An exploratory data analysis of word form prediction during word-by-word reading

Thomas P. Urbach[a,1], Katherine A. DeLong[a], Wen-Hsuan Chan[a], and Marta Kutas[a,b]

[a] Department of Cognitive Science, University of California, San Diego, La Jolla, CA, 92037 ; [b] Department of Neurosciences, University of California, San Diego, La Jolla, CA, 92037

This manuscript was compiled on August 11, 2020

In 2005 we reported evidence indicating that upcoming phonological word forms, e.g., *kite* vs. *airplane*, were predicted during reading. We recorded brainwaves (EEG) as people read word-by-word and then correlated the predictability in context of indefinite articles that preceded nouns (*a kite* vs. *an airplane*) with the average event-related brain potentials (ERPs) they elicited (DeLong, Urbach, and Kutas, 2005). Amid a broader controversy about the role of word form prediction in comprehension, those findings were recently challenged by a failed putative direct replication attempt (Nieuwland, et al., 2018: 9 labs, 1 experiment, 2.6e4 observations). To better understand the empirical justification for positing an association between prenominal article predictability and scalp potentials, we conducted a wide-ranging exploratory data analysis (EDA), pooling our original data with extant data from two followup studies (1 lab, 3 experiments, 1.2e4 observations). We modeled the time course of article predictability in the single-trial data by fitting linear mixed-effects regression (LMER) models at each time point and scalp location spanning a 3 second interval before, during, and after the article. Model comparisons based on Akiake Information Criteria (AIC) and slope regression ERPs (rERPs, Smith and Kutas, 2015) provide substantial empirical support for a small positive association between article predictability and scalp potentials approximately 300–500 ms after article onset, predominantly over bilateral posterior scalp. We think this effect may reasonably be attributed to prediction of upcoming word forms.

language | prediction | EEG | rERP | EDA

Psycholinguistic theories of language comprehension generally endorse the near immediate "incremental" construction of structured representations of meaning, as words, phrases, sentences, and discourses rapidly unfold over time (1). New information must be integrated with this evolving semantic representation and some accounts further posit predictive or preparatory mechanisms that facilitate processing and help the system keep up with the input (2–4). The hypothesis that the comprehension system actively predicts is difficult to test experimentally. The challenge is to find evidence of predictive processing that cannot plausibly be attributed to rapid integration. For instance, given a sentence context like, *The day was breezy so the boys went outside to fly* ___, knowledge of the world and English make some continuations more predictable (*a kite*) and others less so (*an airplane*). It is possible that the supporting context leads the processor to predict (anticipate, expect) the word *kite* before it arrives, in which case on-line measures sensitive to experimental manipulations of processing difficulty, e.g., self-paced reading times, eye movements, event-related brain potentials (ERPs) and magnetic fields (ERFs), might show an experimental effect in the expected direction, i.e., faster reading times, shorter gaze durations, or reduced N400 ERP/Fs for *kite* vs. *airplane*. However, if the effects observed *at* these nouns could with equal justification be attributed to violated predictions or integration difficulty (or both), these findings are compatible with, but do not constitute strong evidence for prediction, and parsimony favors integration mechanisms alone which are necessary on any account.

The crux of the experimental challenge is time: strong tests that information is *pre*-dicted come from measurements made *before* it actually arrives. Seminal laboratory studies measuring eye-movements while listening to meaningful sentences in a controlled visual environment (5–7), found that people tended to glance at mentioned objects quickly or even prior to hearing a likely word, indicating rapid language-driven anticipation of upcoming semantic or conceptual content. To date, the clearest evidence for prediction of specifically linguistic information comes from paradigms that recruit sequential dependencies wherein one type of grammatical element such as a word or morphological marking regularly co-occurs with another element. The seminal ERP studies (8, 9), were conducted by Wicha, Bates, Moreno, and Kutas using grammatical gender agreement between indefinite articles and nouns in Spanish, e.g., feminine *una canasta* ("a basket") vs. masculine *un costal* ("a sack"). If a Spanish sentence is likely to continue about a

Significance Statement

Complex biological systems do not merely react, they anticipate. In 2005, the human language comprehension system was considered an exception. We concluded not, based on our recordings of electrical brain activity measured before the critical words arrived during sentence reading, described in a now widely cited report. This, and the emergence of the "statistical crisis" in psychology led to a large-scale replication attempt that failed. This prompted us to revisit the issue by analyzing our original data and two replication-extensions with an exploratory data analysis (EDA) approach, enabled by advances in scientific computing technology. Our original conclusion was supported: brains can anticipate specific upcoming words. We offer this as a case study in EDA for cognitive neurophysiology, more generally.

Author contributions: K.A.D., M.K., and T.P.U. designed EEG Experiments 1 and 2; K.A.D. and M.K. designed EEG Experiment 3. K.A.D. constructed the materials and conducted Normative Experiments 1-6 and EEG Experiments 1-3. W.C. analyzed data and contributed analytic tools. T.P.U. suggested the design for EEG Experiment 1, contributed analytic tools, conducted the secondary analyses, and wrote the report. M.K., K.A.D, and W.C., contributed substantive editorial revisions. T.P.U. is responsible for errors.

The authors declare no conflict of interest. This report describes de novo secondary analyses of extant behavioral and electrophysiological data; these findings have not previously been published in any form. Materials, data not related to an identifiable natural person, and reproducible analyses are available in the Open Science Foundation repository OSF: UDCK.

[1] To whom correspondence may be addressed. E-mail: turbach@ucsd.edu

basket, the corresponding indefinite article is likely to be *una* not *un*, and vice versa if the likely continuation is about a sack. Since the two forms of the indefinite article have the same meaning ("some singular thing"), they should be equally easy or difficult to integrate. Wicha et al. recorded electrical brain potentials at the scalp (electroencephalogram, EEG) as people read sentences word-by-word on a computer screen, and found small differences between the average ERPs elicited by articles that were compatible vs. incompatible with the grammatical gender of the likely continuation. These effects varied with the particulars of the experimental design: incompatible articles elicited an N400-like relative negativity when the referent of the likely noun was depicted with a line drawing (8, 10) and a relative positive deflection around 500–700 ms when the continuations were orthographic words (9). With other lexical variables controlled by the experimental design, the difference between *un* and *una* is plausibly attributed to a mismatch between the grammatical gender of the article and the gender of the likeliest continuation, indicating that the continuation had been predicted before it was encountered.

Subsequent studies have used related sequential dependency designs to probe other languages for evidence of prediction, e.g., via case-marking in Dutch (11), grammatical gender in Dutch (12–14, but see 15), Polish (16), and German (17). For these types of experimental designs, the nature of the linguistic dependency constrains the inferences that can be drawn about what information is anticipated (discussed in 3, 18). English does not mark grammatical gender or case agreement on nouns but does attest a phonological dependency between alternate forms of the indefinite article *a* which precedes consonant-sound-initial words and *an* which precedes vowel-sound-initial words: *a kite* vs. *an airplane*. We recruited this sequential dependency in previous work (19, hereafter, DUK05), recording scalp potentials while people read sentences like, *The day was breezy so the boys went outside to fly [a kite/an airplane] in the park.*, one word at a time on a computer screen. We observed a positive correlation between the predictability, in context, of the indefinite articles that preceded the nouns *a kite* vs. *an airplane* and the average ERPs they elicited 200–500 ms over bilateral central and posterior scalp. Since the *a/an* alternation depends on the initial speech sound of the next word, we took the systematic association between the ERP amplitude and offline article cloze probability to suggest "that individuals can use linguistic input to pre-activate representations of upcoming words in advance of their appearance" (19, p. 1119), and "Our observation of an ERP expectancy effect at the article leads us to conclude that predictions can be for specific phonological forms—words beginning with either vowels or consonants. In this sense, we propose that prediction can be highly specific, at least under some circumstances" (19, p. 1119-1120).

Controversy has emerged recently regarding the strength of evidence for word form prediction in variations of the *a/an* design. For instance, we did not observe the effect in younger adults with sentences at a faster presentation rate (20, Experiment 2, 3.3 words per second) or in older adults at two words per second (21) and other groups have reported statistically reliable (22), marginal (23, Experiment 2), and null results (23, Experiment 1). A recent large-scale study by Nieuwland and colleagues proposed to resolve the question by re-using the experimental materials and design of the original DUK05 *a/an* study (healthy younger adults reading two words per second in central vision) and analyzing EEG data collected from nine laboratories around Great Britain (24, hereafter NIET18). That report makes four main points: (1) it is important to replicate experimental findings; (2) the prenominal article correlation with grand average ERPs reported in DUK05 could be a spurious statistical result; (3) with the same stimuli, generally similar procedures, more participants (N=338), and more appropriate statistical analyses, they failed to observe a reliable effect at the prenominal article with either the potentially problematic average ERP correlation analysis or planned and post-hoc single-trial linear mixed-effect regression (LMER) model analyses; (4) if there is such an effect, it is relatively small. We concur. The value of replication is uncontroversial, although rather than simply running the same experiment over and over, there may be more to learn from replication and extension as illustrated by the followup studies DeLong conducted in the lab between 2005 and 2010 and that we have analyzed anew for this report. We recognize the limitations of inferences drawn from correlations between averages and thus analyze single trial EEG data with LMER models for this report. It is also clear that NIET18 failed to observe an effect of prenominal article predictability with the pre-registered LMER analysis of scalp potentials averaged across six scalp locations and a 300 ms post-stimulus interval. However, when the existence of such an effect is in question, there seems little reason to suppose that the most informative general answer is to be had by selecting one temporal interval and a small set of scalp locations in advance and drawing inferences about what is or is not going on throughout the brain as comprehension processes evolve from the analysis of this aggregated snapshot. In what follows, we propose alternatives that build on the strengths of the NIET18 analysis and aim to overcome some of its limitations.

