Limits on prediction in language comprehension:

A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology

Mante S. Nieuwland[1,5], Stephen Politzer-Ahles[2,10], Evelien Heyselaar[3], Katrien Segaert[3], Emily Darley[4], Nina Kazanina[4], Sarah Von Grebmer Zu Wolfsthurn[4], Federica Bartolozzi[5], Vita Kogan[5], Aine Ito[5,10], Diane Mézière[5], Dale J. Barr[6], Guillaume Rousselet[6], Heather J. Ferguson[7], Simon Busch-Moreno[8], Xiao Fu[8], Jyrki Tuomainen[8], Eugenia Kulakova[9], E. Matthew Husband[10], David I. Donaldson[11], Zdenko Kohút[12], Shirley-Ann Rueschemeyer[12], Falk Huettig[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[3] School of Psychology, University of Birmingham, Birmingham, United Kingdom

[4] School of Experimental Psychology, University of Bristol, Bristol, United Kingdom

[5] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

[6] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom

[7] School of Psychology, University of Kent, Canterbury, United Kingdom

[8] Division of Psychology and Language Sciences, University College London, London, United Kingdom

[9] Institute of Cognitive Neuroscience, University College London, London, United Kingdom

[10] Faculty of Linguistics, Philology & Phonetics; University of Oxford, Oxford, United Kingdom

[11] Department of Psychology, University of Stirling, Stirling, United Kingdom

[12] Department of Psychology, University of York, York, United Kingdom

Corresponding author:

Mante S. Nieuwland, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. E-mail: mante.nieuwland@mpi.nl, phone: +31-24-3521911

ABSTRACT

In current theories of language comprehension, people routinely and implicitly predict upcoming words by pre-activating their meaning, morpho-syntactic features and even their specific phonological form. To date the strongest evidence for this latter form of linguistic prediction comes from a 2005 *Nature Neuroscience* landmark publication by DeLong, Urbach and Kutas, who observed a graded modulation of article- and noun-elicited electrical brain potentials (N400) by the pre-determined probability that people continue a sentence fragment with that word ('cloze'). In a direct replication study spanning 9 laboratories (*N*=334), we failed to replicate the crucial article-elicited N400 modulation by cloze, while we successfully replicated the commonly-reported noun-elicited N400 modulation. This pattern of failure and success was observed in a pre-registered replication analysis, a pre-registered single-trial analysis, and in exploratory Bayesian analyses. Our findings do not support a strong prediction view in which people routinely pre-activate the phonological form of upcoming words, and suggest a more limited role for prediction during language comprehension.

In the last decades, the idea that people routinely and implicitly predict upcoming words during language comprehension turned from a highly controversial hypothesis to a widely accepted assumption. Initial objections to prediction in language were based on a lack of empirical support[1], incompatibility with traditional bottom-up models and contemporary interactive models of language comprehension[2-3], and the projected futility of prediction in a generative system where sentences can continue in infinitely many different ways[4]. Current theories of language comprehension, however, reject such objections and posit prediction as an integral and inevitable mechanism by which comprehension proceeds quickly and incrementally[5-7]. Prediction, the context-based pre-activation of an upcoming word, is thought to occur at all levels of linguistic representation (semantic, morpho-syntactic and phonological/orthographic) and serves to facilitate the word's integration into the unfolding sentence- or discourse-representation. In this line of thought, language is yet another domain in which the brain acts as a prediction machine[8], hard-wired to continuously match sensory inputs with top-down, grammatical or probabilistic expectations based on context and memory.

What promoted linguistic prediction from outlandish and deeply contentious to ubiquitous and somewhat anodyne? One of the key and most compelling pieces of empirical evidence for linguistic prediction to date comes from a landmark *Nature Neuroscience* publication in 2005 by DeLong, Urbach and Kutas[9], whose approach exploited the English rule whereby the indefinite article is phonologically realized as *a* before consonant-initial words and as *an* before vowel-initial words. In their experiment, participants read sentences of varying degree of contextual constraint that led to expectations for a particular consonant- or vowel-initial noun. This expectation was operationalized as a word's cloze probability (cloze), calculated in a separate, non-speeded sentence completion task as the percentage of continuations of a sentence fragment with that word[10]. For example, the sentence ''The day

was breezy so the boy went outside to fly..." is continued with 'a' by 86% of participants, and when presented together with the article 'a', it is continued with 'kite' by 89% of participants. In the main experiment, word-by-word sentence presentation enabled DeLong and colleagues to examine electrical brain activity elicited by articles that were concordant with the highly expected but yet unseen noun ('a' before 'kite'), or by articles that were incompatible with the highly expected noun and heralded a less expected one ('an' before 'airplane'). The dependent measure was the N400 event-related potential (ERP), a negative ERP deflection that peaks at approximately 400 ms after word onset and is maximal at centroparietal electrodes[11]. The N400 is elicited by every word of an unfolding sentence and its amplitude is smaller (less negative) with increasing ease of semantic processing[12]. DeLong et al. found that the N400 amplitude was smaller with increasing cloze probability of the word both at the noun and, critically, at the article. The systematic, graded N400 modulation by article-cloze was taken as strong evidence that participants activated the nouns in advance of their appearance, and that the disconfirmation of this prediction by the less-expected articles resulted in processing difficulty (higher N400 amplitude).

The results obtained with this elegant design warranted a much stronger conclusion than related results available at the time. Previous studies that employed a visual-world paradigm had revealed listeners' anticipatory eye-movements towards visual objects on the basis of probabilistic or grammatical considerations[13-14]. However, predictions in such studies are scaffolded onto already-available visual context, and therefore do not measure purely pre-activation, but perhaps re-activation of word information previously activated by the visual object itself[15]. DeLong and colleagues examined brain responses to information associated with concepts that were not pre-specified and had to be retrieved from long-term memory 'on-the-fly'. Furthermore, DeLong and colleagues were the first to muster evidence for highly specific pre-activation of a word's phonological form, rather than merely its

semantic[16] or morpho-syntactic features[17-18]. Crucially, as their demonstration involved semantically identical articles (function words) rather than nouns or adjectives (content words) that are rich in meaning, the observed N400 modulation by article-cloze is unlikely to reflect difficulty interpreting the articles themselves. And, most notably, DeLong and colleagues were the first to examine brain activity elicited by a range of more- or less-predictable articles, not simply most- versus least-expected. Based on the observed correlation, they argued that pre-activation is not all-or-none and limited to highly constraining contexts, but occurs in a graded, probabilistic fashion, with the strength of a word pre-activation proportional to its cloze probability. Moreover, they concluded that prediction is an integral part of real-time language processing and, most likely, a mechanism for propelling the comprehension system to keep up with the rapid pace of natural language.

