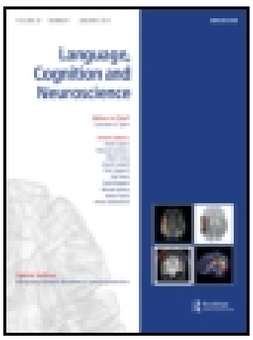

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

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

Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016)

Katherine A. DeLong, Thomas P. Urbach & Marta Kutas

To cite this article: Katherine A. DeLong, Thomas P. Urbach & Marta Kutas (2017): Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016), Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2017.1279339

To link to this article: http://dx.doi.org/10.1080/23273798.2017.1279339

Published online: 16 Jan 2017.

LANGUAGE, COGNITION AND NEUROSCIENCE, 2017
http://dx.doi.org/10.1080/23273798.2017.1279339

Routledge
Taylor & Francis Group

COMMENTARY

Is there *a* replication crisis? Perhaps. Is this *an* example? No: a commentary on Ito, Martin, and Nieuwland (2016)

Katherine A. DeLong[a], Thomas P. Urbach[a] and Marta Kutas[a,b,c,d]

[a]Department of Cognitive Science, University of California, San Diego, CA, USA; [b]Center for Research in Language, University of California, San Diego, CA, USA; [c]Department of Neurosciences, University of California, San Diego, CA, USA; [d]Kavli Institute for Brain and Mind, University of California, San Diego, CA, USA

ARTICLE HISTORY Received 7 November 2016; Accepted 19 December 2016

The *Language, Cognition and Neuroscience* article by Ito, Martin, and Nieuwland (2016a), "How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects", reports an attempted conceptual – not direct – replication of Martin et al. (2013), which derives from DeLong, Urbach, and Kutas (2005). The title poses an important question, although readers may ultimately find the answer unsatisfying. In discussing the replicability crisis in psychology, Pashler and Harris (2012) write: "If a conceptual replication attempt fails, what happens next? Rarely, it seems to us, would the investigators themselves believe they have learned much of anything." However, Ito et al. represent themselves as the rare exception, stating: "These observations [a critique of Martin et al., 2013], along with the failure to replicate the article-effect in the current study, suggest that article-elicited N400 effects may not have high external validity." Not only do Ito et al. believe they have learned something from their (ostensible) failure to replicate, what they propose derives solely from the interpretation of null results – an approach that generally violates sound scientific reasoning. In addition, the framing of their study as a failure to replicate is a crucial mischaracterisation, one of several missteps in exposition, scholarship, experimentation, and theoretical inference. The reader who uncritically accepts their assertions and conclusions will be led into serious error.

We begin with the most troubling aspect of Ito et al.'s report: the unjustified and inconsistent manner in which null results are used as evidence (or not). The primary example appears in their central argument against linguistic prediction – the argument from which their report gets its title. A basic principle in null hypothesis testing is that tests that fail to reach statistical significance do not constitute evidence that the null hypothesis is true. The authors are aware of this and rightly state (twice) that "… null results can never prove a negative (i.e. that our participants did not predict upcoming words)". And yet, they proceed to advance arguments that violate this principle anyway. For example, the authors interpret putative null results to more and less expected *a*'s and *an*'s in the article N400 time window as evidence that "… prediction effects are perhaps not in fact representative of how people comprehend language in natural settings." Conceding awareness of the principle does not constitute an exemption. It is not clear why the authors expect the intelligent reader to join them in deliberately making this error.

Furthermore, the authors' approach to interpreting null results and failures to replicate is inconsistent even within the scope of their report. Whereas they place great theoretical weight on their failure to find a statistically reliable effect at the pre-nominal articles (even when, as noted, doing so is unjustified by their own admission), other null results are brushed aside for no discernible reason. This inconsistency is patent in their treatment of the null N400 effect at the critical nouns for bilinguals in Experiment 1 (see their Figure 1). There is ample evidence in the literature that non-native speakers continue to exhibit noun N400 cloze probability effects in their less dominant language, although sometimes at longer latencies (see Frenck-Mestre, 2005; Kutas, Moreno, & Wicha, 2009, for reviews). A number of studies have reported bilinguals' N400 sensitivity to contextually facilitated nouns using presentation rates identical or similar to the 500 ms stimulus onset asynchrony (SOA) used in Ito et al.'s Experiment 1 (e.g. Ardal, Donald, Meuter, Muldrew, & Luce, 1990; Moreno & Kutas, 2005;