The key empirical premise in the argument for word form prediction based on the *a/an* experimental design is that indefinite article predictability, operationalized as cloze probability, is positively associated with the amplitude of scalp potentials elicited by the articles around 400 ms poststimulus over central and posterior scalp, i.e., that article N400 ERP amplitude correlates inversely with cloze probability. Accordingly, we investigated this association in three EEG data sets recorded in *a/an*-design experiments previously conducted in our laboratory: the original DUK05 experiment and two replication-extension experiments that revised and extended the stimulus materials and experimental conditions. In all three experiments, healthy young adults read sentences two words per second in central vision as in the original DUK05 report and NIET18. In contrast with the absence of evidence reported in NIET18, our exploratory LMER modeling of the single-trial EEG data moment-by-moment at 26 scalp locations finds empirical support for the hypothesized association, which, in turn, may reasonably be attributed to prediction of upcoming word forms.

Exploratory EEG data analysis with regression ERPs. The data from these three experiments have already been analyzed in a number of other ways, published and unpublished (see SI Appendix, Table S1), and the results are known. These circumstances rightly prompt concern about circular analyses, multiple comparisons, and *p*-hacking when choosing which and how among the many available hypotheses to test with confirmatory null hypothesis tests (e.g., 25–28). Since accept-or-reject-at-α confirmatory null hypothesis testing is not appropriate, we present a series of data-driven exploratory analyses along with what Tukey terms *rough confirmatory* assessments of strength of evidence, i.e., a flexible *data investigation* in the sense he contrasts with the rigid steps of *data processing* and confirmatory hypothesis tests (29–31). Consequently, in concept and execution, the analyses reported herein have more in common with the iterative phases of model development, diagnosis, evaluation, and selection found in applied statistical modeling than boiler-plate data processing that passes from EEG recordings to results through a predetermined sequence of steps and declares victory by rejecting (or failing to reject 15, 23, 24), a null hypothesis at $p < .05$. Researchers intrigued or outraged by this approach will find an engaging manifesto in Tukey's "Badmandments" (32, Prologue), a clear overview for psychologists in Behrens (33), and methodological guidance in standard texts, e.g., Cohen, et al. (34, Ch. 4, 10), Fox (35, Data Craft: Ch. 2-4), and Kutner et al. (36, Ch. 9-10, Fig 9.1).

Our exploratory analyses used the same class of LMER models as NIET18 and differ primarily in that we evaluated a greater variety of models and modeled the data at a higher spatial and temporal resolution in the regression ERP (rERP) framework recently described and motivated by Smith and Kutas (37, and references therein for related approaches). For these analyses we sweep an LMER model across the single trial EEG and fit the data for all subjects and items at each time point of the digital recording. As Smith and Kutas point out, modeling the EEG data in this manner is a generalization of conventional sum-and-divide time-domain averaging. For a set of n single-trial EEG epochs (segments of the recording), each time-aligned to an experimental event of interest, the time-domain average $ERP(t) = \frac{1}{n}\sum_{i=1}^{n} EEG_i(t)$ at time, t, is mathematically identical to the estimated intercept, $\hat{\beta}_0$, of an intercept-only linear model of the same data, $EEG(t) = \beta_0 + \epsilon$, fit by ordinary least-squares regression. This means plotting, measuring, analyzing, and interpreting time-domain average ERP waveforms and the time series of estimated linear model intercepts, $\hat{\beta}_0(t)$, are literally one and the same. This approach generalizes to more complex models, notably multiple regression models that may include continuous and categorical predictor variables, and other classes of models including linear mixed-effects models. For models with multiple predictor variables, e.g, $EEG(t) = \beta_0 + \beta_1 X_1 + \ldots + \beta_J X_J + \epsilon$, fitting the model yields a time series of estimated coefficients, $\hat{\beta}_j(t)$, for each regressor, X_j, the waveforms Smith and Kutas dubbed regression ERPs (rERPs). Furthermore, besides the estimated model parameters, fitting a model at each time point also yields the corresponding time series of residual errors and derived quantities such as error variance, coefficient standard errors and confidence intervals, and goodness-of-fit measures. Modeling time series data is nothing new; the key insight of the regression ERP framework is that the logic of conventional event-related time-domain averaging extends to event-related time-domain modeling more generally, and thereby to the investigation of event-related brain activity by methods and procedures from applied statistical data modeling developed to fit, diagnose, compare, and interpret different models. The end game is to determine which model(s), among the many possible, are likely to better or best account for systematic relationships between predictor and response variables, i.e., between experimental variables and event-related brain activity. Determining the existence and form of these associations is the first (though not last) step in causal inference.

Approach

To investigate the association, if any, between the predictability of articles and the brain responses they elicit during word-by-word reading, we swept LMER models across single trial EEG recordings before, during, and after the onset of articles that vary in cloze probability. We make inferences based on the time course and scalp distribution of model goodness-of-fit measures and regression ERPs. Details and further discussion appear in the Methods and Supplementary Information (SI). The analysis reproduction recipe, open-source scripts, and additional figures are available online at OSF: UDCK (38).

EEG data: 3 experiments. After the original study reported in DUK05, DeLong and colleagues continued to investigate aspects of predictive processing in younger and older adults. For this report we selected two studies conducted between 2005 and 2010 that incorporated the *a/an* prenominal indefinite article manipulation and extended the original study design with additional conditions and materials (see SI Appendix, Table S1 for a summary and references). The rationale for selecting these particular studies is that they tested healthy young adults reading two words per second in central vision which affords a close comparison between and across the original DUK05 and NIET18 studies. Furthermore, the additional materials developed by revising and extending the DUK05 materials fill in gaps in the distribution of contextually supported noun and the corresponding pre-nominal article cloze values in the DUK05 materials. This makes the pooled data sets appropriate for modeling article cloze probability as a continuous predictor. So for this report, we pooled the data from these three studies and modeled approximately twelve thousand single trial epochs (Table 1), recorded at 26 scalp locations spanning the interval from about 1.5 seconds before to 1.5 seconds after the critical article (see Methods and SI Appendix, EEG Experimental Procedures).

Table 1. EEG Experiment participants, items, and article cloze

				observed article cloze		
E	P	I	N	M	SD	range
1	32	80	2136 (0.16)	0.38	0.35	0.0 - .97
2	32	160	4668 (0.07)	0.44	0.41	0.0 - 1.0
3	24	240	5232 (0.08)	0.39	0.38	0.0 - 1.0
all	88	320†	12043 (0.10)	0.408	0.389	0.0 - 1.0

E = EEG Experiment. P = Number of participants, I = Number of items in the experimental design for modeling item as a random variable. Each item corresponds to the context prior to the critical article and provides one cloze value for *a* and one for *an* (see Supporting Information for article cloze distributions and data exclusions). N = number of single trials analyzed after excluding EEG artifacts (proportions in parentheses) and stimulus irregularities (0.01). The observed article cloze mean (M) and standard deviation (SD) on each row are computed for the single trial data on that row and may be used to transform estimated regression coefficients for standardized article cloze back to the original cloze scale of 0–1. †Experiment 3 used 160 of the same pre-article item contexts as Experiment 2 and added 80 new ones 80 + 160 + 80 = 320 distinct items. Modeling item random effects takes this into account (see SI Appendix, Stimulus and item coding).

Modeling: linear mixed-effects regression ERPs. To characterize the time-course and scalp distribution of article cloze effects in the regression ERP framework, we swept each of the LMER models in Table 2 across the single trial EEG data and computed the lme4::lmer() profiled maximum likelihood (ML) fit for the 1.2e4 observations at each time point and each channel (39). For exposition, Table 2 presents the models in the formula language of lme4 which specifies LMER models in two parts: the "fixed effect" predictor terms and the "random effect" terms enclosed in parentheses. This syntax aligns with a matrix equation specification of the model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$, that shows the observed response variable $\boldsymbol{y}$ modeled in two parts as the sum of β-weighted regressors for fixed effects ($\boldsymbol{X\beta}$) and b-weighted regressors for random effects ($\boldsymbol{Zb}$). For an introduction to LMER modeling in psychology experiments see the development of Equation (9) in 40 and see 39 for a formal treatment of the model and fitting algorithms.

To highlight the approach in this report, we can unpack $\boldsymbol{X\beta}$ as the column vectors, $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{x}_{\text{cloze}}]$, a column of 1's and the per-item article cloze values, and the scalar coefficients, $\boldsymbol{\beta} = [\beta_0, \beta_{\text{cloze}}]$ for the intercept and article cloze:

$$\boldsymbol{EEG} = \beta_0 \boldsymbol{1} + \beta_{\text{cloze}} \boldsymbol{x}_{\text{cloze}} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \qquad [1]$$

The analyses that follow map neatly onto the terms of Equation 1. First, to select random effects for subjects, items and experiments, we compared models with different $\boldsymbol{Zb}$ (Figure 1). Second, to evaluate evidence for an association between article cloze and scalp potentials we compared (full) models like Equation 1 that include the article cloze regressor, $\boldsymbol{x}_{\text{cloze}}$, with corresponding (reduced) models that do not (Figure 2). Third, the *linear mixed-effect regression ERP* (lmerERP) waveforms are the estimated coefficients for the intercept, $\hat{\beta}_0$, and article cloze $\hat{\beta}_{\text{cloze}}$ over time for each EEG channel (Figure 3).

