first, and no justification was given. It is also unclear why the non-significant analyses were not reported in DUK05, and/or why both analyses were not

consistently reported. Another example is the thesis Chapter 3 reporting effects at a faster SOA (300 ms), again failing to replicate the pre-activation gradient. DUK17 argue here that a marginally significant gradient effect was observed in a subgroup of 11 experienced readers when using the cloze probability of the nouns, but no justification is given for analysing reading experience as a categorical variable rather than as a continuous measure, which the measure is (N.B. in less-experienced readers, the effect went into the opposite direction, so use of continuous measure would probably not show any hint of an effect). Importantly, also no justification was given for using noun cloze rather than article-cloze. Moreover, to our knowledge, these findings have not been replicated.

We also point out that in all the studies with this manipulation, if there is any visible effect of cloze on article-elicited ERPs at all, it already appears to exist well before the N400 time window (see Figure 3), and in DUK05 it already appears in the pre-stimulus baseline window (Figure 3(a), right panel). These differences question whether the observed differences are indeed N400 modulations, or slow signal drift that existed before the articles were presented (a pre-stimulus ERP difference that is not associated with the presentation of the critical word, i.e. a "baseline problem").

Chapters 3 and 4 of the thesis used the A/AN manipulation, but the results associated with that manipulation are not published. Based on these unpublished results, DUK17 argue that the questions that we posed about the potential importance of fillers and reading rate, were "already answered." We do not understand this response. Notably, even in published work, A/AN effects have not been reported when they could have occurred given the manipulation, but these could be cases where the effect was detected, but not reported. DeLong, Quante, and Kutas (2014) reported noun data from materials in which the A/AN manipulation was present; the A/AN manipulation was present in DeLong and Kutas (2016), though no article data were reported. Again, these points are not meant to imply that these cases are not replications – they may well be – but rather to simply state that the observation of the A/AN prediction effect was not reported, and thus, to our knowledge, no published replication of it exists.

Importantly, since the publication of DUK17, a direct replication attempt of DUK05 (Nieuwland et al., 2017) has yielded large-sample evidence for the non-replicability of the A/AN prediction effect. In that study, involving 9 laboratories and a sample more than 10 times larger than that of DUK05, there was no statistically significant effect of article-cloze on N400 activity, both in a pre-registered analysis that followed DUK05 and in a pre-registered, linear mixed-effects analysis. The original DUK05 results were the benchmark impetus for incorporating a particular version of prediction into models of language processing (Dell & Chang, 2014; Pickering & Garrod, 2013). Thus, the failure to replicate such an important finding, after multiple attempts and given the new likelihoods these have yielded, must naturally draw into question the centrality of probabilistic pre-activation of a predicted word's initial sound during reading.

We also note that the conclusions of DUK05 can be questioned on methodological grounds from the reported analyses alone. Their main claim of phonological pre-activation was based on an observed correlation showing smaller (i.e. less negative) article-elicited N400s with increasing cloze probability of the articles. They computed this correlation using 10 point cloze bins, each of which contained articles that fell in the 0–10% cloze bin, the 10–20% cloze bin, and up to the 90–100% cloze bin. They first computed ERP participant-averages for all trials in each bin and then used these values to compute bin averages. The correlation analysis thus involved only 10 data points per electrode channel, namely the bin-average ERPs as a function of the bin-average cloze. Our first concern with this approach is that treating items and subjects as fixed rather than random potentially inflates false-positive rates, due to the confounding of the overall cloze effect with by-subject and by-item variation in the effect (Barr, 2013). Strictly speaking, their analysis does not allow generalisation beyond the specific items and subjects in their experiment. Moreover, their analysis also disregarded variance at the level of the bin, which may have been substantial because the number of items in each bin was highly unbalanced (see Figure 4 for the count of items in each bin). The lowest-cloze bin, for articles, contained 58 items (the equivalent of about 36% of the items), whereas for example the highest-cloze bin contained only 7 items (about 4% of the items). Assuming, for the sake of the argument, zero trial-loss, the ERP values for the lowest-cloze bin would be calculated based on 928 items, whereas the values for the highest-cloze bin would be based on 112 items, more than 8 times less than that for the lowest-cloze bin. However, in their analysis, DUK05 disregard the differences in variance associated with such item count differences.