CONTACT Katherine A. DeLong kadelong@ucsd.edu

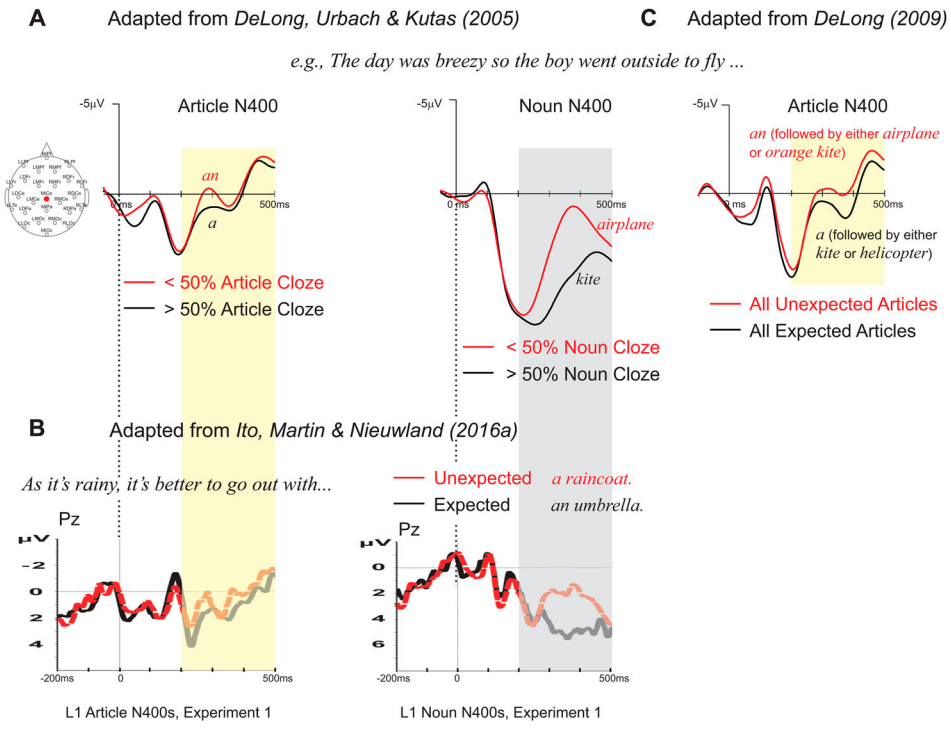

Figure 1. (A) ERPs taken from Figure 1 of DeLong et al. (2005), showing waveforms at the vertex recording site for articles and nouns separately. Conditions were created according to median splits on article and noun cloze probabilities in that study, and plots focus on the N400 time windows over which correlational analyses were conducted (200–500 ms post-word onset). (B) ERPs adapted from Figure 1, Experiment 1 of Ito et al. (2016a), showing expected and unexpected articles and nouns for L1 participants. Descriptively, ERP patterns from the two studies are strikingly similar. (C) Data taken from Figure 4.5 of DeLong's dissertation (2009), plotting ERPs at the vertex electrode over the article N400 time window for all Expected articles (those articles consistent with the highest cloze probability nouns, but which could be followed by either a high or low cloze probability noun), and all unexpected articles (those articles inconsistent with the highest cloze probability nouns, but which in the experiment could be followed by either a low cloze probability noun or an adjective followed by the highest cloze noun). Despite what Ito et al. (2016a) suggest would be a decreased "cue value" for the indefinite articles in this study, there is a reliable N400 article expectancy effect [$F(1, 31) = 4.52$, $p = .0416$] for the ANOVA calculated for 32 participants, with 2 levels of Article Expectancy and 26 levels of Electrode.

Weber-Fox & Neville, 1996). Even the authors themselves (Ito, Martin, & Nieuwland, 2016b) have reported such effects in bilinguals at a 500 ms SOA. However, since Ito et al. (2016a) do not cite any of these studies, we point them out here so that readers will be aware that N400 noun effects are well attested under conditions similar to Ito et al.'s Experiment 1.