Model evaluation: Akiake Information Criterion and Δ_i. To have the same metric for comparing larger sets of models en masse and model pairs (41), we evaluated models on estimated Akiake Information Criterion (AIC). In outline, the general form of the AIC $= -2\log(\mathcal{L}) + 2K$ rewards goodness-of-fit through the maximized likelihood, $\mathcal{L}$, of the model given the data, while penalizing model complexity in proportion to the number of model parameters, K. Better fitting models of the same data have larger likelihoods, hence smaller $-2\log(\mathcal{L})$ (*deviance*). Simpler models have fewer parameters, i.e., smaller K. So, among a set of models of the same data, the better fitting, simpler model(s), M_i, have lower AIC values than worse fitting and/or more complex models. We evaluated the degree of empirical support for models in a set according to Burnham and Anderson's heuristics for $\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$, the difference between the AIC for model, M_i, and the minimum AIC among models being compared: "models having $\Delta_i \leq 2$ have substantial support (evidence), those in which $4 \leq \Delta_i \leq 7$ have considerably less support, and models having $\Delta_i > 10$ have essentially no support" (42, p. 270-271). Critically, these heuristics treat AIC differences less than 2 as meaningless for model selection, i.e., they characterize evidential ties, and begin to look for AIC differences around 4 or greater to differentiate alternative models. Taken together, the AIC and heuristics comprise a practical general framework for investigating—comparing and selecting among—sets and pairs of models with fixed and random effects (see SI Appendix, AIC model selection).

Table 2. Linear mixed-effects models as lme4 formulae

Random effects	
maximal	
M0	cloze + (cloze | expt) + (cloze | subject) + (cloze | item)
drop 1 slope	
M1	cloze + (cloze | expt) + (cloze | subject) + (1 | item)
M2	cloze + (cloze | expt) + (1 | subject) + (cloze | item)
M3	cloze + (1 | expt) + (cloze | subject) + (cloze | item)
drop 2 slopes	
M4	cloze + (cloze | expt) + (1 | subject) + (1 | item)
M5	cloze + (1 | expt) + (cloze | subject) + (1 | item)
M6	cloze + (1 | expt) + (1 | subject) + (cloze | item)
drop 3 slopes	
M7	cloze + (1 | expt) + (1 | subject) + (1 | item)
drop 1 random term	
M8	cloze + (1 | subject) + (1 | item)
M9	cloze + (1 | expt) + (1 | subject)
M10	cloze + (1 | expt) + (1 | item)
Article cloze fixed-effect comparisons	
Keep It Maximal (KIM)	
M5	cloze + (1 | expt) + (cloze | subject) + (1 | item)
M5r	(1 | expt) + (cloze | subject) + (1 | item)
Keep It Parsimonius (KIP)	
M7	cloze + (1 | expt) + (1 | subject) + (1 | item)
M7r	(1 | expt) + (1 | subject) + (1 | item)
Experiment as a fixed effect	
Keep It Maximal (KIM)	
M11	cloze + expt + (cloze | subject) + (1 | item)
M11r	expt + (cloze | subject) + (1 | item)
Keep It Parsimonius (KIP)	
M12	cloze + expt + (1 | subject) + (1 | item)
M12r	expt + (1 | subject) + (1 | item)
Experiments 1, 2, and 3 modeled separately	
Keep It Maximal (KIM)	
M13	cloze + (cloze | subject) + (1 | item)
M13r	(cloze | subject) + (1 | item)
Keep It Parsimonious (KIP)	
M14	cloze + (1 | subject) + (1 | item)
M14r	(cloze | subject) + (1 | item)

Note: Fixed and random intercepts are implicit and modeled by default.

Random effects selection. There is some debate in the recent mixed-effects modeling literature about whether maximal or parsimonious random effects are appropriate for hypothesis testing with LMER models (43, 44). The debate turns in part on how the decision to include, e.g., random slopes in addition to random intercepts, impacts the rate of incorrect null hypothesis rejections (Type I errors) vs. loss of power and failure to reject the null hypothesis (Type II errors). We took the present project as an opportunity to evaluate the consequences of the decision as a case study of exploratory data analysis. Specifically, among the 11 candidate models with random effects ranging from maximal to minimal, $M0, \ldots, M10$ (Table 2), we selected two for further investigation according to different decision rules. "Keep It Maximal" (KIM): select the maximal random effects for which the model converges reliably. "Keep It Parsimonious" (KIP): select the simplest random effects for which the model converges reliably and does not have substantially less support than the alternatives ($\Delta_{\text{M}_i} \geq 4$).

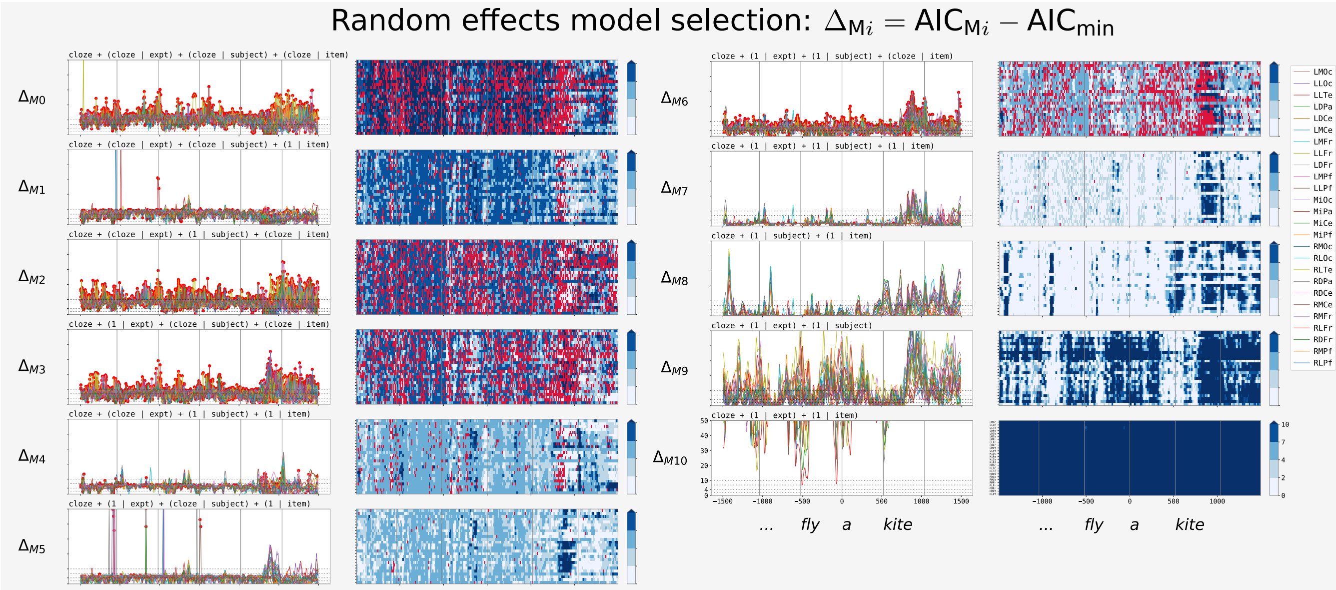

Fig. 1. The time course and scalp distribution of AIC $\Delta_{\mathrm{M}i}$ comparisons among models in the set $\{M0, \ldots, M10\}$ (Table 2). Each panel, $\Delta_{\mathrm{M}i}$, indicates how the AIC for model M_i compares with the best supported model (minimum AIC) among the eleven candidates at each time point and channel: $\Delta_{\mathrm{M}i} = \mathrm{AIC}_{\mathrm{M}i} - \mathrm{AIC}_{\mathrm{min}}$. Since there is always some minimum AIC, somewhere among the models $\Delta_{\mathrm{M}i} = 0$. As the panels show, this varies by time point and channel. The x-axis is time in milliseconds, vertical lines indicate stimulus word onsets, critical article onset is at 0. The rainbow line plots show the time course of $\Delta_{\mathrm{M}i}$ (y-axis) for each channel in colors given by the channel legend; horizontal lines indicate the Burnham and Anderson Δ_i heuristic intervals bounded by 2, 4, 7, and 10. A few values for M9 and most for M10 are above 50 and not shown. The adjacent blue and red raster plots show the same data: darker colors correspond to larger $\Delta_{\mathrm{M}i}$ values; shading levels correspond to the heuristic intervals. EEG channels are arrayed on the y-axis in the order given by the channel color legend: the top 11 rows are left hemiscalp, the next four are midline, the bottom 11 rows are right hemiscalp. At a glance, the lightest patches among the raster plots indicate the best- (or equally well-) supported model(s) in the set ($0 \leq \Delta_{\mathrm{M}i} \leq 2$) and darker patches indicate that the model is less well supported than an alternative ($\Delta_{\mathrm{M}i} > 2$). Times and channels where lme4::lmer() fitting generated a warning are indictated with red. Models M5 and M7 were selected for further investigation based on the Keep it Maximal and Keep it Parsimonious selection rules, respectively. These results are for models fit to approximately 1.2e4 single trial observations at 8 ms intervals and 26 EEG channels (Table 1).

Evidence for an article cloze effect: Δ_{M} and lmerERPs. The critical empirical question is whether there is an association between article cloze and scalp potentials generated by brain activity in response to encountering those articles. We approached this in two ways based on fitting the models selected by the KIM and KIP decision rules: 1) we computed Δ_{M} and Δ_{Mr} for the full and corresponding reduced model pairs taking $\Delta_{\mathrm{Mr}} > 4$ as indicative of substantially less support for the reduced model; 2) we examined the magnitude and confidence intervals of the article cloze (slope) regression ERPs for the full model.