DeLong et al.'s study has had an immense impact on psycholinguistics, neurolinguistics and beyond. It is cited by authoritative reviews[19-25] as delivering decisive evidence for probabilistic prediction of words all way up to their phonological form. Moreover, as a demonstration of pre-activation of phonological form (sound) during reading, it is often cited as evidence for 'prediction through production'[6-7,16], the hypothesis that linguistic predictions are implicitly generated by the language production system. To date, DeLong et al. has received a total of 648 citations (Google Scholar), roughly averaging 1 citation per week over the past decade, with an increasing number of citations in each subsequent year. The results also settled an ongoing debate in the neuroscience of language by providing the clearest evidence that the N400 component, which for 25 years had been taken to directly index the high-level compositional processes by which people integrate a word's meaning with its context[26-27], reflected non-compositional processes by which word information is accessed as a function of context.

But how robust are gradient effects of form prediction? In over a decade that has passed since the publication by DeLong and colleagues, that pattern of results has not been successfully replicated, neither directly nor conceptually[28]. In a subsequent study[29], DeLong and colleagues performed two experiments using the same article and noun manipulation but effects of the articles were not reported, only effects of the nouns. In at least two unpublished data sets[30], DeLong and colleagues failed to replicate the correlation between article-N400 and cloze probability. Similar studies with the a/an manipulation and with cloze as categorical (high/low) variable have yielded unclear results[28]. Studies with other pre-nominal manipulations, namely of morpho-syntactic features, also show inconsistent results[17-18,31-32], yielding qualitatively different patterns for the same manipulation in highly similar experiments.

As the tremendous scientific impact of the DeLong et al. findings is at odds with the apparent lack of replication attempts, we here report a direct replication study. Inspired by recent demonstrations for the need for large subject-samples in psychology and neuroscience research[33-34], our replication spanned 9 laboratories each with a sample size equal to or greater than that of the original. Our replication attempt also seeks to improve upon DeLong et al.'s data analysis. Their correlation analysis reduced an initial pool of 2560 data points (32 subjects who each read 80 sentences) to 10 grand-average values, by averaging N400 responses over trials within 10 cloze probability decile-bins (cloze 0-10, 11-20, et cetera), per participant and then averaging over participants, even though these bins held greatly different numbers of observations (for example, the 0-10 cloze bin contained 37.5% of all data). These 10 values were correlated with the average cloze value per bin, yielding numerically high correlation coefficients with large confidence intervals (for example, the Cz electrode showed a statistically significant *r*-value of 0.68 with a 95% confidence interval ranging from 0.09 to 0.92). On the one hand, by discretizing cloze probability into deciles and not distinguishing

various sources of subject, item, and trial-level variation, this analysis potentially compromises power; on the other, treating subjects as fixed rather than random potentially inflates false positive rates, due to the confounding of the overall cloze effect with by-subject variation in the effect[35-36].

In our replication study, we followed two pre-registered analysis routes: a *replication analysis* that duplicated the DeLong et al. analysis, and a *single-trial analysis* that modelled variance at the level of item and subject. We expected to replicate the effect of cloze on noun-elicited N400s[9,12], but this alone would not offer evidence for pre-activation. But observing a reliable effect of cloze on article-elicited N400s in our replication analysis and, in particular, in our single-trial analysis, would constitute powerful evidence for the pre-activation of phonological form during reading.

RESULTS

We first obtained offline cloze probabilities for all target articles and nouns. These values resembled those of the original study (median for articles = 29%, for nouns = 40%; both range 0-100%). In the subsequent ERP experiment, different participants (N=334) read the sentences word-by-word from a computer display at a rate of 2 words/s while we recorded their electrical brain activity at the scalp.

Replication analysis

We sorted the articles and nouns into 10 bins based on each word's cloze probability. For each laboratory, ERPs per bin were averaged first within, then across, participants. The average of the cloze values per bin was then correlated with mean ERP amplitude in the N400 time window (200-500 ms), yielding a correlation coefficient (r-value) per EEG channel. This analysis yielded a very different pattern than DeLong et al. observed (Fig. 1). In no laboratory did article-N400 amplitude become significantly smaller (less negative) as article-cloze probability increased (in fact, in most laboratories the pattern went into the

opposite direction). Only in one laboratory (Lab 2) did the associated p-value of the correlation coefficient dip below 0.05 (uncorrected for multiple comparisons) at a few left-frontal electrodes, not at the central-parietal electrodes where DeLong et al found their N400 effects. Moreover, in 2 laboratories (Labs 3 and 5), a statistically significant effect was observed in the opposite direction, larger (more negative) article-N400 amplitude with increasing cloze probability. For the nouns, the pattern was more similar to the DeLong et al. results. In six laboratories (Lab 2, 3, 4,6, 7, and 9), noun-N400 amplitude at central-parietal or parietal-occipital electrodes became smaller with increasing noun-cloze, and in two other laboratories (Lab 5 and 8) the effects clearly went in the expected direction without reaching statistical significance.