Since a reliable N400 effect at the more vs. less predictable nouns was not observed for the bilinguals in Experiment 1, it is a null result and, given the literature, a failure to (conceptually) replicate. In this respect, the noun results are the same as the (putative) null prenominal article N400 results. By the same (incorrect) inference from null results they use to argue against predictive language processing, their null noun N400 finding in bilinguals should argue against ease of semantic access, facilitated integration, or whatever functional theory of N400 processing they ascribe to. It should also count as evidence that contextual predictability effects are only observed under special experimental conditions perhaps not representative of natural language, that the findings are part of an ongoing replication crisis, etc.. Although the authors are logically committed to these erroneous inferences, we in no way suppose that they or anyone else should accept them. Indeed, quite the opposite: they should be categorically rejected. And likewise, the inferences the authors purport to draw from the null results at the prenominal articles should also be rejected.

This is not to say that the null noun N400 effect for bilinguals in Experiment 1 is unimportant. To the contrary, we think these results are both striking and critical, though for reasons that work against some of Ito et al.'s conclusions. The authors, however, do not appear to share our view of the importance of *this* null effect. They neither discuss it in any depth, nor interpret it within the context of the wider literature. Rather, the authors merely suggest that the failure to observe a noun N400 in bilinguals in Experiment 1 "… could be because the reading rate of 500 ms SOA was too fast

for them." This is a logical possibility but empirically implausible. Whereas the authors cite several bilingual sentence ERP studies that have, like their Experiment 2, used slower SOAs, they fail to mention any of the aforementioned directly relevant reports that militate against the speculation that rate is the relevant variable. And only a single sentence of discussion is devoted to the null noun N400 finding: " … non-native speakers in our experiment appeared to be insensitive, at least in the initial stages of semantic processes reflected in N400 activity, to the predictability both of the article and of the noun." In our view, an alternative interpretation must at least be considered. When robust N400 effects at open class words are widely replicated across labs, languages, comprehenders, and presentation rates, their absence under similar conditions undermines the face validity of experimental results. In short, since bilinguals in Experiment 1 did not even show the well-attested N400 noun effect, there are no grounds whatsoever for supposing that other failures to find effects are anything more than that – failures to find effects.

Based on (what Ito et al. inappropriately/inaccurately conclude is) a failure to replicate results from one particular prediction experimental paradigm, they call into question linguistic prediction more generally. Moreover, they do so despite reports, as they acknowledge, of "widely replicated prediction effects" which go beyond the *a*/*an* ERP results (a paradigm used only by DeLong et al. and Martin et al.). If, indeed, their aim is to argue against prediction as an "explanatory mechanism in language processing", then at a minimum, all of the pre-nominal prediction paradigm studies – conceptually similar to the *a*/*an* studies – are inarguably relevant. And yet, Ito et al. either do not cite or do not discuss the seminal work of Wicha and colleagues (using determiner-noun gender agreement in Spanish: Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2003, 2004) that pioneered this experimental design, the work of van Berkum and colleagues (using gender-marked pre-nominal adjectives in Dutch: Otten & Van Berkum, 2008; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005), or that of Szewczyk and colleagues (using animacy marked pre-nominal adjectives in Polish: Szewczyk & Schriefers, 2013). All of these report reliable ERP differences to words preceding more and less expected nouns, indicating that information about likely upcoming words or word features can be activated prior to their receipt. In other words, they provide some ERP evidence of predictive processing, albeit not all with identical ERP effects. By choosing to ignore these studies, Ito et al. also choose to limit the allowable conclusions as to what can be inferred about the specific type of prediction targeted by the *a*/*an* paradigm; namely, prediction of specific word forms – what many would consider the "strongest" form of prediction.

Despite these limiting choices, it appears Ito et al. intend to draw broad inferences and challenge the general role of prediction in theories of language comprehension. That said, it is not entirely clear what the authors intend to challenge. The article begins, "In current theories of language production and comprehension, prediction plays an outsized role as the mechanism by which language processing can occur quickly, incrementally, and rather effortlessly." It appears the authors intend to refer to some larger, unspecified, class of language comprehension theories in which prediction plays an "outsized" role. However, there are no further particulars regarding this, nor any hint about what, in their view, makes a role of prediction outsized vs. undersized vs. just-right-sized in such a theory. The first sentence is thus a vague gesture in the general direction of a potential problem, and we are unsure how to determine what proposition is being articulated or how it might be evaluated.