The possible outcomes and interpretations of this regression ERP modeling are straightforward. If the article cloze and scalp potentials are unrelated, including article cloze in the model should have little impact on the goodness-of-fit and Δ_{M} for the full vs. reduced model should be around 2 because of the AIC penalty for the additional parameter. And in this same case, the article cloze (slope) rERP waveforms should tend to be around 0 plus or minus random variation, i.e., the X-Y trend line for article cloze (X) vs. EEG (Y) at each point in time should tend to be flat. Alternatively, if there is an approximately linear association between article cloze probability and scalp potentials, the deviance term of the AIC for the full model should be smaller. In this case, the extent to which Δ_{Mr} for the reduced model is greater than 2 indicates the degree to which the full model is better supported by the data after adjusting for its increased complexity, with $\Delta_{\mathrm{Mr}} > 4$ indicating a substantial difference in support. Furthermore, the time course and scalp distribution of the Δ_{Mr} values and lmerERPs are important. To support the inference that the potentials are generated by a brain response to the article, an AIC Δ_{Mr} effect should be evident in the interval after article onset and not before. Likewise, the article cloze (slope) rERP waveforms should tend to hover around 0 prior to article onset and then deviate from zero afterwards, with the polarity of the deviation, positive or negative, indicating the direction of the association (correlation).

Taken together, the full vs. reduced model pair Δ_i values and the magnitude of the lmerERPs relative to their confidence intervals are the basis of our evaluation of the strength of evidence for an article cloze effect, the rough confirmatory analysis in Tukey's sense. In Tukey's view (31, p. 24), strong confirmatory null hypothesis testing requires designing, executing, and analyzing an experiment to ask and answer one question, thereby reducing the entire project to a single bit of information—1 or 0, significant or not (32, p. 277). By contrast, our exploratory modeling aims to gauge where and when and to what extent—if any—there is evidence to support a linear approximating model of the relationship between article cloze and scalp potentials.

Results

The following summarizes the main findings in the critical interval from 1.5 s before article onset up to the onset of the following word. Note that Figures 1, 2, and 3 display the 3 s of data modeled which spans the two words after the article.

Random-effects selection. The LMER models M0, M1, ..., M10 (Table 2) hold constant the intercept and fixed-effect of article and vary the random effects. Figure 1 shows there is no unique best supported model with minimum AIC at all time points and EEG channels, i.e., no single model where $\Delta_{\mathrm{M}i} = \mathrm{AIC}_{\mathrm{Mi}} - \mathrm{AIC}_{\mathrm{min}} = 0$. However, some models were much less supported than others in the 1.5 s pre- to 0.5 s post-article interval and we selected two for further investigation. First, in accord with both decision rules, we ruled out models with substantial numbers of fitting warnings (M0, M1, M2, M3, M4, and M6), each of which included item or experiment random slopes for article cloze. Of those remaining, in accord with the Keep It Maximal decision rule, we selected M5 with random intercepts for experiment, subject, and item and a random slope for subjects as the model with the maximal random effects that reliably converged, KIM M5: `cloze+(1|expt)+(cloze|subject)+(1|item)`. We examined the remaining models with simpler random effects and, unsurprisingly, found intervals of substantially less support ($\Delta_{\mathrm{M}i} > 4$) for models that dropped any one of the experiment, subject, or item random variables entirely (M8, M9, M10). Consequently, in accord with the Keep It Parsimonious rule we selected model M7 with random intercepts for experiment, subject, and item as the model with the most parsimonious random effects that was well-supported by the design and the data, KIP M7: `cloze+(1|expt)+(1|subject)+(1|item)`. Neither the KIM (M5) nor KIP (M7) models were entirely free of fitting warnings, but these were scattered irregularly across the times and channels and few in number, especially during the interval of interest. Although Keep It Maximal and Keep it Parsimonious decision rules may represent different extremes, in this particular instance, the models selected, M5 and M7, differed only in whether or not to include an article random slope for subjects.

Evidence for an article cloze effect. With the KIM (M5) and KIP (M7) models selected for further investigation, we turned to the research question of primary interest: is there evidence of an association between article predictability and scalp potentials? We addressed this by pairwise AIC model comparisons between the full and reduced KIM (M5, M5r) and KIP (M7, M7r) models (Figure 2) in conjunction with the values of the estimated coefficients for the article cloze predictor in the full models, $\hat{\beta}_{\mathrm{cloze}}$, i.e., the article cloze regression ERPs (Figure 3B).

We note first that Δ_{M5} and Δ_{M7} for the full KIM and KIP models, respectively, accord with the definitions of AIC and Δ_{M}. These values range between 0 and 2 at all times and channels (Figure 2, top row), except for a few anomalous values where the fitting failed to converge for the maximal model M5. These expected results support the face validity of the AIC estimates and Δ_{M} calculations which appear to be generally well-behaved for these models and data.

The key evidence for an article cloze effect is observed at those scalp locations and times where the reduced models Δ_{M5r} and Δ_{M7r} values are > 4, indicating a substantial decrease in goodness-of-fit when the article cloze predictor is omitted from the model. For these reduced models (Figure 2, middle and bottom rows), there are two intervals of immediate interest: the prestimulus interval (-1.5–0 s), and the critical article (0–0.5 s). The interval spanning the words immediately following the article (0.5–1.5 s), is relevant as well, albeit less directly, as we touch on in the Discussion.

Prestimulus $\Delta_{\boldsymbol{M}}$**.** During the 1.5 s preceding the onset of the critical article, values for the reduced KIM model range between 0 and 2 (Figure 2A, Δ_{M5r}) with occasional irregular values above 2 (indicated by the darker blue speckles) and, again, a few anomalously large AIC values coincident with model fitting warnings. The findings for the reduced KIP model with the parsimonious random effects are similar (Figure 2B, Δ_{M7r}) except that there are fewer fitting warnings and no anomalous Δ_{M7r} excursions. Since $\Delta_{\mathrm{M}} \leq 2$ for the most part during the prestimulus interval, and rarely > 4, we conclude that support for the full and reduced models does not differ substantially in this interval for either the KIM or KIP random effects. This evidential tie in the prestimulus interval is instructive for what it does not show. Given the design of the experiment, and the epoch centered on the entire 1.5 s prestimulus baseline, an effect of article predictability should be evident upon encountering the article but not before. If the modeling showed an article cloze effect prior to article onset, it could indicate something amiss in the design or execution of the experiments, the model specification or fitting, or the model comparison metric. In so far as we can determine with the present approach, examination of the 1.5 s of prestimulus activity for the 26 scalp locations at 8 ms intervals reveals no clear indication of these potential defects. Consequently, we suppose that article cloze effects observed in the interval following article onset may reasonably be attributed to a brain response to the article.

Critical article Δ_M**.** Following the onset of the critical *a/an* indefinite articles, the AIC differences between the full and reduced models do not appear to be dramatically different from those in the prestimulus interval until about 300 ms poststimulus. Then, between around 300 ms and the onset of the next word, AIC values for the reduced models, M5r and M7r, are systematically larger, predominantly over bilateral posterior scalp, peaking around 400 ms (Figure 2A and 2B, bottom row, magenta highlight). This increase was not observed over anterior scalp. The results for the KIM and KIP models are similar: the KIP model Δ_{M7r} values are slightly larger in some cases, there are fewer fitting warnings, and no anomalously large AIC values. For both the KIM and KIP comparisons, there appears to be an oscillation around 10 Hz in the reduced models (Δ_{M5r}, Δ_{M7r}) during the interval 300–500 ms poststimulus, and perhaps earlier, over posterior scalp. These oscillations may indicate residual alpha band noise EEG though the possibility of an event-related 10 Hz amplitude modulation should not be overlooked. These oscillations make evaluation of the time course of AIC differences on a scale below about a tenth of a second precarious, but the slower phasic response is evident with or without the oscillations. We interpret this phasic increase in Δ_{M5r} and Δ_{M7r} above 4 for the KIM and KIP pairwise model comparisons as empirical support—rough confirmation—of a systematic association between article cloze and scalp potentials 300–500 ms over posterior scalp. This effect is the crux of the argument for word form prediction.

