Are word predictability effects really linear? A critical reanalysis of key evidence

C. Shain[1], C. Meister[2], T. Pimentel[3], R. Cotterell[2], R. Levy[1] [1]MIT, [2]ETH, [3]Cambridge

In the effort to understand the cognitive processes involved in human language comprehension, there has emerged a divide between *procedural* theories that focus on storage, retrieval, and structure building in working memory (e.g., [6]) and *inferential* theories that focus on rational allocation of probability among possible sentence interpretations (e.g., [3]). These two views stress different aspects of the comprehension problem (respectively, representations are computed vs. which representations to compute). They are thus potentially reconcilable (e.g., [8]), and their empirical predictions are often aligned [5]. Nevertheless, these views diverge in important ways. One prominent such divergence concerns whether processing cost is linear or logarithmic in word predictability [11]: procedural theories tend to construe predictability effects as reflecting preactivation and predict a linear effect [1], but inferential theories construe these effects as reflecting probabilistic updates and predict a logarithmic effect [3] (see [9] for extended discussion). All naturalistic reading studies of this question to-date [11, 4, 7, 12, 14, 9] have supported the (super)logarithmic effect predicted by the inferential view. However, Brothers & Kuperberg [1] investigated this question using a controlled experiment based on cloze norms, and reported a linear effect. Thus, different experimental approaches favor different conclusions, making it difficult to synthesize across the relevant body of evidence.

To address this issue, we reanalyze data from the self-paced reading (SPR) and cross-modal picture naming experiments of [1][1] using not only the cloze and trigram predictability estimates from the original study, but also estimates from GPT-2-small; this was the best-performing model of [9], who compared it to other language models in several datasets and to cloze probabilities on dataset [7]. We conduct GAM-based analyses [13] with both raw-probability and (logarithmic) surprisal scales on both experiments, as in the original study, but we prioritize SPR, since this task more closely resembles incremental language comprehension. We find a marked difference in results (**Fig 1A**). The cloze effect indeed shows a linear decrease on a raw-probability scale and plateaus on a surprisal scale, but the GPT (and trigram) effect shows a sharp superlinear increase as raw probability approaches zero and a linear increase on a surprisal scale over 8 or more nats of surprisal. In cross-validated permutation tests of model generalization, both experiments show a significant effect of GPT surprisal over and above GPT probability but not vice versa, and we find no significant degradation in the key SPR experiment from using GPT predictability relative to cloze. We explore the relationship between cloze and GPT predictability and find important differences: (i) although the item ranks from both estimates are correlated ($r = 0.55$), there is also substantial reorganization of relative predictability across items (**Fig 1B**), (ii) GPT improves on cloze in key cases, as evidenced by the existence of high cloze items with low GPT predictability and long reading times (**Fig 1C**) and the small but significant tendency for items ranked more surprising by GPT to have longer reading times (**Fig 1D**), and (iii) qualitative analysis of GPT's predictions suggests sensitivity to continuations whose human-subjective probability is likely under-estimated by cloze, including coordination, modifiers, metaphors, and effects of plausible extrasentential context (**Table 1**). Our results show that the linear effect found by [1] depends critically on analysis choices, and their data in fact favor the opposite conclusion when reanalyzed using methods more similar to those used in naturalistic studies. Given (i) the preponderance of naturalistic evidence in favor of logarithmic predictability effects and (ii) our finding that the data from [1] are also consistent with this conclusion when analyzed differently, we argue that the best available synthesis of the evidence is that predictability effects primarily reflect the costs of probabilistic inference, rather than predictive preactivation.

[1] Brothers & Kuperberg's meta-analysis was not reanalyzed due to lack of publicly available item-level data.