This murkiness of exposition continues in regard to which authors and studies are intended to bear the brunt of Ito et al.'s argumentation, given that, to our knowledge, none of the pre-nominal word ERP studies have made the strong claim that prediction is required for language comprehension. This is an idea Ito et al. aim to discredit, writing "it is not clear that the existence of such effects warrants the conclusion that prediction is a necessary computation in language processing." But surely it is perfectly clear. Such effects do not warrant the conclusion that prediction is a necessary computation in language comprehension. We wonder who said that they did.

Ito et al. cite Huettig and Mani (2016) as asserting that many researchers "explicitly or implicitly appear to support the notion that prediction is necessary to understand language", but Huettig & Mani, in turn, present no evidence for this. Views on the role of prediction in language are legion and can only be determined with considerable attention to detail (for discussion on this topic, see Kuperberg & Jaeger, 2016). Although we cannot speak for others, DeLong et al. (2005) intentionally wrote, "We believe that this sort of anticipation is an integral (perhaps inevitable) part of real-time language processing" and not "a necessary (perhaps inevitable)" part of real-time language processing. Indeed, DeLong, Troyer, and Kutas (2014) detail specific examples of limits to predictive processing, in particular across the lifespan and modulated by various individual factors.

In challenging a (putative) received view of prediction as a central explanatory mechanism in language

processing, Ito et al.'s assertion that predictability effects may be limited to certain experimental designs (e.g. to high constraint contexts) is not a new claim. Whether or to what extent this is true was already at issue in, e.g. Fischler and Bloom (1979) and Mitchell and Green (1978), and the field has gone back and forth since. The experimental difficulty in demonstrating prediction during language comprehension has been coming up with compelling evidence, given that processing disruptions at a low probability word are consistent with either: (a) violating a prediction, (b) difficulty integrating that word into the ongoing sentence in the absence of any prediction, or (c) both. That is what differentiates the conceptually similar Wicha, Bates, et al. (2003, 2004), Van Berkum et al. (2005), DeLong et al. (2005), and Martin et al. (2013) studies from earlier work. In fact, Kutas, DeLong, and Smith (2011, p. 196), as well as DeLong et al. (2005, p. 1117), outline the argument above and the distinction between inferences that can be drawn from nominal and pre-nominal effects, concluding that "Clearly, an argument for information getting pre-activated would be strengthened if it could be demonstrated that predictions were being formulated prior to target words."

Mindful of the criticism that prediction may be restricted to high constraint contexts, DeLong et al. (2005) sought evidence for graded pre-activation by correlating pre-nominal article N400 amplitudes and article cloze probabilities, finding reliable inverse correlations. Ito et al. do not report any correlations or any analyses treating expectancy as a continuous variable. Instead, they conducted ANOVAs, found a marginal N400 effect at the articles in Experiment 1, though none at the slower SOA in Experiment 2, and took both as null effects, which they interpreted as a failure to replicate. They also conducted a linear mixed-effects model analysis on the articles in Experiment 1 that resulted in a null effect, but which again treated cloze probability as a categorical, not a continuous, variable.

The analysis used in DeLong et al., with individual electrode correlations calculated using 10 data points, was one way of investigating what we hypothesised was a graded response to more and less predicted input. The correlations that DeLong et al. (2005) observed at the articles and nouns showed the same: (1) directionality–pattern of decreasing ERP negativity with increasing cloze probability, (2) timing–occurring in canonical visual N400 time windows, and (3) scalp distribution–highest correlations over central-posterior scalp sites. Indeed, these similarities are what make these results so striking and impactful. Neither of the categorical analyses conducted by Ito et al. attempted to replicate – directly or conceptually – the correlational analysis of DeLong et al. (2005), or more generally test the hypothesis that pre-nominal article N400 effects are graded as a function of predictability. Thus, far from failing to conceptually replicate, the analyses Ito et al. report do not even attempt to address what we take to be one of the main points about predictive processing – that it is *not* all-or-none.

It is also worth noting that the number of experimental trials in Ito et al. was 64 (32 per condition), whereas both the DeLong et al. and Martin et al. studies used 80 experimental items (40 per condition for Martin). Ito et al. report an average 22% data rejection rate, yielding an average of only 25 artifact-free trials per condition. Because averaging across fewer trials reduces the signal-to-noise ratio of the ERPs (other things equal), these lower counts may be troubling, given that the article N400 prediction effect is reported to be smaller than that for nouns. Taken with the fact that the reported .06/.05 article *p*-values are already marginally significant in Experiment 1, it is unclear how the hypothesis tests and conclusions drawn might have been different if modest increases in the number of items per condition or number of participants had increased statistical power.