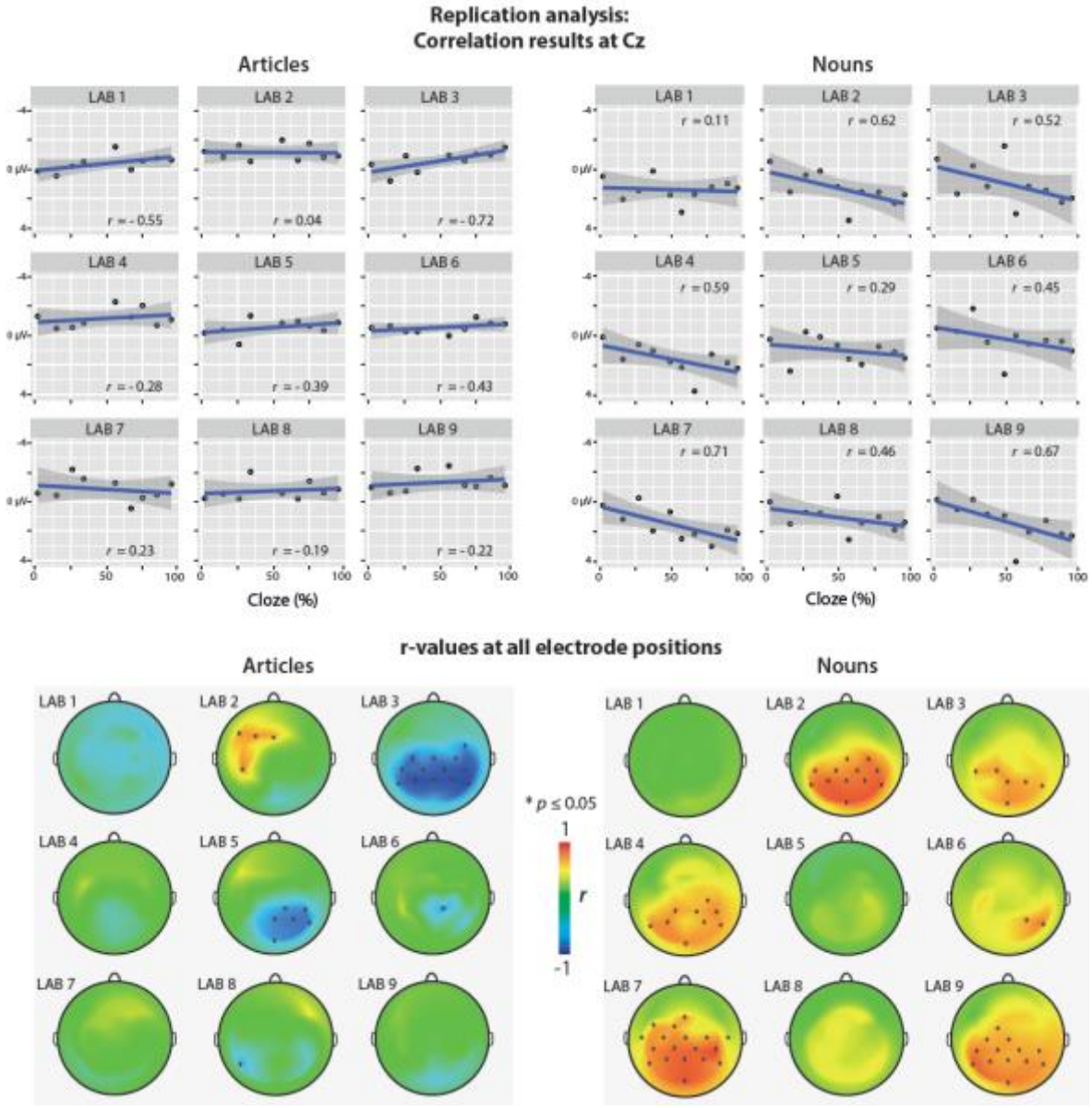

Figure 1. Replication analysis. *Correlations between N400 amplitude and article/noun cloze probability per laboratory. N400 amplitude is the mean voltage in the 200-500 ms time window after word onset. A positive value corresponds to the canonical finding that N400 amplitude became smaller (less negative) with increasing cloze probability. Here and in all further plots, negative voltages are plotted upwards. Upper graph: Scatter plots showing the correlation between cloze and N400 activity at electrode Cz. Lower graph: Scalp distribution of the r-values for each lab. Asterisks (*) indicate electrodes that showed a statistically significant correlation (not corrected for multiple comparisons).*

Single-trial analysis

We first performed baseline correction by subtracting the average amplitude in the 100 ms time window before word onset. Baseline-corrected ERPs for relatively expected and unexpected words and difference waveforms are shown in Fig. 2. Then, for the data pooled across all laboratories, we used linear mixed effects models to regress the N400 amplitude (in a spatiotemporal region of interest selected *a priori* based on the DeLong et al. results) on cloze probability. For the articles, the effect of cloze was not statistically significant at the α=.05 level, β = .29, CI [-.08, .67], $\chi^2(1) = 2.31$, p = .13 (see Fig. 3, left panel), with β referring to the N400 difference in microvolts associated with stepping from 0% to 100% cloze. The effect of cloze on N400 amplitude did not significantly differ between laboratories, $\chi^2(8) = 7.90$, p = .44. For the nouns, however, higher cloze values were strongly associated with smaller N400s, β = 2.22, CI [1.76, 2.69], $\chi^2(1) = 56.50$, $p < .001$ (see Figure 3, right panel). This pattern did not significantly differ between laboratories, $\chi^2(8) = 11.59$, p = .17. The effect of cloze on noun-N400s was statistically different from its effect on article-N400s, $\chi^2(1) = 31.38$, $p < .001$.

Exploratory (i.e., not pre-registered) analyses: We noticed small ERP effects of cloze before article-onset in laboratories 1, 3, 4, 6, 8 and 9, and a slow drift effect of cloze immediately at article onset in laboratory 8 (Supplementary Figures showing all electrodes are available on https://osf.io/eyzaq). An analysis in the 500 to 100 ms time window *before* article-onset indeed revealed a non-significant effect of cloze that resembled the pattern observed *after* article-onset, β = .16, CI [-.07, .39], $\chi^2(1) = 1.82$, p = .18. We therefore performed tests which used longer baseline time windows to better control for pre-article voltage levels, or which used the pre-registered baseline and applied a 0.1 Hz high-pass filter to better control for slow signal drift (while presumably not affecting N400 activity). All three tests reduced the initially observed effect of article-cloze (200 ms baseline, β = .25, CI

[-.12, .62], $\chi^2(1) = 1.35$, $p = .19$; 500 ms baseline, $\beta = .14$, CI [-.25, .53], $\chi^2(1) = 0.46$, $p = .50$; 0.1 Hz filter: $\beta = 0.09$, CI [-.22, .41], $\chi^2(1) = 0.33$, $p = .56$), suggesting that the results obtained with the pre-registered analysis at least partly reflected the effects of slow signal drift that existed before the articles were presented.