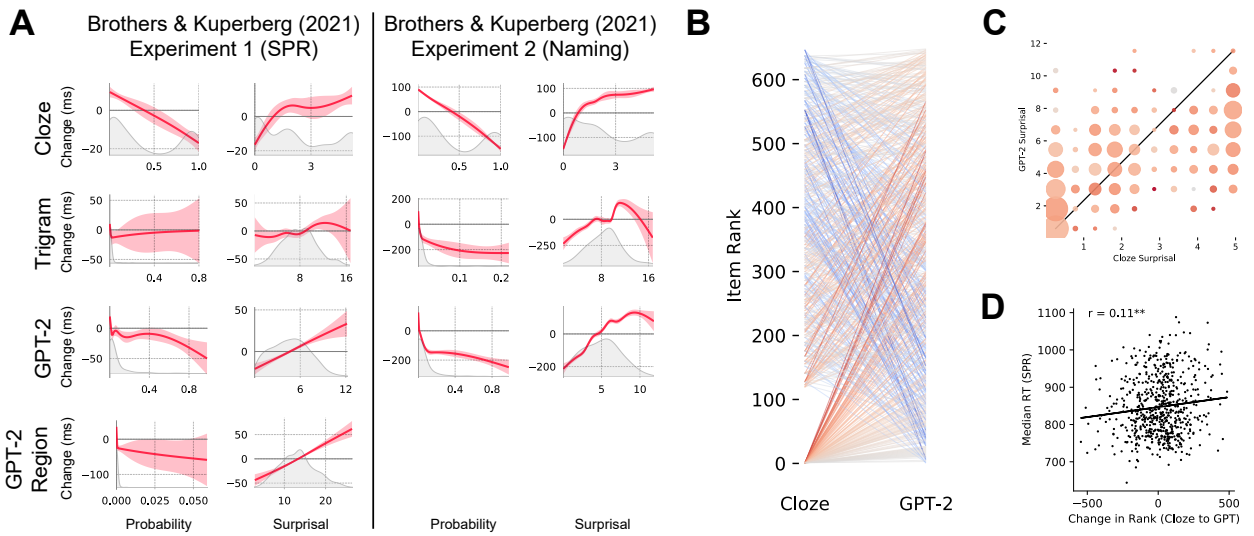

Figure 1: **A.** GAM-estimated functional form of effects in Brothers' & Kuperberg's [1] Experiment 1 (SPR) and Experiment 2 (cross-modal picture naming) across language model types (human cloze and trigram from the original study, GPT-2 predictability of the critical word, and GPT-2 predictability of the whole 3-word critical region—only relevant to SPR). Plots cover the full empirical range of predictor values in each training dataset. Kernel density plots show the distribution of predictor values in the training data over the plotted range. Uncertainty intervals show the $\pm$ 2 standard errors used as a default in plots from `mgcv` [13]. **B.** Change in rank for Brothers' and Kuperberg's 648 stimuli as a function of cloze probability (left) vs. GPT-2 probability (right). Color reflects the slope of the change in rank (red for positive, blue for negative). **C.** The distribution of items and reading latencies in Brothers and Kuperberg's SPR experiment as a function of cloze surprisal vs. GPT-2 surprisal. Point size represents the number of items that fall into the corresponding region of the cloze surprisal vs. GPT-2 surprisal coordinate space. Point color represents the median RT of those items in the SPR experiment (darker red indexes longer RTs). **D.** Median RT by item in Brothers and Kuperberg's SPR experiment as a function of the item's rank change in surprisal as estimated by GPT-2, relative to cloze (positive x-axis value means the item was ranked higher in surprisal by GPT-2 than by cloze).

Coordination	*The web had been spun by the large **and*** (high-cloze: *spider*)
Modifiers	*My son saw the bird pecking at the soil using its **own*** (high-cloze: *beak*)
Metaphors	*The web had been spun by the large **multinational*** (high-cloze: *spider*)
Plausible extrasentential context	*Before proposing to his girlfriend he would need a **passport*** (high-cloze: *ring*) (e.g., perhaps the partners live in different countries)

Table 1: Example alternative continuations (in **bold**) predicted by GPT-2-small for Brothers & Kuperberg's [1] high-cloze contexts. To generate these examples, we passed Brothers & Kuperberg's contexts through GPT and queried the top-10 most probable continuations, which we then qualitatively analyzed for patterns. These examples illustrate the diversity of linguistic inputs that a general-purpose sentence comprehension system must contend with during typical reading, diversity that models must also contend with in training but that cloze productions may not represent well due to task-specific strategic effects [10], such as a prototypicality bias [2].

References

[1] Brothers, T. and Kuperberg, G. R. *JML*, 2021.

[2] Frade, S., Santi, A., and Raposo, A. *Behavior Research Methods*, 2023.

[3] Hale, J. In *NAACL*, 2001.

[4] Hoover, J. L., Sonderegger, M., Piantadosi, S. T., and O'Donnell, T. J. *Open Mind*, 2023.

[5] Levy, R. *Cognition*, 2008.

[6] Lewis, R. L. and Vasishth, S. *Cognitive Science*, 2005.

[7] Luke, S. G. and Christianson, K. *Cognitive Psychology*, 2016.

[8] Rasmussen, N. E. and Schuler, W. *Cognitive Science*, 2018.

[9] Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. P. *PNAS*, to appear.

[10] Smith, N. J. and Levy, R. In *CogSci*, 2011.

[11] Smith, N. J. and Levy, R. *Cognition*, 2013.

[12] Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and Levy, R. P. *TACL*, 2023.

[13] Wood, S. N. *Generalized Additive Models: An Introduction with R*, 2006.

[14] Xu, W., Chon, J., Liu, T., and Futrell, R. In *Findings of EMNLP*, 2023.