For a more direct comparison with Ito et al.'s Experiment 1 results, we conducted an ANOVA on two subsets of critical articles from DeLong et al. (2005), whose average cloze probabilities mirrored those of the Expected (75.1%) and Unexpected (14.5%) article conditions in Ito et al. These article bins were comprised of 57 higher and 57 lower cloze probability articles that yielded average cloze probabilities of 75.3% (SD = 10.5) and 14.2% (SD = 13.7), respectively. An ANOVA between 200 and 500 ms post-article onset for the 32 participants with 2 levels of Cloze Probability (higher, lower) and 26 levels of Electrode yielded non-significant results [$F(1, 31) = 2.37$, $p = .13$], with mean amplitude values of −0.60 μV for higher and −1.03 μV for lower cloze probability articles. In addition, we conducted an ROI analysis similar to that done by Ito et al. and did not find significant article effects. Although our laboratory does not use a 10–20 system, we used electrodes over similar scalp areas as Ito et al. and Martin et al. (Frontal: LDFr, RDFr, LMFr, RMFr; Central: LDCe, RDCe, LMCe, MiCe; Parietal: LDPa, RDPa, MiPa; see Figure 1 for electrode scalp locations). We found no significant differences ($p < .05$) over Frontal, Central, or Parietal channels, from either 200 to 500 ms (the N400 time window from DeLong et al.) or 250 to 400 ms (the N400 time window used by Ito et al.). There was one marginally significant effect in the expected direction between 250 and 400 ms at Frontal sites [$F(1, 31) = 3.27$, $p = .0805$]. These particular analyses, like Ito et al.'s, do not provide

evidence for pre-activation of pre-nominal articles *per se*, but given the reliable inverse correlation (a more sensitive measure), it is clear that important information about this relationship is lost with categorical binning for the ANOVA.

Regardless of the statistical tests performed, the ERPs reported at critical test words in Experiment 1 (refer to Ito et al., Figure 1, L1 participants) and those in DeLong et al. (2005) show qualitatively similar N400 morphologies (Figure 1(A,B) directly compares ERP waveforms from these two experiments). In our estimation then, Ito et al.'s findings at the articles in Experiment 1 are consistent both with those that we have observed and those of Martin et al. (2013): specifically, there is no failure to replicate. As for the article results of Ito et al. Experiment 2 when the SOA is longer (700 ms), again, unexpected article N400s were numerically more negative than those to expected articles, although the difference was not statistically reliable (unlike results reported by Martin et al. for monolinguals). At best, that is one study in favour and one not (null effect), at the longer SOA. We do, however, wonder why Ito et al. did not find an article prediction effect in Experiment 2, when they cite other work testing longer SOAs (e.g. Dambacher et al., 2012, in a personal communication; Wlotko & Federmeier, 2015) as indicating that prediction is more likely at these slower rates. In any case, the absence of an article prediction effect at the longer SOA is yet another null finding that warrants further investigation.

Indeed, the particular SOA likely does matter for prediction, but we already knew that. For example, in DeLong's dissertation (2009, Chapter 3), in a faster SOA (300 ms) version of the DeLong et al. (2005) study, a conditional effect of article expectancy was, again, not observed in the N400 time window [$F(1, 31) = .99$, $p = .33$], see Figure 2(A). A marginally significant graded N400-like prediction pattern was observed at the pre-nominal articles, but only for a subset of participants, i.e. more experienced (Figure 2(C)) but not less experienced readers (Figure 2(B)) as assessed by an offline measure of print exposure, and associated with the cloze probability of the subsequent critical noun, rather than the article cloze probability. Input speed thus may be a limiting factor for when pre-nominal article effects of prediction of specific word forms may be observed. Does this mean that prediction does not occur when input is speeded? Not necessarily. It may occur, at some level or for some set of features, or even at the level of word forms for a subset of readers. The approach taken in DeLong's dissertation, examining predictive ERPs in different comprehender groups, was an exploratory analysis that calls for further exploration. The findings do not necessarily undermine a role for predictive processing but rather suggest bounds on when prediction of particular word forms might be observed.