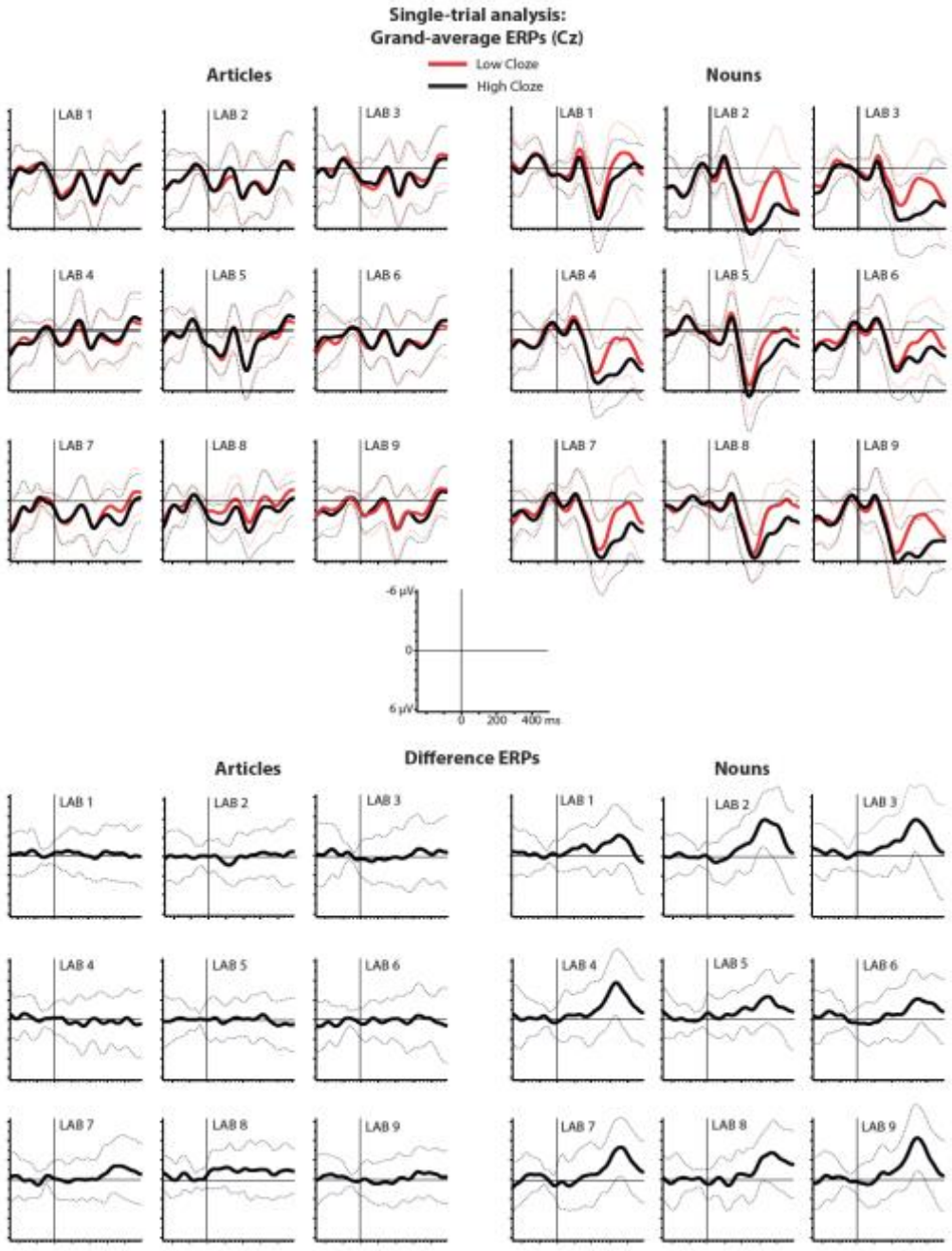

Figure 2. Single-trial analysis. *Grand-average ERPs elicited by relatively expected and unexpected words (cloze higher/lower than 50%) and the associated difference waveforms at electrode Cz. Standard deviations are shown in dotted lines.*

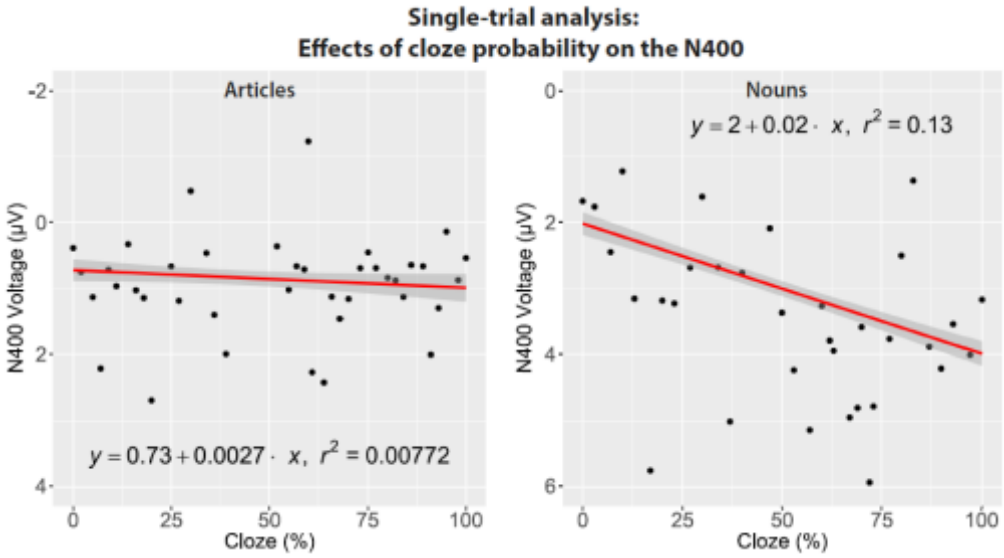

Figure 3. Single-trial analysis*. Relationship between cloze and N400 amplitude as illustrated by the mean N400 values per cloze value, along with the linear fit equation, regression line and 95% confidence interval. N400 amplitude is the average voltage across 6 central-parietal channels (Cz/C3/C4/Pz/P3/P4) in the 200-500 ms window after word onset for each trial. As per the linear fit equation, a change in article cloze from 0 to 100 is associated with a drop in N400 amplitude of 0.27 μV (95% confidence interval: -.04 to .57), whereas a change in noun cloze from 0 to 100 is associated with a drop in N400 amplitude of 2 μV (95% confidence interval: 1.69 to 2.25). We note that these smoothed values, obtained for plotting purposes, differ slightly from the output of the single-trial analysis.*

Exploratory Bayesian analyses

For the articles, our pre-registered analyses yielded non-significant *p*-values, indicating failure to reject the null-hypothesis that cloze has no effect on N400 activity. To better adjudicate between the null-hypothesis (H_0) and an alternative hypothesis (H_1), we performed exploratory Bayes factor analysis for correlations. The obtained Bayes factor quantifies the evidence that there is or is not an effect in the direction reported by DeLong et al. (see Fig. 4). For the articles, this yielded strong evidence for the null-hypothesis, with BF_{01} values up to 32 (at the Cz electrode depicted by DeLong et al., $BF_{01} = 21$), and strongest evidence at the posterior channels. For the nouns, we obtained extremely strong evidence for

the alternative hypothesis, particularly at posterior channels, with BF_{10} values up to 4,807,400 (at Cz, $BF_{10} = 4016$).

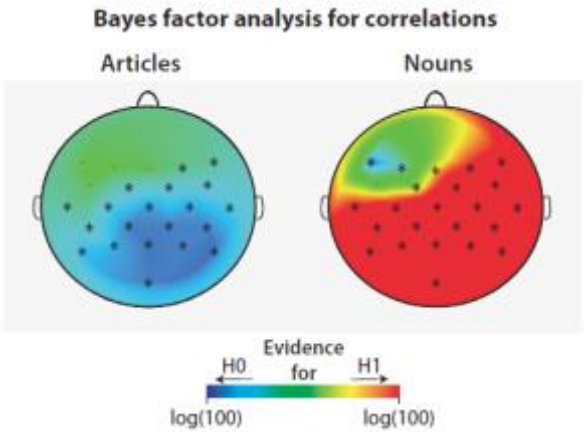

Figure 4. Bayes factor analysis. *Quantification of the obtained evidence for the null-hypothesis (H_0) that N400 is not impacted by cloze, or for the alternative hypothesis (H_1) that N400 is impacted by cloze with the direction of effect reported by DeLong et al. Scalp maps show the common logarithm of the one-sided default Bayes factor for each electrode, capped at log(100) for presentation purposes. Electrodes that yielded at least moderate evidence for or against the null-hypothesis (Bayes factor of ≥ 3) are marked by an asterisk.*

Next, we computed Bayesian mixed-effect model estimates (*b*) and 95% credible intervals (CrI) for our single-trial analyses, using priors based on DeLong et al. In none of our article-analyses did zero lie outside the obtained credible interval, 100 ms baseline: *b* = .31, CrI [-.06 .69]; 200 ms baseline: *b* = .28, CrI [-.11 .64]; 500 ms baseline: *b* = .17, CrI [-.22 .55]; 0.1 Hz filter: *b* = .11, CrI [-.22 .50]. For the nouns, zero was not within the credible interval, *b* = 2.24, CrI [1.77 2.70]. These Bayesian analyses further confirm our failure to replicate the DeLong et al. article-effect and successful replication of the noun-effect.