A final note on the role of filler materials; that no fillers were reported in DUK05 is relevant, because DUK17 argue that "a stimulus set in which over half the items have been rated implausible, as they were in the Ito et al. study, does little to allay concerns about potential adoption of strategic processes." We agree with this concern in general, but we point out that a similar

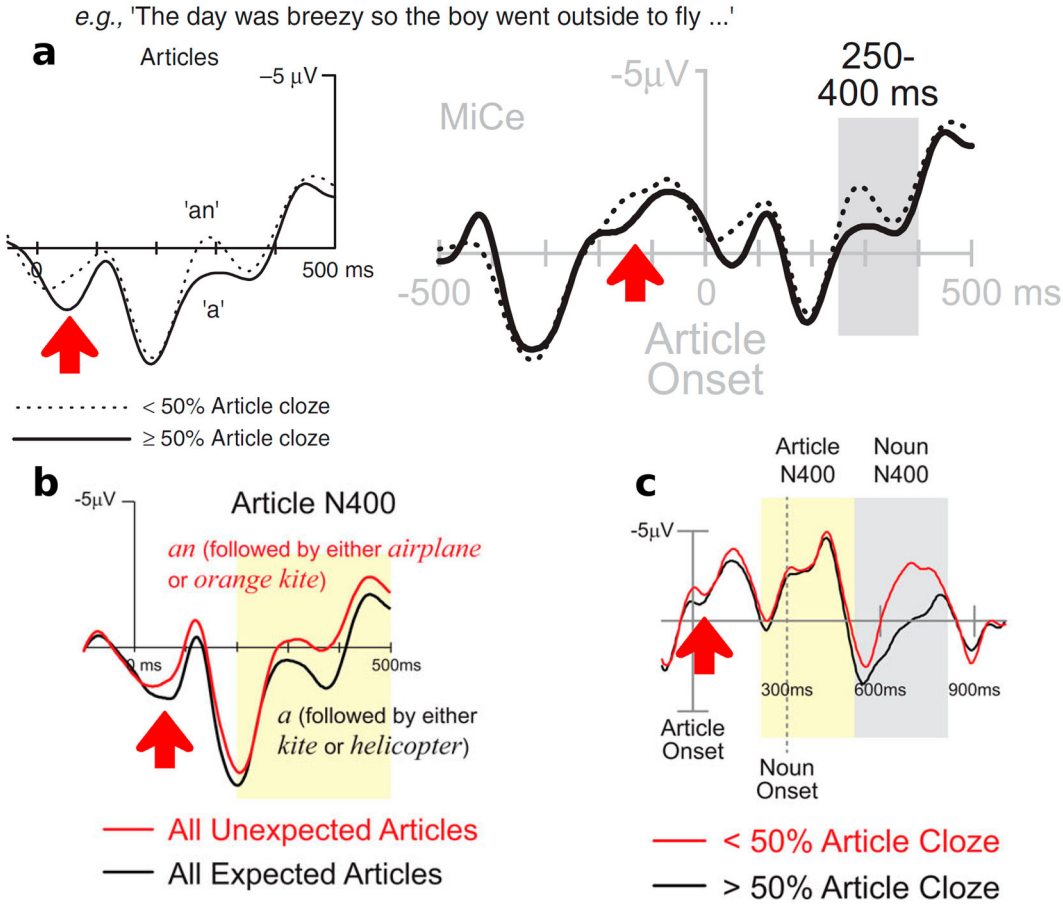

Figure 3. ERP plots taken from (a) DeLong et al. (2005; left) and DeLong et al. (2012; right) for the same data with a longer pre-stimulus window and (b and c) DeLong et al. (2017). In all the plots, there is a baseline ERP difference between unexpected and expected articles (indicated with a red arrow in each plot), calling into question whether the later effects are genuine N400 modulations or arising from slow potential drift.

concern could be raised against DUK05. DUK05 presented their sentence materials as plausible and report no fillers. However, in DUK17 previously undisclosed filler materials are described as being used in DUK05 and as being plausible. However, in neither case do DUK05 or DUK17 substantiate these descriptions with