Inexplicably, Ito et al. take their (ostensible) null article results to argue against prediction in natural language processing except under specific experimental designs. They attribute the Martin et al. (2013) and DeLong et al. (2005) data to the absence of filler items: "This may have caused participants to pay extra attention to the *a/an* manipulation, thereby inadvertently encouraging participants to engage in predictive processing." Although not noted in DeLong et al. (2005) – an admittedly important but inadvertent omission on our part – filler sentences *were* interspersed with experimental items throughout ERP recording in that experiment. These filler sentences comprised 59% of the total sentence stimuli presented (116 out of the 196 items per participant) and were generally plausible, with more and less typical direct object continuations to unique Subject–Verb combinations (e.g. "Bakers slice bread/pizza in a special cutting machine."). In addition, there were 40 other indefinite article (*a/an*) instances not associated with the experimental manipulation. Also, the DeLong et al. stimuli had a broad range of article and noun cloze probabilities, with critical words that were sentence medial – not sentence final, as in Ito et al. – and were presented for the same duration and at the same SOA as the other words of the sentences in which they were embedded (unlike Ito et al., in which the critical nouns appeared for a longer duration than the preceding words of the sentence). All these factors in combination point to the experimental manipulation in DeLong et al. being less obvious than in Ito et al., and less likely to have (strategically or unconsciously) led participants to adopt a more predictive approach. On our view of prediction, contextually based pre-activation of linguistic information is a natural consequence of how stored information is accessed during the comprehension process. It is as yet unclear *if* and *how* local experimental factors such as stimulus proportions might impact the degree to which words are unconsciously activated during online language comprehension. Certainly, the use of 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.

Finally, on the subject of prediction in natural language settings, Ito et al. state:

> … we think that cue-reliability may be an important factor in experiments without fillers, where each sentence contains an article that reliably confirms or disconfirms the sentence-final expected noun. In such an experimental setting, participants' realisation about this

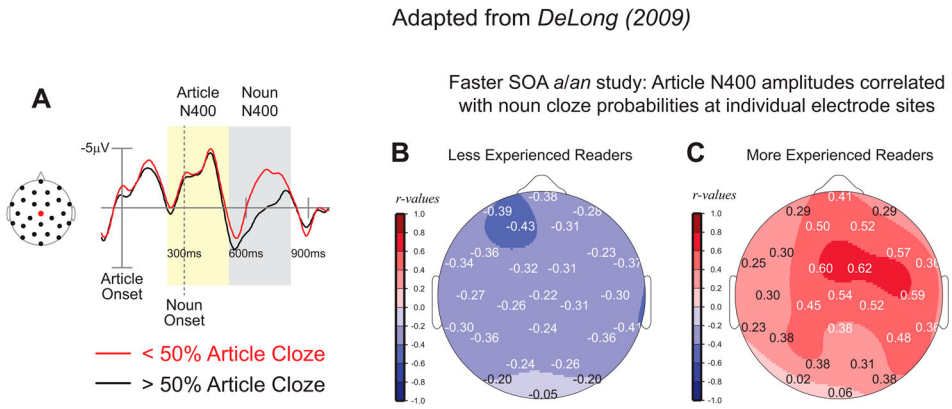

Figure 2. ERPs taken from Figures 3.2 and 3.10 of DeLong's dissertation (2009), with Chapter 3 describing a faster SOA (300 ms) version of DeLong et al. (2005). (A) ERP plot of the article+noun time window sorted by high (≥50%) and low (<50%) article cloze probability, with the N400 windows (200–500 ms post-word onset) highlighted in yellow for the articles and grey for the nouns. There was no significant conditional article N400 effect based on article cloze probability in this study. Panels (B,C) map *r*-values at individual electrode sites, between 200 and 500 ms post-article onset, correlating article mean ERP amplitude with upcoming noun cloze probability for more (C) and less (B) experienced reader participant groups (as assessed through offline testing). Positive *r*-values (in red) indicate more positive article ERP amplitudes (smaller N400s) with increasing noun cloze (a more N400-like pattern). Negative *r*-values (in blue) indicate increasing ERP positivity with decreasing cloze. At the faster presentation rate, only the group assessed as more experienced readers (C) showed correlation values similar to prediction effects observed in DeLong et al. (2005), albeit correlated with upcoming noun cloze probabilities.

> pattern might have boosted their sensitivity to the articles. However, this may not be the case in experiments where articles do not always occur, or when they do not reliably cue the upcoming noun because they match an adjective instead (e.g. As it's rainy, it's better to go out with a big umbrella). In our view, in natural language settings, articles may not be very reliable cues to upcoming nouns, which means that pre-activation of word form may not be a common phenomenon.