Control experiment

Lack of a statistically significant, article-elicited prediction effect could reflect a general insensitivity of our participants to the a/an rule. We ruled this out in an additional experiment that followed in the same experimental session. Participants read 80 short

sentences containing the same nouns as the replication experiment, preceded by a correct or incorrect article (e.g., "David found a/an apple..."), presented in the same manner as before. In each laboratory, nouns following incorrect articles elicited a late positive-going waveform compared to nouns following correct articles (see Fig. 5), starting at about 500 ms after word onset and strongest at parietal electrodes. This standard P600 effect[37] was confirmed in a single-trial analysis, $\chi^2(1) = 83.09$, $p < .001$, and did not significantly differ between labs, $\chi^2(8) = 8.98$, $p = .35$.

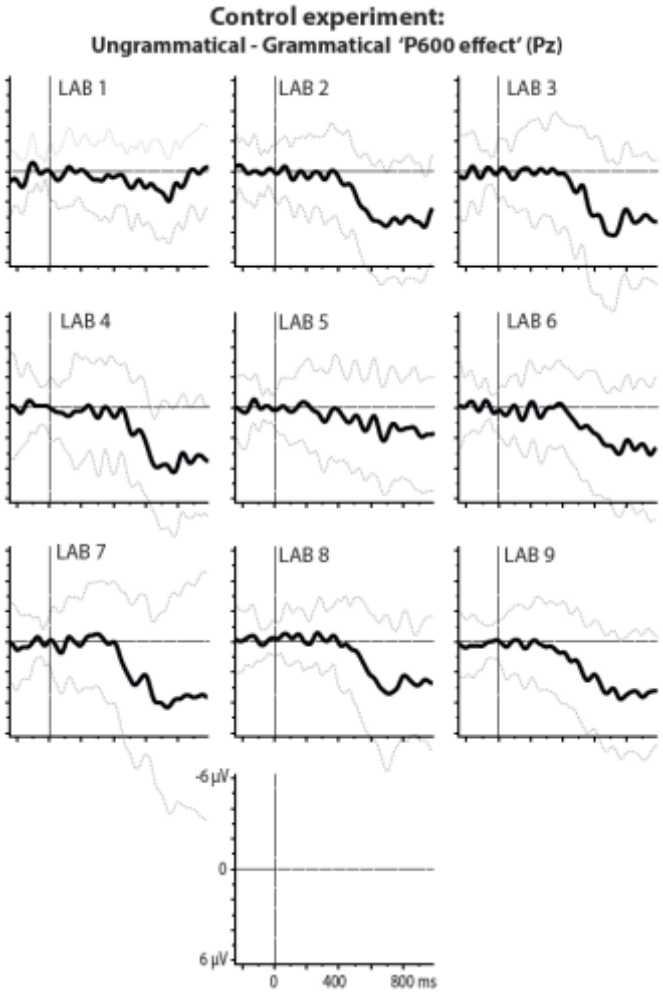

Figure 5. Control experiment. *P600 effects at electrode Pz per lab associated with flouting of the English a/an rule. Plotted ERPs show the grand-average difference waveform and standard deviation for ERPs elicited by ungrammatical expressions ('an kite') minus those elicited by grammatical expressions ('a kite').*

DISCUSSION

In a landmark study, DeLong, Urbach and Kutas observed a statistically significant, graded modulation of article- and noun-elicited electrical brain potentials (N400) by the pre-determined probability that people continue a sentence fragment with that word (cloze). They concluded that people probabilistically pre-activate upcoming words to a high level of detail, including whether a word starts with a consonant or vowel. Our *direct replication* study spanning 9 laboratories failed to replicate the crucial effect of cloze on article-elicited N400 activity, but successfully replicated its effect on noun-elicited N400 activity. This pattern of failure and success was observed in a pre-registered replication analysis that duplicated the original study's analysis, a pre-registered single-trial analysis that modelled variance at the level of item and subject, and exploratory Bayesian analyses. A control experiment confirmed that our participants were capable of applying the a/an rule to the nouns used in the replication experiment.

Our findings carry important theoretical implications by effectively removing a crucial cornerstone of the 'strong prediction view' held by current theories of language comprehension[5-7]. The strong prediction view entails two key claims. The first is that people pre-activate words at all levels of representation in a routine and implicit (i.e., non-strategic) fashion. Pre-activation is thus not limited to a word's meaning, but includes its grammatical features and even what it looks and sounds like. This would put language on a par with other cognitive systems that attempt to predict the inputs to lower-level ones[8]. The second claim is that pre-activation occurs at all levels of contextual support and gradually increases in strength with the level of contextual support. When contextual support for a specific word is high, like at a 100% cloze value, the word's form and meaning is strongly pre-activated. When contextual support for a word is low, like when it is one amongst 20 words each with a 5% cloze value, pre-activation is distributed across multiple potential continuations.

However, even then, a word's form and meaning are pre-activated, just weakly so. The strength of pre-activation is probabilistic, that is, linked to estimated probability of occurrence.

DeLong and colleagues, and many scientists with them[19-25], took their results as the only evidence to support both these claims. Indeed, theirs was – and still is - the only study to date that measured pre-activation at the prenominal articles *a* and *an* that do not differ in their semantic or grammatical content, and the only study that observed a graded relationship between cloze and N400 activity across a range of low- and high-cloze words, rather than merely a difference between low- and high-cloze words. Given that the use of these articles depends on whether the next word starts with a vowel or consonant, their results thus seemed like powerful evidence that participants probabilistically pre-activated the initial sound of upcoming nouns.

However, we convincingly show that there is no statistically significant effect of cloze on article-elicited N400 activity, using a sample size more than ten times that of the original, and a statistical analysis that better accounts for sources of non-independence than the correlation approach. If an effect of cloze on article-N400s exists at all, it is practically irrelevant for a theory of language comprehension, because it would be so small that it cannot be reliably detected even in an expansive multi-laboratory approach, let alone in the typical sample size in psycholinguistic and neurolinguistic experiments (roughly, $N = 30$). In contrast, we observed a strong and statistically significant effect of cloze on noun-elicited activity in all our analyses, overwhelmingly replicating that finding from the original study alongside others[12]. Where does this pattern of failure and success leave the strong prediction view?

Following the experimental logic of DeLong et al, we can conclude that people do not routinely pre-activate the initial phoneme of an upcoming word, or perhaps any other word

form information, but do pre-activate its meaning. Without pre-activation of the initial phoneme, the specific instantiation of the article does not cause people to revise their prediction about the meaning of the upcoming noun, thus lacking any impact on processing. Crucially, this conclusion is incompatible with the strong prediction view, because it suggests that pre-activation does not occur to the level of detail that is often assumed[19-25]. Our results are also incompatible with an alternative interpretation[38-39] of the DeLong et al. findings that people predict the article itself together with the noun, and they pose a serious challenge to the theory that comprehenders predict upcoming words, including their initial phonemes, through implicit production[6-7,16]. Crucially, the idea that prediction is probabilistic, rather than all-or-none, is now questionable, given that there is no other published report of a pre-activation gradient. Although other studies have claimed prediction of form[38] or a prediction gradient[40], no such study has indisputably demonstrated pre-activation, i.e., effects occurring *before* the noun. Effects that are observed upon, rather than before the noun, do not purely index pre-activation but index a mixture of attentional and memory retrieval processes instigated by the noun itself[15,22,32]. Therefore, there is currently no clear evidence to support probabilistic pre-activation of a noun's phonological form.