Article cloze lmerERPs. Whereas the full vs. reduced model AIC comparisons indicate when (around 300–500 ms poststimulus) and where (bilateral posterior scalp) there is evidence of an article cloze effect, the magnitude and polarity of the estimated rERP slope coefficients characterize the magnitude and

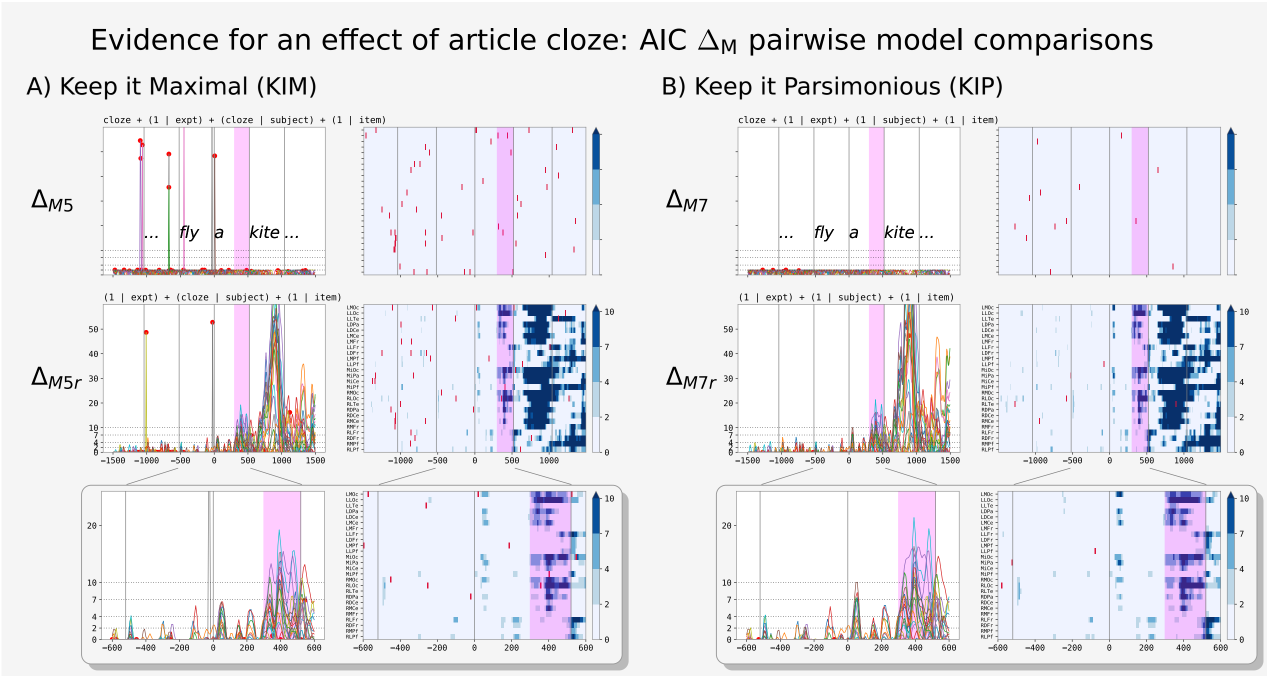

Fig. 2. AIC Δ_{M} pairwise full vs. reduced LMER model comparisons. A) Keep it Maximal (KIM): full (M5) vs. reduced (M5r). B) Keep it Parsimonious (KIP): full (M7) vs. reduced (M7r). Axes, scales, and data are as in Figure 1. The top two rows shows AIC Δ_{M} and Δ_{Mr} for the full and reduced models respectively across the 3 s epoch, article onset at 0. The bottom row inset zooms in to show AIC Δ_{Mr} for the reduced model at the critical prenominal article in more detail. For both comparisons, during the 1.5 s interval preceding the critical article, the full and reduced models are equally supported, Δ_{M} and $\Delta_{\mathrm{Mr}} < 2$, with a few idiosyncratic exceptions. During the interval around 300 - 500 ms following the article onset (highlighted in magenta), the reduced models are substantially and systematically less supported at bilateral posterior scalp locations, Δ_{M5r} and $\Delta_{\mathrm{M7r}} > 4$, as indicatd in panels A and B by traces above 4 in the rainbow line plots and darker blue bands in raster plots.

direction of the association under the assumption of a linear relationship. We found that the magnitude and confidence intervals for the KIM and KIP intercept ($\hat{\beta}_0$) and article cloze ($\hat{\beta}_{\mathrm{cloze}}$) lmerERPs are essentially indistinguishable over the entire 3 s epoch (see SI Appendix, fig. S4) and we present results here for the KIM model only (Figure 3).

The model intercept lmerERPs ($\hat{\beta}_0$) are the rERP analog of grand mean average ERPs. These show the morphology characteristic of visual evoked potentials, a series of six transient responses to the six words presented two per second over the three second epoch (Figure 3A). For the critical article cloze lmerERPs ($\hat{\beta}_{\mathrm{cloze}}$) we found that prior to the onset of the article, they hover around 0 and the 95% confidence intervals for the point estimates generally span 0 (Figure 3B). Then, following the onset of the critical article, we observed a biphasic positive response. The first phase begins around 300 ms after the article, is larger predominantly over posterior scalp, increases to a peak around 400 ms and then decreases until shortly after the onset of the following word. The polarity of this deflection indicates a positive association, i.e., as cloze probability of an article increases, scalp potentials over posterior scalp become more positive. This interval, about 300–500 ms post-article, is the first time in the epoch where the lower bound of the 95% confidence interval for the article cloze rERP is above 0 for sustained periods. A second, larger phasic positive deflection was observed, peaking around 400 ms after the word following the article, with a time course and scalp distribution corresponding to the larger second phase of increased AIC Δ_{Mr} for the reduced KIM and KIP models that emerges after the onset of the word following the article (Figure 2, panels A and B, second rows).

In sum, we observed what appears to be a systematic event-related lmerERP response to the article with a polarity, latency, and scalp distribution that coincide with previously reported reductions in N400 ERP amplitude with increasing cloze probability. We interpret this as direct evidence that the brain response to the article systematically covaries with the predictability of the indefinite articles *a* and *an*. To the extent the predictability of the article is dependent on the predictability of the not-yet-presented noun and its initial speech sound, the positive-going phasic article cloze lmerERP response is reasonably interpreted as indirect evidence for word form prediction.

Interim Summary

When we modeled about twelve thousand EEG single trials moment by moment at 26 scalp locations with appropriate linear mixed effects models, we found that models that include article cloze probability as a predictor variable do a substantially better job accounting for the variability in potentials recorded over posterior scalp around 300–500 ms after the onset of the article. The face validity of the modeling generally, and pairwise AIC model comparison results in particular, are bolstered by the facts that 1) $\Delta_{\mathrm{M}} \leq 2$ for the full models are in line with theory, 2) the full and reduced models are equally supported during the prestimulus interval when no difference is expected, and 3) the direction of the observed positive association between article cloze probability and scalp potentials characterized by the slope regression ERPs agrees with the previously reported reductions in average N400 ERP

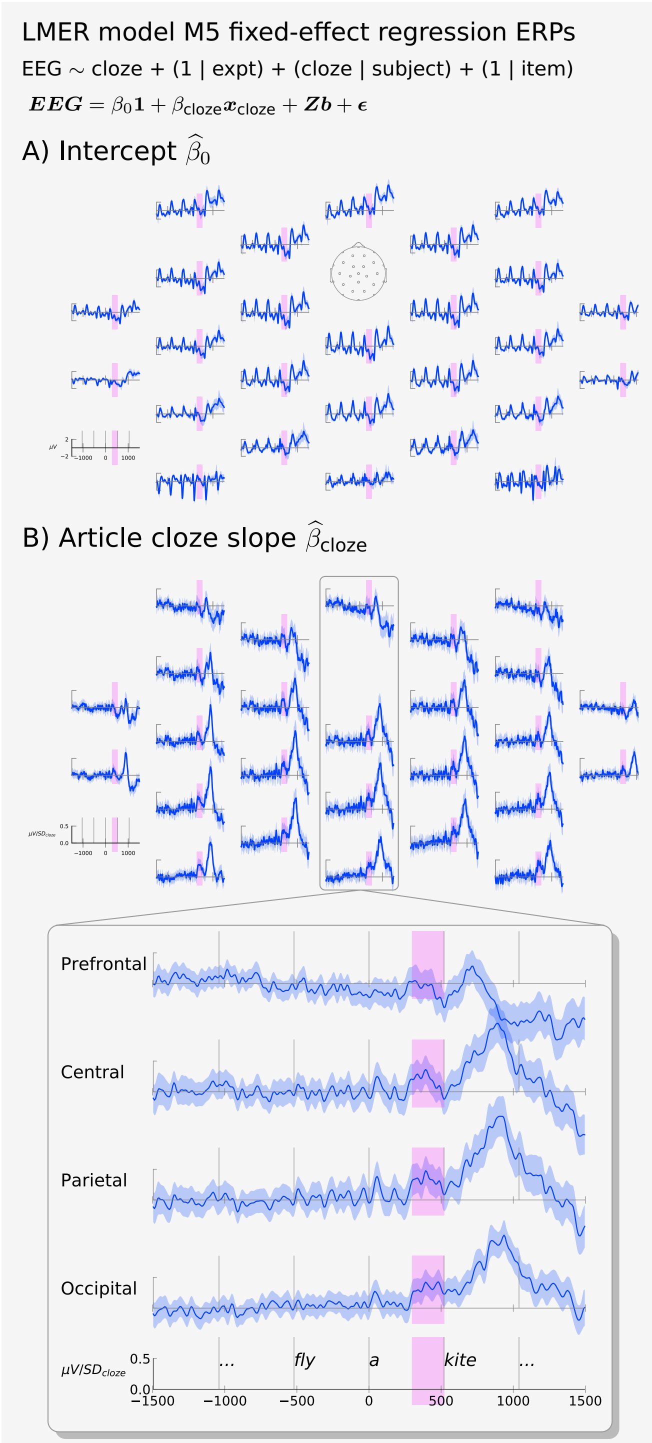

Fig. 3. Model M5 linear mixed-effects regression ERPs (3 seconds, 26 channels). Solid lines plot the estimated regression parameter over time (ms) relative to critical article onset at 0, bands indicate 95% confidence intervals, positive values are plotted up. Anterior to posterior scalp locations are arrayed top to bottom in each panel. A) Intercept lmerERPs ($\hat{\beta}_0$) are analogs of grand mean average ERPs and show the characteristic morphology of visual evoked potential responses, sharply defined transient peaks and troughs, especially prominent over lateral occipital scalp. B) Article cloze lmerERPs ($\hat{\beta}_{\text{cloze}}$), characterize the slope of the straight line relationship between standardized article cloze and scalp potentials as it evolves over time. The y-axis is μV per unit standardized cloze. The cloze lmerERPs show a transient positive response, predominantly over bilateral posterior scalp, around 300–500 ms after article onset (magenta highlight) and not before, indicating a positive association between cloze probability and scalp potentials in response to the critical prenominal articles.

amplitude with increasing cloze probability (19, 45). As best we could determine, for these data, the perhaps contentious choice to fit models with maximal or parsimonious random effects made little difference for characterizing the time course, scalp distribution, or strength of empirical support for the article cloze effect based on model comparisons or for estimating the fixed-effect of article cloze, i.e., the magnitude and precision of the lmerERP estimates.