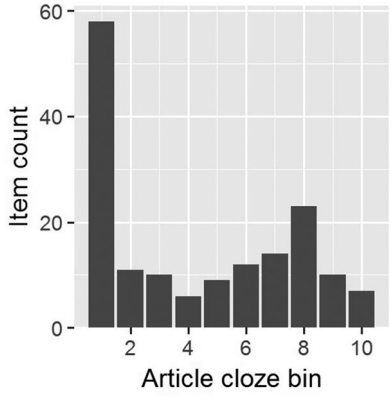

Figure 4. The number of items (articles) in each cloze bin used by DeLong et al. (2005).

plausibility norms that quantitatively illustrate plausibility. Moreover, plausibility norms reported in other papers by DeLong and colleagues suggest that half of the DUK05 materials may have been implausible (though perhaps not anomalous). DeLong et al. (2014) used similar sentences with article-noun combinations as DUK05 and reported that the unexpected article-noun combinations were rated, on average, as 2.8 on a 5-point plausibility scale. Urbach and Kutas (2010, Experiment 1) performed an ERP study using the materials described by DUK17 as the fillers for DUK05. DUK17 mention only one somewhat plausible example ("Bakers slice pizza in a special cutting machine"), but the materials contain many implausible sentences like "Instructors evaluate grapes by giving them a test," "Bartenders mix metaphors while talking to the patrons," and "Weathermen report weddings that are expected in the next few days." No plausibility ratings were obtained, but in Experiment 2, where quantifier versions of these sentences were rated (e.g. "A large number of instructors evaluate grapes … "), those sentences received a 2.2 on a

5-point plausibility scale. In other words, a closer look at other studies by DeLong et al. that used the exact same or similar materials and alleged fillers of DUK05 suggests a different story about plausibility of those materials. This matters not only because DUK05 and DUK17 describe their materials as being plausible without any norming, but also because concerns about potential adoption of strategic processes apply just as well to DUK05, not just our own work. In our study, the fillers were indeed often anomalous, but the critical nouns were in fact equally plausible in the expected and unexpected condition, probably unlike those in the DUK05 materials. More generally, we doubt that effects that are strongly driven by the presence or absence of fillers, or by any other aspects of the fillers, such as plausibility, are the kinds of effects that a general theory of the architectures and mechanisms of language processing should be based on.

5. Other article-elicited pre-activation effects

We emphatically do not argue against a general role for prediction in language processing and certainly do not do so based on our current results. If the A/AN prediction effect does not replicate, that fact does not negate prediction effects elicited by other pre-nominal article manipulations. DUK17 argue that pre-nominal effects from A/AN articles, gender and animacy are considered to be strong evidence for prediction compared to effects found on nouns. However, it is also important to acknowledge the key differences between the A/AN manipulation and gender- or animacy-based computation. Unlike Dutch or Spanish gender-marked articles, for example, English indeterminate articles do not agree with the upcoming noun. When articles and nouns agree in gender regardless of intervening words, an article that is gender-inconsistent with the predicted noun can immediately disconfirm that noun is upcoming. However, English indeterminate articles are only informative about the initial phoneme of the next word. Thus, they cannot reliably disconfirm prediction of a noun, because there is no phonological dependency between the a/an article and the noun when the predicted noun does not come immediately after the article. Estimates from natural language corpora suggest that the probability that a/an articles are followed immediately by a noun is only about 33% in both American and British English (The Corpus of Contemporary American English, Davies, 2008; "The British National Corpus", 2007), meaning that using these articles to confirm or disconfirm the prediction of a given word would not be a very efficient strategy.

But even for gender-marked articles, there is inconsistency in the type of article effects that have been obtained in terms of ERP componentry, and this too signals that there will likely be problems with replicating these effects. For example, when studying grammatical gender agreement between article and noun in Spanish, Wicha and colleagues sometimes report N400 effects for articles that mismatch the gender of an expected noun, and sometimes P600 effects (Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2012). Utilising adjective-noun or article-noun gender agreement in Dutch, three published studies each reported a qualitatively different ERP effect (Otten & Van Berkum, 2009; Otten, Nieuwland, & Van Berkum, 2007; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). An account of what causes these differences is currently lacking, but the specific materials and design could play a role. For example, the studies by Wicha and colleagues sometimes mixed reading with picture comprehension and always required participants to read a large number of semantically and/or grammatically anomalous sentences (such that unexpected articles could cue upcoming anomalies), whereas the studies by Van Berkum and colleagues involved more subtle manipulations of noun expectancy. But even with highly similar materials and manipulations, failures to replicate the same ERP effect exist in this modest literature (e.g. Otten et al., 2007; Van Berkum et al., 2005). These inconsistencies warrant future investigations, because they signal something important about the language comprehension system. Simply taking them all to index the same representations or processes seems misplaced, and our message is simply that more caution with theoretical inference is needed given the nature of the evidence.