Our results, however, do not necessarily exclude phonological form pre-activation, and we temper our conclusion with a caveat stemming from the a/an manipulation. For this manipulation to 'work', people must specifically predict the initial phoneme of the next word, and revise this prediction when faced with an unexpected article. However, because articles are only diagnostic about the next word, not about whether the expected noun appears at all, an unexpected article does not disconfirm the upcoming noun, it merely signals that another word would come first (e.g., 'an old kite'). This opens up explanations for why the a/an manipulation 'fails'. Maybe people don't predict the noun to follow immediately, but at a later point; the unexpected article then does not evoke a change in prediction. Predictions

about a specific position may be disconfirmed too often in natural language to be viable. This idea is supported by corpus data (Corpus of Contemporary American English and British National Corpus,), showing a mere 33% probability that a/an is followed by a noun. Alternatively, people predict the noun to come next, but only revise their prediction about its position while retaining the prediction about its meaning. So perhaps a revision of the predicted meaning, not the position, is what modulates N400 activity. This is not unreasonable given the well-established association between N400 activity and processing meaning[11-12,21]. In both of these hypothetical scenarios, people do not revise their prediction about the upcoming noun's meaning when they don't have to. This raises an important challenge to formalize or model linguistic prediction, as current endeavors[6,7,42] assume a sequence of predictions limited to each subsequent word.

Our results can be straightforwardly reconciled with effects reported for other pre-nominal manipulations, such as those of Dutch or Spanish article-gender[17-18,31-32]. Unlike a/an articles, gender-marked articles can immediately disconfirm the noun, because article- and noun-gender agrees regardless of intervening words (e.g., the Spanish article 'el' heralds a masculine noun). Revising the prediction about the noun presumably results in a semantic processing cost, thereby modulating N400 activity. Although gender-marked articles do not consistently incur the exact same type of effect[17-18,31-32] and have only been observed at very high cloze values, previous studies suggest that a noun's grammatical gender can be pre-activated along with its meaning. Compared to this gender-manipulation, the English a/an manipulation tests a stronger version of the prediction view, namely that people predict which word comes next *and*, once disconfirmed, revise their prediction about this word altogether.

What do our results say about prediction during natural language processing? Like the conclusions by DeLong et al., ours are limited by the generalization from language comprehension in a laboratory setting. On one hand, a rich conversational or story context

may enhance predictions of upcoming words, and listeners may be more likely to pre-activate the phonological form of upcoming words than readers. On the other hand, our laboratory setting offered particularly good conditions for prediction of the next word's initial sound to occur. Each article was always immediately followed by a noun, unlike in natural language. Moreover, compared to natural reading rates our word presentation rate was slow, which may facilitate predictive processing[38]. In natural reading, articles are hardly fixated and often skipped[41]. In short, arguments can be made both for and against phonological form prediction in natural language settings, and novel avenues of experimentation are needed to settle this issue.

To conclude, we failed to replicate the main result of DeLong et al., a landmark study published more than ten years ago that has not been replicated since. Our findings thus highlight the importance of direct replication attempts in the neurosciences and language sciences, disciplines that are subject to the same circumstances that gave rise to the replication crisis in psychology and elsewhere[33]. Our findings also challenge one of the pillars of the 'strong prediction view' in which people routinely and probabilistically pre-activate information at all levels of linguistic representation, including phonological form information such as the initial phoneme of an upcoming noun. Consequently, there is currently no convincing evidence that people pre-activate the phonological form of an upcoming noun, and we take our findings to suggest a more limited role for prediction during language comprehension.

REFERENCES

[1] Zwitserlood, P. The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition* **32**, 25-64 (1989).

[2] Kintsch, W. The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* **95**, 163-182 (1988).

[3] Marslen-Wilson, W. & Tyler, L.K. The temporal structure of spoken language understanding. *Cognition* **8**, 1-71 (1980).

[4] Jackendoff, R. *Foundations of language: brain, meaning, grammar, evolution* (Oxford University Press, Oxford; New York, 2002).

[5] Altmann, G.T. & Mirkovic, J. Incrementality and Prediction in Human Sentence Processing. *Cogn. Sci.* **33**, 583-609 (2009).

[6] Dell, G.S. & Chang, F. The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Phil. Trans. Royal Soc. London Series B, Biol. Sci.* **369**, 20120394 (2014).

[7] Pickering, M.J. & Garrod, S. An integrated theory of language production and comprehension. *Behav. Brain Sci.* **36**, 329-347 (2013).

[8] Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181-204 (2013).

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

[10] Taylor, W.L. "Cloze Procedure": A new tool for measuring readability. *Journalism Quart.* **30**, 415-433 (1953).

[11] Kutas, M. & Hillyard, S.A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* **207**, 203-205 (1980).

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

[13] Altmann, G.T. & Kamide, Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* **73**, 247-264 (1999).

[14] Dahan, D. & Tanenhaus, M.K. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *J. Exp. Psychol. Learn. Mem. Cog.* **30**, 498-513 (2004).

[15] Huettig, F. Four central questions about prediction in language processing. *Brain Res.* **1626**, 118-135 (2015).

[16] Federmeier, K.D. & Kutas, M. A rose by any other name: Long-term memory structure and sentence processing. *J. Mem. Lang.* **41**, 469-495 (1999).

[17] Van Berkum, J.J., Brown, C.M., Zwitserlood, P., Kooijman, V. & Hagoort, P. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cog.* **31**, 443-467 (2005).

[18] Wicha, N.Y., Moreno, E.M. & Kutas, M. Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* **16**, 1272-1288 (2004).

[19] Altmann, G.T. & Mirkovic, J. Incrementality and Prediction in Human Sentence Processing. *Cogn. Sci.* **33**, 583-609 (2009).

[20] Pickering, M.J. & Clark, A. Getting ahead: forward models and their place in cognitive architecture. *Trends. Cogn. Sci.* **18**, 451-456 (2014).

[21] Lau, E.F., Phillips, C. & Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nature Rev. Neurosci* **9**, 920-933 (2008).

[22] Hagoort, P. The core and beyond in the language-ready brain. Manuscript in press at *Neurosci. & Biobehav. Rev*. (2017).

[23] Federmeier, K.D. Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* **44**, 491-505 (2007).

[24] Christiansen, M.H. & Chater, N. The Now-or-Never bottleneck: A fundamental constraint on language. *Behav. Brain Sci.* **39**, e62 (2016).

[25] Kuperberg, G.R. & Jaeger, T.F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32-59 (2016).

[26] Brown, C. & Hagoort, P. The processing nature of the n400: evidence from masked priming. *J. Cogn. Neurosci.* **5**, 34-44 (1993).

[27] van Berkum, J.J., Hagoort, P. & Brown, C.M. Semantic integration in sentences and discourse: evidence from the N400. *J. Cogn. Neurosci.* **11**, 657-671 (1999).

[28] Ito, A., Martin, A. E., & Nieuwland, M. S. How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Lang. Cogn. Neurosci*. 1-12 (2016).