Followup Analyses

Since exploratory data investigation arrives at conclusions through an iterative process of evaluating assumptions and alternatives, we conducted a number of followup analyses, summarized briefly here (see SI for further details and discussion).

Influential data diagnosis. A general issue for the interpretation of estimated regression model coefficients is whether subsets of extreme or outlying observations exert a disproportionate influence on estimates and exaggerate (or obscure) patterns seen in the bulk of the data. For modeling the time course of the article cloze effect, this question is whether the morphology of the lmerERP waveforms in particular, is driven by a subset of unrepresentative data. Mixed-effects modeling is computationally intensive and influence diagnostics based on model refitting are intractable for data on the scale of this analysis at present so we fell back to ordinary least squares (see SI Appendix, Influential data diagnosis). We identified and excluded a subset of about 5% of the single trial epochs that contained the highest proportion of potentially influential observations. We then re-fit the KIP and KIM models to this trimmed data set and computed how much the amplitude of the intercept and article cloze lmerERPs changed as a consequence of the trimming, i.e. we computed a version of the DFBETAS data diagnostic, adapted for regression ERPs. We assumed that article cloze DFBETAS ± 2 would indicate an unusually large change in the rERP estimate based on a large n Student's t distribution (35). We found there were few DFBETAS excursions of that magnitude and those that occur do so at the peaks and troughs of approximately 10Hz oscillations (see SI Appendix, fig. S6). This oscillation suggests that the epochs identified and excluded contained high amplitude alpha band activity. Crucially, the time course and distribution of Δ_{Mr} values for the reduced KIM (M5r) and KIP (M7r) models of the trimmed data still show the phasic increase over posterior scalp around 300–500 ms and the article cloze lmerERPs ($\hat{\beta}_{\text{cloze}}$), show the corresponding positive deflection (see SI Appendix, fig. S7). So it appears the article cloze effect observed in the initial analysis are not driven entirely by this subset of potentially influential trials.

Modeling experiment as a fixed effect. The designs and procedures of EEG Experiments 1, 2, and 3 are sufficiently similar to justify pooling the data for purposes of modeling the brain reponse to the critical indefinite articles, provided systematic variation between the experiments is also accounted for. Since for our purposes, systematic differences between the experiments is nuisance variation and the different numbers of trials in the three experiments make the design substantially unbalanced, we modeled experiment as a random variable. However, views may differ on the appropriate treatment of categorical variables as fixed vs. random and the consequences for drawing model-based inferences, particularly when the number of levels is small (for discussion see 46, p. 20ff and 47, p. 246 and p. 275ff). So we investigated the question by modeling the single trial EEG with article cloze and experiment as fixed-effects, retaining the KIM and KIP random effects for subjects and items, see Table 2 KIM (M11, M11r) and KIP (M12, M12r). We found that fitting full and reduced models with experiment as a fixed effect converged reliably and the pattern of AIC Δ_{M} and Δ_{Mr} for the pairwise full vs. reduced model comparisons and article cloze regression ERPs and their confidence intervals are essentially the same as for models with a random intercept for experiment (see SI Appendix, fig. S8). So in this instance, the choice of fixed- vs. random-effect for the experiment variable was immaterial for inferences about the article cloze effects.

Modeling Experiments 1, 2, and 3 separately. To assess whether the article cloze effect observed for the data pooled across the three experiments was representative of each experiment individually, we split the data by experiment and fit the full and reduced model pairs in Table 2: KIM (M13, M13r), and KIP (M14, M14r). For each experiment, we examined AIC Δ_{M} and Δ_{Mr} measures and the article lmerERPs (see Experiment 1, SI Appendix, fig. S9; Experiment 2, SI Appendix, fig. S10; Experiment 3, SI Appendix, fig. S11). The results were mixed for the AIC model comparisons and somewhat more consistent for the article cloze lmerERPs. For the Experiment 1 data, fitting the full and reduced models with KIM random effects had considerable difficulty converging. Fitting the full and reduced KIP models converged reliably with irregular intervals of $\Delta_{\mathrm{M14r}} > 4$ throughout the 3 s epoch and no clear break in the pattern between the pre- and post-article interval that suggests an event-related brain response to the article. So the AIC model comparisons did not provide clear evidence for a relationship between article cloze and an event-related EEG response in Experiment 1. For the Experiment 2 data, the KIM and KIP models converged reliably with only a modest increase in convergence failures for the KIM models. Overall, the time course and scalp distributions were generally similar to those for models of the data pooled across all three experiments, with scattered idiosyncratic $\Delta_{\mathrm{M13r}} > 4$ in the prestimulus interval and a systematic onset and offset around 300 ms and 500 ms post-article, respectively. For the Experiment 3 data, there are slightly more convergence failures for the KIM models and prestimulus AIC differences for the reduced model are evident, more so for the KIP comparison, though not to the extent observed for Experiment 1. In the critical interval around 300–500 ms post-article, AIC differences larger than in Experiment 1 and smaller than Experiment 2 rise and fall. In all three experiments, the article cloze lmerERPs tended to vary around 0 prior to the critical article onset, after which they showed a small positive deflection followed by a larger one over bilateral posterior scalp. The onset of this rERP response in Experiment 1 appears to be perhaps 100 to 200 ms later than in Experiment 2 and Experiment 3, though the timing in Experiment 1 is obscured by a pronounced oscillation around 10 Hz. In sum, the AIC Δ_{M} results observed for the data pooled across the experiments appear to be more representative of Experiments 2 and 3 than Experiment 1. The pattern of article cloze slope lmerERPs was more consistent and all three experiments showed a similar, albeit more variable, biphasic positive response following the article, similar to that observed for the pooled data.

LMER modeling interval mean amplitude. Whereas the regression ERP analyses described thus far model the moment-by-moment time course of the article cloze effect from 1.5 s before to 1.5 s after the article, experimental EEG studies using event-related designs, including DUK05 and NIET18, often base inferences about event-related brain responses on measurements of scalp potentials aggregated over a specific time interval, e.g., mean amplitude between 200 or 300 and 500 ms poststimulus, relative to mean amplitude in a specified pre-stimulus baseline interval, e.g., 100, 200, or 500 ms. To compare the LMER regression ERP results with interval mean amplitude analyses, we reduced the single trial EEG time series data to four sets of summary measures: mean amplitude in two post-stimulus intervals (200–500 ms, 300–500 ms), each measured relative to a baseline of mean amplitude in two intervals (100 ms and 500 ms prestimulus). We then modeled these single-trial time-averaged mean amplitude measurements by fitting the KIM (M5, M5r) and KIP (M7, M7r) model pairs at each of the 26 EEG channels separately (c.f., NIET18 LMER analyses of mean potentials aggregated in the interval 200–500 ms poststimulus across six centro-parietal scalp locations).

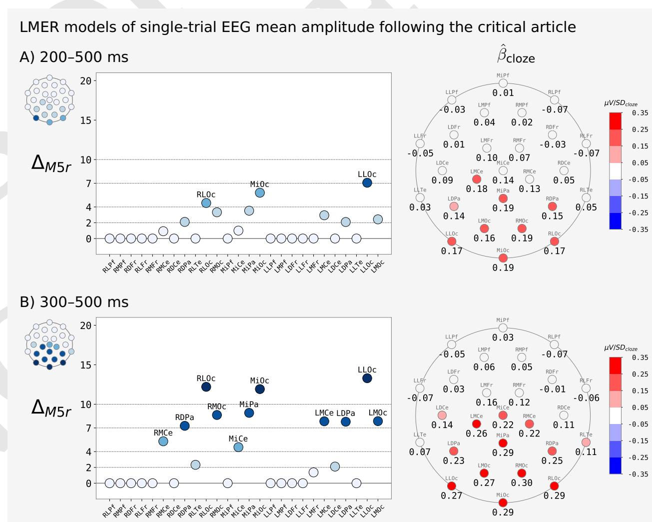

Fig. 4. Comparison of KIM models M5 and M5r of single-trial mean EEG amplitude measured in a longer, earlier-starting interval 200 - 500 ms poststimulus (Panel A) and a shorter, later-starting interval 300 - 500 ms poststimulus (Panel B). The left column shows the AIC Δ_{M5r} values for the pairwise full (M5) vs. reduced (M5r) KIM model comparison. Δ_{M5} for the full model (not shown) were between 0 and 2 as expected for this comparison. The right column shows the magnitude of the estimated fixed-effect coefficient for article cloze, $\hat{\beta}_{\mathrm{cloze}}$, positive values in red, with filled circles only at locations where the 95% confidence interval for the estimate did not include 0. Like the temporally fine-grained regression ERP models, this single-trial LMER modeling indicates a positive association between article cloze and potentials over bilateral posterior scalp around 400 ms postimulus, albeit more robust for the shorter and later interval 300 - 500 ms poststimulus. Results in this figure are for poststimulus potentials measured relative to mean amplitude in a 500 ms prestimulus baseline; results for measurements relative to a 100 ms prestimulus baseline were similar. See OSF: udck19_pipeline_5.html for these and additional analyses.