We also question what evidence there is to suggest pre-activation is an integral part of language comprehension. DUK05 appear to come to this conclusion based on observed similarity of article-N400 and noun-N400 modulations by cloze, and on the observed gradient (suggesting effects at low-cloze values too, not just at high cloze). However, in a subsequent single-trial reanalysis in the DUK05 data (DeLong, Groppe, Urbach, & Kutas, 2012), the article- and noun effects looked very different in terms of timing and distribution, and a gradient effect could also result from the averaging of items in which pre-activation occurred in an all-or-none fashion (see Van Petten & Luka, 2012), or simply from a difference between a cluster of high-cloze articles and low-cloze articles. Judging from the DUK05 article-correlation data (Figure 1(b), left panel, in DeLong et al. 2005), it is unclear whether an article-cloze effect would be found if only low-cloze trials were analysed. We therefore do

not believe that the DUK05 data unambiguously support the conclusion that, for example, pre-activation of form and meaning of nouns with a 10% cloze value is double the strength of the pre-activation of form and meaning of nouns with a 5% cloze value. To be clear, however, we are not questioning the existence of *a* role of prediction in language processing, but rather, critically evaluating the evidence for (1) the situations in which prediction can be observed, and (2) the representational granularities at which prediction can occur. We note that both of these points have resulted in major architectural claims in models of language processing over the last decade (for a review, see Huettig, 2015; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2013), making consideration of the empirical evidence vital, in our view.

6. Conclusion

In their opening paragraph, DUK17 cite Pashler and Harris (2012), who make a strong case for the importance of direct replication research, amongst other things. Using this citation, DUK17 suggest that we should not have learned anything from our replication attempt. But the point that Pashler and Harris make is that a focus on conceptual replication in an academic field can appear to confirm the reality of a non-existent phenomenon, whereas in our paper we in fact question the reality of the A/AN effect. Pashler and Harris also describe the frequent opposition to replication, sometimes by prominent and accomplished researchers who see the replication crisis as overblown. We think this opposition is visible in psycholinguistics, too. Conceptual and direct replication research is unfortunately very sparse (but see Nieuwland et al., 2017; Rommers, Meyer, & Huettig, 2013; Zwaan & Pecher, 2012; see also Jäger, Engelmann, & Vasishth, 2017), and even novel but incremental contributions are often considered insufficient for publication.

Based on the results reported in Ito, Martin, et al. (2016a) and on the additional observations presented in this rebuttal, we conclude that our study *is* a failure of conceptual replication of Martin et al. (2013) and, by extension, of DeLong et al. (2005). Simply put, we see the A/AN prediction effect as reported by DUK05 as hard to replicate – to our knowledge there is no published replication of the A/AN prediction effect, although there are several instances in the literature where the effect could have been observed or tested for (DeLong & Kutas, 2016; DeLong et al., 2014), and the largest scale attempt at replication to date (Nieuwland et al., 2017) did not yield a statistically significant A/AN prediction effect. The A/AN prediction effect reported by DeLong et al. may not be a real effect, in other words, may be a Type I error. Alternatively, this effect is so small it could not be reliably detected in our replication attempts. Even if a small effect were real, but difficult to detect because of its size, one could question whether such an effect should be regarded as stalwart evidence that people probabilistically pre-activate phonological information and that this activation plays a meaningful role in everyday language comprehension. The same question could be raised if an effect can only be detected under limited circumstances (e.g. with a certain type of fillers, in a certain population at a certain time, in very high-cloze sentences) – which themselves do not arise as readily outside the lab. We think that such inference is problematic, and we are sceptical that such a small effect should be taken as evidence that phonological pre-activation is an integral part of language processing in the wild. Nonetheless, the literature, including current influential theories and models of language processing (Dell & Chang, 2014; Pickering & Garrod, 2013), have given the effects reported in DUK05 precisely this interpretation.