[29] DeLong, K.A., Quante, L. & Kutas, M. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia* **61**, 150-162 (2014).

[30] DeLong, K. A. (2009). *Electrophysiological explorations of linguistic pre-activation and its consequences during online sentence processing*. Doctoral dissertation. San Diego: University of California.

[31] Otten, M., Nieuwland, M.S. & Van Berkum, J.J. Great expectations: specific lexical anticipation influences the processing of spoken language. *BMC Neurosci.* **8**, 89 (2007).

[32] Otten, M. & Van Berkum, J. Discourse-based word anticipation during language processing: Prediction or priming? *Disc. Processes* **45**, 464-496 (2008).

[33] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

[34] Button, K.S., *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365-376 (2013).

[35] Clark, H.H. Language as Fixed-Effect Fallacy - Critique of Language Statistics in Psychological Research. *J. Verb. Learn. Verb. Behav.* **12**, 335-359 (1973).

[36] Barr, D.J., Levy, R., Scheepers, C. & Tily, H.J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255-278 (2013).

[37] Osterhout, L. & Holcomb, P.J. Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* **31**, 785-806 (1992).

[38] Ito, A., Corley, M., Pickering, M.J., Martin, A.E. & Nieuwland, M.S. Predicting form and meaning: Evidence from brain potentials. *J. Mem. Lang.* **86**, 157-171 (2016).

[39] Van Petten, C. & Luka, B.J. Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* **83**, 176-190 (2012).

[40] Smith, N.J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302-319 (2013).

[41] O'Regan, K. Saccade size control in reading: evidence for the linguistic control hypothesis. *Percept. Psychophys*. **25**, 501–509 (1979).

ONLINE METHODS

Experimental design and materials

The original materials from DeLong et al. were adapted from American to British spelling and underwent a few minor changes to ensure their suitability for British participants. The complete set of materials and the list of changes to the original materials are available online (Supplementary Table 1 and 2). The materials were 80 sentence contexts with two possible continuations each: a more or less expected indefinite article + noun combination. The noun was followed by at least one subsequent word. All article + noun continuations were grammatically correct. Within each participant, each article + noun combination served once as the more expected continuation and the other time as the less expected continuation, in different contexts. We divided the 160 materials in two lists of 80 sentences such that each list contained each noun only once. Each participant was presented with only one list (thus, each context was seen only once). One in four sentences was followed by a yes/no comprehension question, 86% of which, on average, were answered correctly by participants.

Article-cloze and noun-cloze ratings were obtained from a separate group of student volunteers from the University of Edinburgh who did not participate in the ERP experiment. We obtained article-cloze ratings from 44 participants for 80 sentence contexts truncated before the critical article. Noun-cloze ratings were obtained by first truncating the sentences after the critical articles, and presenting two different, counterbalanced lists of 80 sentences to 30 participants each, such that a given participant only saw each sentence context with the expected or the unexpected article.

Participants

Participants were students from the University of Birmingham, Bristol, Edinburgh, Glasgow, Kent, Oxford, Stirling, York, or from the participant pool of University College

London or Oxford University, who received cash or course credit for taking part in the ERP experiment. All participants (N = 356; 222 women) were right-handed, native English speakers with normal or corrected-to-normal vision, between 18–35 years (mean, 19.8 years), free from any known language or learning disorder. Eighty-nine participants reported a left-handed parent or sibling. Participant information and EEG recording information per laboratory is available online (Supplementary Table 3).

Procedure

After giving written informed consent, participants were tested in a single session, with written sentences presented in the center of a computer display, one word at a time (200 ms duration, 500 ms stimulus onset asynchrony). Participants were instructed to read sentences for comprehension and answer yes/no comprehension questions by pressing hand-held buttons. The electroencephalogram (EEG) was recorded from minimally 32 electrodes.

The replication experiment was followed by a control experiment, which served to detect sensitivity to the correct use of the a/an rule in our participants. Participants read 80 relatively short sentences (average length 8 words, range 5-11) that contained the same critical words as the replication experiment, preceded by correct or incorrect articles. As in the replication experiment, each critical word was presented only once, and was followed by at least one more word. All words were presented at the same rate as the replication experiment. There were no comprehension questions in this experiment. After the control experiment, participants performed a Verbal Fluency Test and a Reading Span test; the results from these tests are not discussed here.

Data processing

Data processing was performed in BrainVision Analyzer 2.1 (Brain Products, Germany). We performed one pre-registered replication analysis that followed the DeLong et al. analysis as closely as possible and one pre-registered single-trial analysis (Open Science Framework,

https://osf.io/eyzaq). First, we interpolated bad channels from surrounding channels, and down-sampled to a common set of 22 EEG channels per laboratory which were similar in scalp location to those used by DeLong et al. For one laboratory that did not have all the selected 22 channels, 12 virtual channels were created using topographic interpolation by spherical splines. We then applied a 0.01-100 Hz digital band-pass filter (including 50 Hz Notch filter), re-referenced all channels to the average of the left and right mastoid channels (in a few participants with a noisy mastoid channel, only one mastoid channel was used), and segmented the continuous data into epochs from 500 ms before to 1000 ms after word onset. We then performed visual inspection of all data segments and rejected data with amplifier blocking, movement artifacts, or excessive muscle activity. Subsequently, we performed independent component analysis[42], based on a 1-Hz high-pass filtered version of the data, to correct for blinks, eye-movements or steady muscle artefacts. After this, we automatically rejected segments containing a voltage difference of over 120 µV in a time window of 150 ms or containing a voltage step of over 50 µV/ms. Participants with fewer than 60 article-trials or 60 noun-trials were removed from the analysis, leaving a total of 334 participants (range across laboratories 32-42) with, on average, 77 article-trials and 77 noun-trials.

Replication analysis

We applied a 4th-order Butterworth band-pass filter at 0.2-15 Hz to the segmented data, averaged trials per participant within 10% cloze bins (0-10, 11-20, etc. until 91-100), and then averaged the participant-averages separately for each laboratory. Because the bins did not contain equal numbers of trials (the intermediate bins contained fewest trials), not all participants contributed a value for each bin to the grand average per laboratory. For nouns and articles separately, and for each EEG channel, we computed the correlation between ERP amplitude in the 200-500 ms time window per bin with the average cloze probability per bin.

Single-trial analysis

We did not apply the 0.2-15 Hz band-pass filter, but we performed baseline-correction by, for each trial, subtracting the mean voltage of the -100 to 0 ms time window from the data. This common procedure corrects for spurious voltage differences before word onset, generating confidence that observed effects are elicited by the word rather than differences in brain activity that already existed before the word. Baseline correction is a standard procedure in ERP research[43], and often used in the Kutas Cognitive Electrophysiology Lab, but was not used in DeLong et al. The alternative approach taken by DeLong and colleagues, applying a strong filter (0.2-15 Hz) to the data, carries the risk of filter-induced data distortions[44]. Here, we did not apply this filter but we opted for baseline correction using the 100 ms pre-stimulus baseline period. We based this procedure on a review of the published work from the Kutas Cognitive Electrophysiology Lab, which found that the 100 pre-stimulus baseline period was most often used in similar studies.