Consistent with the lmerERP time-course analysis, modeling the potentials averaged across these temporal intervals also found a positive association between article cloze, with a posterior scalp distribution (Figure 4). Across the different combinations of model random effects, baseline intervals, and N400 measurement intervals, only the poststimulus measurement interval had much impact on the results

(OSF: udck19_pipeline_5.html). Regardless of the random effects or prestimulus baseline interval, the magnitudes of the estimated article cloze coefficients for the longer and earlier 200–500 ms poststimulus interval measurements tend to be around $\frac{1}{3}$ smaller than for the measurements made 300–500 ms poststimulus (Figure 4A vs Figure 4B). Attenuated article effects in the 200–500 ms post-article interval are consistent with the time-course regression ERP modeling which found no clear evidence of the article effect before 300 ms poststimulus.

Lurking variables and spurious lmerERPs. Another general issue for the interpretation of an estimated regression model coefficient is the spurious effect that can result from a "lurking" variable, i.e., a variable that is causally related to the response variable and correlated with the predictor but omitted from the model (for discussion, see SI Appendix, pp. 6-8). If the article cloze lmerERPs in Figure 3 are driven purely by correlation with some causal factor unrelated to the form of the indefinite article, interpreting them as support for word form prediction would be unwarranted. The impact of a lurking variable on a regression coefficient can be quantified as the omitted variable bias (e.g., 35, pp. 111-112), which we used to investigate the impact of a variable known to be correlated with article cloze but unrelated to the form of the indefinite article[*]. Since our normative stimulus testing was free response, the proportion of indefinite articles goes down as the proportion of non-article responses, (e.g., bare plurals, adjectives, definite articles), goes up. The article and non-article cloze probabilities are negatively correlated ($r = -0.264, p < 0.0001$, see SI Appendix, fig. S12). We modeled the non-article cloze rERP (see SI Appendix, fig. S13), and found that despite this correlation, the omitted variable bias does not account for the article cloze lmerERP (see SI Appendix, fig. S14). Numerous variables are associated with article cloze and scalp potentials to some degree. However, unless the correlations are strong and the omitted variable regression ERPs are large, the bias is small and thus unlikely to account for the article cloze effect.

Discussion

The project reported herein aims to shed light on the recent theoretical controversy about whether the human language comprehension mechanism anticipates the phonological form of upcoming words. The crucial empirical question is whether processing at the prenominal articles, *a/an*, varies with their predictability since, other things equal, the factor responsible for the form of the indefinite article is the initial speech sound of a not-yet-encountered word. Because of this phonological dependency, direct evidence of an effect of predict-*ability* at the article may be reasonably interpreted as indirect evidence that, by then, upcoming noun word forms were predict-*ed*.

To investigate the time course of the electrical brain activity we modeled single-trial EEG recorded before, during, and after presentation of pre-nominal indefinite articles (*a/an*), in three experiments that manipulated the predictability (cloze probability) of nouns in sentence contexts read by healthy younger adults at two words per second in central vision. Our interim conclusion was that models that include article cloze probability as a continuous predictor do a substantially better job accounting for the variability in potentials recorded over bilateral posterior scalp around 300–500 ms after the onset of the article than do models that omit this variable. Since this was not the case during the 1.5 s prior to the article, we interpreted these results as evidence of a systematic association between article cloze probability and scalp potentials generated by the brain response to the article. The latency, polarity, and scalp distribution of this article cloze effect is generally consistent with the association between cloze probability and scalp potentials (19, 45).

Exploratory investigation of alternatives indicated that evidence for the association does not appear to depend on the choice of maximal (KIM) or parsimonious (KIP) random effects, to be driven by the influence of a subset of unrepresentative data, or to depend on whether the experiment variable is modeled as a fixed or random effect. That said, the article cloze effect appears to be markedly smaller (less variability accounted for, lower amplitude slope lmerERPs) than a corresponding effect at the following word (Figure 2 and Figure 3, immediately after the magenta highlight). In this experimental design (. . . *a kite* . . .), article cloze probability is correlated, though not perfectly, with noun cloze probability. The larger Δ_{Mr} and lmerERP effects for the article cloze predictor variable on the following word are likely a consequence of this relationship but cannot be strictly attributed to the contextually supported nouns because in a subset of materials in Experiment 2, a phonologically legal adjective is interposed between the article and noun, *an orange kite*. Given the high proportion of nouns relative to adjectives in the combined data, it is reasonable to suppose that modeling potentials elicited by the nouns with noun cloze as a predictor variable would find similar, if not larger effects, but testing this speculation is tangential to the present aims and beyond the scope of this report. Although the comparison is imperfect, in all of the models investigated, the magnitude of the transient article cloze rERP response at the article was smaller than at the following word. In this respect the pattern is consistent with other studies that recruit sequential dependency experimental designs to test for prediction in language comprehension and report relatively small and variable ERP effects at the probe word (8, 9, 11–14, 16).

LMER modeling the single-trial data for each experiment separately found that article cloze slope lmerERPs for all three experiments showed a biphasic positive response following the article, similar to that observed for the pooled data, albeit more variable. The AIC Δ_{M} patterns for the individual experiment pairwise model comparisons were similar to the pooled data for two of the data sets, Experiment 2 and Experiment 3 to a lesser extent, but not Experiment 1. This is not entirely surprising since there are roughly twice as many single-trial observations in Experiments 2 and 3 as in Experiment 1 (Table 1). It may be that the two-part stimulus presentation procedure and/or the additional materials developed for Experiments 2 and 3 afford a better opportunity to observe a small article cloze effect with a single trial LMER analysis than do the procedures and materials used for the DUK05 study. While the regression ERP modeling does not show clear evidence of an article effect for the Experiment 1 data considered on its own, the findings are consisent with the stronger support provided by the replication and extension studies that followed. We also modeled single trial mean amplitude in the post-article intervals 200–500 ms and 300–500 ms with the same KIM and

[*] We thank an anonymous reviewer for suggesting this example.

KIP LMER models used for the time course modeling. The choice of KIM vs. KIP model and choice of measurement relative to a shorter (100 ms) vs. longer (500 ms) prestimulus baseline interval had a negligible impact on the results, but in all cases, the magnitude of the article cloze effect was markedly smaller for the 200–500 ms poststimulus interval.

Taken together, this pattern of findings may be relevant to understanding the failure to observe an effect of article cloze reported in NIET18. That study tested only the smaller set of *a/an* items and single sentence RSVP presentation used for the study reported in DUK05 (Experiment 1 in this report), whereas we found that the article cloze effects may be more readily observed in the followup Experiments 2 and 3 with the expanded sets of items and two-part stimulus presentation. The LMER analyses reported in NIET18 were conducted on single-trial mean amplitudes in the interval 200–500 ms post-article, averaged over six centro-parietal electrode locations, whereas our time course modeling at each scalp location found the article cloze effect to have a more posterior distribution and somewhat later onset (Figures 2 and 4). The LMER model pairs compared in NIET18 for the likelihood ratio tests of the null hypothesis assumed maximal random effects with correlated random intercepts and slopes for subjects and items whereas we found that in pairwise AIC model comparisons, the article cloze effect was at times slightly attenuated for the maximal relative to parsimonious model (Figure 2, Δ_{M5r} vs. Δ_{M7r}). So although the decisions made in conducting and analyzing the study reported in NIET18 are defensible for purposes of conducting a direct replication of the DUK05 study, they may be suboptimal for answering the scientific question of interest about word form prediction.

The failure of the NIET18 report to observe a prenominal cloze probability effect in a much larger data sample with generally similar design parameters as the DUK05 report raised the possibility that there is no such systematic relationship between prenominal article cloze and electrical brain activity at all. This is the primary research question that our project was designed to address, using an exploratory analysis approach. To answer this specific question, we selected data from experiments similar to both DUK05 and NIET18: *a/an*, designs testing young adults reading at two words per second in central vision. This selection affords meaningful comparisons among the studies but it also means the results do not answer looming secondary questions about how various experimental variables such as presentation rate or age, among others, might impact the model fits and morphology of article cloze regression ERP waveforms. Still less does the analysis answer broader questions about the generalizability of the findings in the way a meta-analysis might. Although we pooled data across multiple studies, ours is a forensic EEG data investigation, not a meta-analysis. And, considered in its entirety, the pattern of results from the lmerERP modeling we conducted does appear to provide direct evidence of an association (quantitative relationship) between prenominal article cloze and scalp potentials. Of course, the time course, scalp distribution, and polarity of article cloze slope lmerERPs, i.e., the estimated $\hat{\beta}_{cloze}$ coefficients are key to this interpretation. And of course, if a model omits (any) relevant predictor variables, estimates of the coefficients for variables that are included may be biased and, in turn, inferences drawn from the model may be wrong; we never know with certainty whether a model omits relevant predictors. Interpreting our findings as evidence of a structural relation between the predictability of the stimulus and the brain response it elicits requires the stronger assumption that there are no serious lurking variables. This caveat applies to all regression modeling. All the more reason to systematically explore the data, "look for what can be seen, even if not anticipated." (48, p. 24).

Conclusions

In contrast with the large scale null result reported in NIET18, our moderately large scale LMER modeling of single-trial EEG moment-by-moment at 26 scalp locations finds direct empirical support for an association between the predictability of prenominal indefinite articles and the brain's response to encountering them in word-by-word reading. This effect may reasonably be attributed to prediction of upcoming word forms in answer to the question of scientific interest. The exploratory modeling reported herein illustrates an approach to experimental EEG data analysis that may prove a useful complement to confirmatory null hypothesis testing.