Note

1. Even if we had failed to replicate a previously published noun effect, the null-hypothesis significance testing (NHST) and frequentist statistics dictate that failure to observe "real" effects (Type II error) and observation of "false" effects (Type I error; Gelman, 2015) are part and parcel of the scientific endeavour. Had we failed to observe a noun effect (which we did not), such a case would not diagnostically indicate there were technical problems with our experiment or data, but rather that, in NHST terms, we restricted our acceptable error to a probability less than 0.05. This uncertainty in observation or sampling error is at the heart of scientific inference and we adamantly maintain that Type I and II error and sampling error be considered in the conclusions we all draw from the effects we observe.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Andrea E. Martin was supported by a Future Research Leaders grant from the Economic and Social Research Council of the UK [grant number ES/K009095/1].

ORCID

Aine Ito http://orcid.org/0000-0003-4408-8801

References

Ardal, S., Donald, M. W., Meuter, R., Muldrew, S., & Luce, M. (1990). Brain responses to semantic incongruity in bilinguals. *Brain and Language*, *39*(2), 187–205. doi:10.1016/0093-934X(90)90011-5

Baayen, H. R., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. doi:10.3389/fpsyg.2013.00328

Boekel, W., Wagenmakers, E. J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. doi:10.1016/j.cortex.2014.11.019

The British National Corpus. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/cpr.xml?ID=reference

Bürkner, P.-C. (in press). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*.

Davies, M. (2008). *The corpus of contemporary American English: 450 million words, 1990–present*. Retrieved from http://corpus.byu.edu/coca/

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B*, *369*, 20120394. doi:10.1098/rstb.2012.0394

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, *121*(3), 226–239. doi:10.1016/j.bandl.2012.02.006

DeLong, K. A., & Kutas, M. (2016). Hemispheric differences and similarities in comprehending more and less predictable sentences. *Neuropsychologia*, *91*, 380–393. doi:10.1016/j.neuropsychologia.2016.09.004

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*(1), 150–162. doi:10.1016/j.neuropsychologia.2014.06.016

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), 1117–1121. doi:10.1038/nn1504

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 & Nieuwland (2016). *Language, Cognition, and Neuroscience*. doi:10.1080/23273798.2017.1279339

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1461–1469. doi:10.1037/a0036756

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*(2), 632–643. doi:10.1177/0149206314525208

Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: Off-line and on-line performance. *Bilingualism: Language and Cognition*, *12*(4), 463–483. doi:10.1017/S1366728909990253

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. doi:10.1016/j.brainres.2015.02.014

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. doi:10.1016/j.jml.2015.10.007

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

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016b). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning Memory and Cognition*, doi:10.1037/xlm0000315

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Jeffreys, H. (1961). *The theory of probability*. Oxford: Oxford University Press.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59. doi:10.1080/23273798.2015.1102299

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

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. doi:10.3758/s13428-010-0049-5

Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish-English bilinguals. *Cognitive Brain Research*, *22*(2), 205–220. doi:10.1016/j.cogbrainres.2004.08.010

Nieuwland, M. S. (2016). Quantification, prediction and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334. doi:10.1037/xlm0000173

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … 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:10.1101/111807

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. A. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*, 89. doi:10.1186/1471-2202-8-89

Otten, M., & Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research*, *1291*, 92–101. doi:10.1016/j.brainres.2009.07.042

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. doi:10.1177/1745691612463401

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioural*

and Brain Sciences, *36*, 329–392. doi:10.1017/S0140525X12001495

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–163). Cambridge: Blackwell.

Rommers, J., Meyer, A. S., & Huettig, F. (2013). Object shape and orientation do not routinely influence performance during language processing. *Psychological Science*, *24*(11), 2218–2225. doi:10.1177/0956797613490746

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179. doi:10.1016/j.jml.2010.03.008

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. doi:10.1037/0278-7393.31.3.443

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. doi:10.1016/j.ijpsycho.2011.09.015

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. doi:10.1037/a0036731

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience*, *8*(3), 231–256. doi:10.1162/jocn.1996.8.3.231

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*(3), 165–168. doi:10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2012). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Cognitive Neuroscience*, *16*(7), 1272–1288. doi:10.1162/0898929041920487

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE*, *7*(12), e51382. doi:10.1371/journal.pone.0051382