These are reasonable questions but, again, ones that have already been asked and answered. DeLong (2009, dissertation Chapter 4) conducted a variant of the original DeLong et al. (2005) design using sentences that could continue either with article+noun or article+adjective+noun combinations, such as, "The day was breezy so the boy went outside to fly … a kite/an airplane/a helicopter/an orange kite … ". Contrary to Ito et al.'s prediction, there continued to be a reliable article N400 effect [$F(1, 31) = 4.52$, $p = .0416$], see Figure 1(C), although correlation values were lower and only marginally significant compared to the original DeLong et al. (2005) study. These data patterns are inconsistent with Ito et al.'s cuing account.

In conclusion, we concur that replication studies are of critical importance (see Pashler & Wagenmakers, 2012); however, conducting them requires they be done responsibly, with scholarship, caution, and measured interpretation. In this commentary, we have challenged Ito et al.'s approach, describing how they (1) incorrectly and inconsistently attempted to use null results to prove a negative, (2) ignored the larger prediction literature while selectively focusing on one narrow type of prediction (lexical form), yet then went on to draw much broader conclusions by questioning the existence of a more general predictive language mechanism, (3) framed their argument as challenging the view that prediction is necessary for language comprehension, when to our knowledge this straw-person is not widely endorsed by language researchers (contra Huettig & Mani, 2016), (4) reported questionably null prediction findings and failed to analyse prediction as a graded, rather than all-or-nothing, phenomenon, and (5) put forth some interesting, though not novel, proposals about conditions under which prediction effects might not obtain, but framed these as a rejection of prediction rather than just setting some bounds on when it might occur. We maintain that when Ito et al.'s report is properly assessed, there is no failure to replicate, and thus no (new) evidence to question the viability of a predictive mechanism during language comprehension. If Ito et al. had argued that *a/an* article prediction effects cannot be observed under every experimental manipulation, there could be no serious objection. However, well beyond what their data allow, they concluded "prediction effects are perhaps not in fact representative of how people comprehend language in natural settings." In our opinion, this study offers no positive evidence for this conclusion and makes no substantive contribution to the literature on what factors may

matter for prediction and when. In the final sentence of their report, Ito et al. (2016a) write alarmingly about a replication crisis in psycholinguistics. Informed readers must therefore decide for themselves whether the statistically marginal and putatively null article N400 effects are part of the solution or part of the problem, and, more generally, just what has been learned.

Disclosure statement

No potential conflict of interest was reported by the authors.

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

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852–1870. doi:10.1016/j.neuropsychologia.2012.04.011

DeLong, K. A. (2009). *Electrophysiological explorations of linguistic pre-activation and its consequences during online sentence processing* (Doctoral dissertation). Cognitive Science, UC San Diego. b6301658. Retrieved from: http://escholarship.org/uc/item/4q7520sb

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*(12), 631–645. doi:10.1111/lnc3.12093

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

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 1–20. doi:10.1016/S0022-5371(79)90534-6

Frenck-Mestre, C. (2005). Ambiguities and anomalies: What can eye-movements and event-related potentials reveal about second language sentence processing? In J. Kroll & A. de Groot (Eds.), *Handbook of bilingualism* (pp. 268–284). Amsterdam: Elsevier.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. doi:10.1080/23273798.2015.1072223

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*. Advance online publication. 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*. Advance online publication. doi:10.1037/xlm0000315

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

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York, NY: Oxford University Press.

Kutas, M., Moreno, E., & Wicha, N. (2009). Code-switching and the brain. In B. Bullock & A. Toribio (Eds.), *The Cambridge handbook of linguistic code-switching, Cambridge handbooks in linguistics* (pp. 289–306). New York, NY: Cambridge University Press.

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

Mitchell, D. C., & Green, D. W. (1978). The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology*, *30*(4), 609–636. doi:10.1080/14640747808400689

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

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. doi:10.1080/01638530802356463

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi:10.1177/1745691612465253

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. doi:10.1016/j.jml.2012.12.002

Van Berkum, J. J., 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

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., 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., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing

within a written sentence in Spanish. *Cortex*, *39*(3), 483–508. doi:10.1016/S0010-9452(08)70260-0

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. doi:10.1162/0898929041920487

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. doi:10.1016/j.cortex.2015.03.014