Instead of averaging N400 data for subsequent statistical analysis, we performed linear mixed effects model analysis[45] of the single-trial N400 data, using the "lme4" package[46] in the R software[47]. This approach simultaneously models variance associated with each subject and with each item. Using a spatiotemporal region-of-interest approach based on the DeLong et al. results, our dependent measure (N400 amplitude) was the average voltage across 6 central-parietal channels (Cz/C3/C4/Pz/P3/P4) in the 200-500 ms window for each trial. Scripts and data are publicly available on https://osf.io/eyzaq.

For articles and nouns separately, we used a maximal random effects structure as justified by the design[36], which did not include random effects for 'laboratory' as there were only 9 laboratories, and laboratory was not a predictor of theoretical interest. Z-scored cloze was entered in the model as a continuous variable that had two possible values for each item (corresponding to relatively expected and unexpected words), and laboratory was entered as a deviation-coded categorical variable. We tested the effects of 'laboratory' and 'cloze' through

model comparison with a χ^2 log-likelihood test. We tested whether the inclusion of a given fixed effect led to a significantly better model fit. The first model comparison examined laboratory effects, namely whether the cloze effect varied across laboratories (cloze-by-laboratory interaction) or whether the N400 magnitudes varied over laboratory (laboratory main effect). If laboratory effects were nonsignificant, we dropped them from the analysis to simplify interpretation. For the articles and nouns separately, we compared the subsequent models below. Each model included the random effects associated with the fixed effect 'cloze'[36]. All output β estimates and 95% confidence intervals (CI) were transformed from z-scores back to raw scores, and then back to the 0-100% cloze range, so that the voltage estimates represent the change in voltage associated with a change in cloze probability from 0 to 100.

Model 1: N400 ~ cloze * laboratory + (cloze | subject) + (cloze | item)

Model 2: N400 ~ cloze + laboratory + (cloze | subject) + (cloze | item)

Model 3: N400 ~ cloze + (cloze | subject) + (cloze | item)

Model 4: N400 ~ (cloze | subject) + (cloze | item)

We also tested the differential effect of cloze on article-ERPs and on noun-ERPs by comparing with and without an interaction between cloze and the deviation-coded factor 'wordtype' (random correlations were removed for the models to converge)

Model 1: N400 ~ cloze * wordtype + (cloze * wordtype || subject) + (cloze * wordtype || item)

Model 2: N400 ~ cloze + wordtype + (cloze * wordtype || subject) + (cloze * wordtype || item)

Analysis of the control experiment involved a comparison between a model with the categorical factor 'grammaticality' (grammatical/ungrammatical) and a model without. Our

dependent measure (P600 amplitude) was the average voltage across 6 central-parietal channels (Cz/C3/C4/Pz/P3/P4) in the 500-800 ms window for each trial.

Model 1: P600 ~ grammaticality + (grammaticality | subject) + (grammaticality | item)

Model 2: P600 ~ (grammaticality | subject) + (grammaticality | item)

Exploratory single-trial analyses

We performed an exploratory analysis in the 500 to 100 ms time window *before* the article, using the originally (100 to 0 ms) baselined data, using Model 3 and 4 from the article-analysis. We took this window to cover the full pre-stimulus section of the data epochs not included in the baseline time window. We then performed exploratory analyses with longer (200 ms or 500 ms) pre-articles baselines. These baseline windows were taken because after the 100 ms pre-stimulus time window, these windows were used most frequently in the Kutas laboratory. We also performed an exploratory analysis with the original baseline but an additional 0.1 Hz high-pass filter applied before baseline correction. We used this filter because it is frequently used in the Kutas laboratory and removes slow signal drift without impacting N400 activity (which has a higher-frequency spectrum)[43-44].

Exploratory Bayesian analyses

Supplementing the Replication analysis, we performed a Bayes factor analysis for correlations using as prior the direction of the effect reported in the original study[48]. This test was performed for each electrode separately, after collapsing the data points from the different laboratories. Because we had no articles in the 40-50 % cloze bin, there was a total of 9 and 10 data points per bin for the articles and nouns, respectively. Our analysis was based on a uniform prior distribution. A Bayes factor between 3 and 10 is considered moderate evidence, between 10-30 is considered strong evidence, 30-100 is very strong evidence, and values over 100 are considered extremely strong evidence.

Supplementing the single-trial analyses, we performed Bayesian mixed-effects model analysis using the brms package for R[49], which fits Bayesian multilevel models using the Stan programming language[50]. We used a prior based on the Delong et al. observed effect size at Cz for a difference between 0% cloze and 100% cloze (1.25 μV and 3.75 μV for articles and nouns, respectively) and a prior of zero for the intercept. Both priors had a normal distribution and a standard deviation of 0.5 (given the a priori expectation that average ERP voltages in this window generally fluctuate on the order of a few microvolts; note that these units are expressed in terms of the z-scored cloze values, rather than the original cloze values, such that μ for the cloze prior was actually 0.45, which corresponds to a raw cloze effect of 1.25). We computed estimates and 95% credible intervals for each of the mixed-effects models we tested, and transformed these back into raw cloze units. The credible interval is the range of values of which one can be 95% certain that it contains the true effect, given the data, priors and the model.

METHOD REFERENCES

[42] Jung, T.P., *et al.* Removing electroencephalographic artifacts by blind source separation. *Psychophysiology.* **37**, 163-178 (2000).

[43] Luck, S. J. *An introduction to the event-related potential technique* (MIT press, Cambridge MA, 2014).

[44] Tanner, D., Morgan- Short, K., & Luck, S. J. How inappropriate highpass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology.* ***52***, 997-1009 (2015).

[45] Baayen, R.H., Davidson, D.J. & Bates, D.M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390-412 (2008).

[46] Bates, D., Maechler, M., Bolker, B., & Walker, S. lme4: Linear mixed-effects models using Eigen and S4. *R package version*, **1** (2014).

[47] R CoreTeam, R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.URL ⟨http://www.R-project.org/⟩. (2014)

[48] Wagenmakers, E.J., Verhagen, J. & Ly, A. How to quantify the evidence for the absence of a correlation. *Behav. Res. Meth.* **48**, 413-426 (2016).

[49] Buerkner, P.C. brms: Bayesian Regression Models using Stan. *R package version* **1.4.0** (2016).

[50] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.14.1. http://mc-stan.org. (2016)

Author contributions:

M.S.N. and F.H. designed the research, M.S.N., D.J.B., G.R., and S.P.-A. planned the analysis. E.H., E.D., S.V.G.Z.W., F.B., V.K., A.I., S.B.-M., Z.F., E.K., S.P-A., and Z.K. collected data. M.S.N., K.S., N.K., G.R., H.J.F., J.T., E.M.H., D.I.D., and S.R supervised data collection. M.S.N. and S.P.-A. analyzed the data. M.S.N. drafted the manuscript and received comments from S.P.-A., N.K., K.S., D.J.B., H.J.F., E.M.H, and F.H.

Acknowledgements:

This work was partly funded by ERC Starting grant 636458 to H.J.F.