Materials and Methods

Methods. All normative stimulus testing and EEG studies were conducted under human subjects resarch protocols approved by the University of California, San Diego Institutional Review Board. Volunteers were recruited by flyer and through the campus subject pool. Upon their arrival at the lab, the experimental procedures were explained verbally and participants were presented with a printed consent form describing the procedures and potential risks. Individuals who elected to participate in the study provided their written informed consent and received two hours of course credit, cash payment, or a combination, at their discretion. The normative predictability of the critical pre-nominal indefinite articles and nouns was operationally defined as the relative frequency of production in a sentence fragment completion task (cloze probability) in separate testing with individuals who did not participate in the EEG experiments. Participants in the EEG studies were healthy young adult right-handed native English speakers. Salient differences between the EEG experiments include the number of participants and experimental items (Table 1), the presentation mode (one vs. two sentences per trial), experimental conditions (± prenominal adjectives, ± filler items), counterbalancing scheme, the distribution of cloze probabilities, and normative plausibility of critical nouns (see SI Appendix, Table S1 Synopsis: Experiments 1, 2, and 3). In all three EEG Experiments, sentences containing the critical prenominal articles were read word-by-word at a fixed rate approximately 2 per second and the EEG data acquisition and data processing procedures were the same (see SI Appendix, EEG recording and data processing). Prior to modeling, the EEG data were visually screened for artifacts, smoothed (25 Hz low-pass phase compensated FIR), downsampled to 125 samples per second, centered by subtracting the mean of the 1496 ms prestimulus interval for each channel and re-screened for EEG artifacts by computer algorithm (see SI Appendix, EEG Experimental Procedures for details and OSF: udck19_pipeline_1.html for exclusions tabulated by experiment, participant, and item).

LMER model fitting. For the data pooled across the three experiments, each observation was coded for the experiment, subject, and stimulus item. Each item corresponds to the context prior to the critical article and provides one cloze value for *a* and one for *an* (see SI Appendix, fig. S2 for the distributions of article cloze across and within each design). Prior to modeling the EEG, the article cloze predictor variable was scaled from proportions of response (0.0 - 1.0) to standardized units ("z-scores") by centering and dividing by the standard deviation. The 1.2e4 screened single-trial EEG epochs were stacked into a dataframe (4.5e6 rows = 1.2e4 epochs × 375 samples / epoch), each row indexed for epoch and time stamped relative to article onset, with the experiment, subject, item, standardized cloze values, and the 26 EEG channels in columns. To model these single-trial data we used `fitgrid` (49), an open-source Python package we developed in the lab that implements mixed-effects model fitting via the pymer4 (50) interface to the lmerTest (51) and lme4 (39) R packages (52). With fitgrid, we swept each LMER model in Table 2 across 3 s epochs of data with the critical article in the middle (375 time points, 8 ms intervals = 125 samples/second; 26 electrode locations spaced about 5 cm apart) and collected the lme4::lmer() profiled maximum likelihood fits (REML=FALSE) in a tabular grid. From this grid of model fits, we extracted summary measures returned by lmerTest::lmer( ) for the fit at each time and channel including Akiake Information Criterion, $\hat{\beta}_j$ estimates for the intercept and article cloze lmerERPs and their 95% Wald confidence intervals, and fitting algorithm warnings (49, fitgrid.lmer). The $\hat{\beta}_{\text{cloze}}$ lmerERPs in Figure 3B and interval mean amplitude coefficents in Figure 4 for standarized cloze may be converted to coefficients $\hat{B}_{\text{cloze}}$ on the original cloze scale (μV/cloze) as $\hat{B}_{\text{cloze}} = \hat{\beta}_{\text{cloze}}/SD_{\text{cloze}}$ with the article cloze SD values in Table 1.

Data deposition. Stimulus materials, aggregated behavioral and EEG data, summary measures, data analysis software, and reproduction recipe are deposited in the Open Science Foundation repository OSF: UDCK and licensed under the CreativeCommons Attribution-NonCommercial-NoDerivatives 4.0 International License which may be viewed here: http://creativecommons.org/licenses/by-nc-nd/4.0. Behavioral and EEG data related to an identifiable natural person are maintained under control of the P.I. M.K. and Co-Investigators. Contact the corresponding author for information about further use of the research materials or access to privacy-sensitive data under a written data sharing agreement.

ACKNOWLEDGMENTS. We would like to thank Anna Stoermann for her painstaking work preparing the normative and EEG data and report; Andrey Portnoy for his key original contributions designing and implementing the fitgrid software used for these analyses and Lauren Liao for her work on early prototypes; Nathaniel Smith for sharing his code for reading the lab's EEG data; Melissa Troyer, Anna Stoermann, Reina Mizrahi, Seana Coulson, Megan Bardolph, and Emily Provenzano for critical discussion of EEG data analysis, modeling, and testing; Gregory A. Miller and Kara D. Federmeier for their comments on earlier drafts. This research was supported by the National Institutes of Health Grants 5R01HD022614-27 and 1R01AG048252-01A1.

1. MacDonald MC, Hsiao Y (2018) *Sentence Comprehension*, eds. Rueschemeyer SA, Gaskell MG. (Oxford University Press), 2nd edition.
2. Altmann GTM, Mirkovic J (2009) Incrementality and prediction in human sentence processing. *Cognitive Science* 33(4):583–609.
3. Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31(1):32–59.
4. Kutas M, DeLong KA, Smith NJ (2011) A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future* 190207.
5. Cooper RM (1974) Control of eye fixation by meaning of spoken language: New methodology for real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6(1):84–107.
6. Tanenhaus MK, Spiveyknowlton MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217):1632–1634.
7. Altmann GTM, Kamide Y (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73(3):247–264.
8. Wicha NYY, Bates EA, Moreno EM, Kutas M (2003) Potato not pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters* 346(3):165–168.
9. Wicha NYY, Moreno EM, 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.
10. Wicha NYY, Moreno EM, 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.
11. Van Berkum JJA, Brown CM, 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.
12. Otten M, Nieuwland MS, van Berkum JJA (2007) Great expectations: Specific lexical anticipation influences the processing of spoken language. *Bmc Neuroscience* 8.
13. Otten M, Van Berkum J (2008) Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes* 45(6):464–496.
14. Otten M, Van Berkum JJA (2009) Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research* 1291:92–101.
15. Kochari AR, Flecken M (2019) Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language Cognition and Neuroscience* 34(2):239–253.
16. Szewczyk JM, 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.
17. Nicenboim B, Vasishth S, Rösler F (2020) Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia* 142:107427.
18. DeLong KA, Chan WH, Kutas M (2019) Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology* 56(4):14.
19. DeLong KA, Urbach TP, Kutas M (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* 8(8):1117–1121.
20. DeLong KA (2009) Doctoral dissertation (University of California, San Diego).
21. DeLong KA, Groppe DM, Urbach TP, Kutas M (2012) Thinking ahead or not? natural aging and anticipation during reading. *Brain Lang* 121(3):226–39.
22. Martin CD, et al. (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.
23. Ito A, Martin AE, Nieuwland MS (2016) How robust are prediction effects in language comprehension? failure to replicate article-elicited n400 effects. *Language, Cognition and Neuroscience* pp. 1–12.
24. Nieuwland MS, et al. (2018) Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7:e33468.
25. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *Plos Biology* 13(3).
26. Nelson LD, Simmons J, Simonsohn U (2018) *Psychology's Renaissance*, Annual Review of Psychology, ed. Fiske ST. Vol. 69, pp. 511–534.
27. Gelman A, Loken E (2014) The statistical crisis in science. *American Scientist* 102(6):460–465.
28. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* 12(5):535–540.
29. Tukey JW (1969) Analyzing data: Sanctification or detective work. *American Psychologist* 24(2):83–91.
30. Tukey JW (1972) Data analysis, computation and mathematics. *Quarterly of Applied Mathematics* 30(1):51–65.
31. Tukey JW (1980) We need both exploratory and confirmatory. *The American Statistician* 34(1):23–25.
32. Tukey JW (1986) Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning Badmandments in *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1949-1964*, ed. Jones LV. (CRC Press) Vol. III, pp. 187–389.
33. Behrens JT (1997) Principles and procedures of exploratory data analysis. *Psychological Methods* 2(2):131–160.
34. Cohen J, Cohen P, West S, Aiken L (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. (Lawrence Erlbaum Associates), 3rd edition.
35. Fox J (2008) *Applied regression analysis and generalized linear models*. (Sage Publications).
36. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) *Applied linear statistical models*. (McGraw-Hill Irwin Boston) Vol. 5.
37. Smith NJ, Kutas M (2015) Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*.
38. Urbach TP, DeLong K, Chan W, Kutas M (2019) Empirical support for word form prediction during word-by-word reading: Osf project repository doi:10.17605/OSF.IO/TKSUR.
39. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *2015* 67(1):48.
40. Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4):390–412.
41. Burnham K, Anderson D (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. (Springer-Verlag), 2nd edition.
42. Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* 33(2):261–304.
43. Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3):255–278.
44. Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D (2017) Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94:305–315.
45. Kutas M, Hillyard SA (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307(5947):161–163.
46. Gelman A (2005) Analysis of variance: Why it is more important than ever. *Annals of Statistics* 33(1):1–31.
47. Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical models*. (Cambridge university press).
48. Tukey J (1977) *Exploratory Data Analysis*. (Addison-Wesley Publishing Company, Reading, MA).
49. Urbach T, Portnoy AS (2019) fitgrid 0.4.8, DOI:10.5281/zenodo.3581504.
50. Jolly E (2018) Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software* 3(31).
51. Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13):1–26.
52. R Core Team (2019) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).