Cognition 215 (2021) 104848

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Discussion

How 'rational' is semantic prediction? A critique and re-analysis of Delaney-Busch, Morgan, Lau, and Kuperberg (2019)

Mante S. Nieuwland [a,b,*]

[a] Max-Planck-Institute for Psycholinguistics, The Netherlands
[b] Donders Centre for Cognition, Brain and Behaviour, The Netherlands

ARTICLE INFO

Keywords
Semantic priming
Expectation adaptation
Rational adaptation
Bayesian adaptation
Probabilistic prediction
N400
Forward association

ABSTRACT

In a recent article in *Cognition*, Delaney-Busch et al. (2019) claim evidence for 'rational', Bayesian adaptation of semantic predictions, using ERP data from Lau, Holcomb, and Kuperberg (2013). Participants read associatively related and unrelated prime-target word pairs in a first block with only 10% related trials and a second block with 50%. Related words elicited smaller N400s than unrelated words, and this difference was strongest in the second block, suggesting greater engagement in predictive processing. Using a rational adaptor model, Delaney-Busch et al. argue that the stronger N400 reduction for related words in the second block developed as a function of the number of related trials, and concluded therefore that participants predicted related words more strongly when their predictions were fulfilled more often. In this critique, I discuss two critical flaws in their analyses, namely the confounding of prediction effects with those of lexical frequency and the neglect of data from the first block. Re-analyses suggest a different picture: related words by themselves did not yield support for their conclusion, and the effect of relatedness gradually strengthened in othe two blocks *in a similar way*. Therefore, the N400 did not yield evidence that participants rationally adapted their semantic predictions. Within the framework proposed by Delaney-Busch et al., presumed semantic predictions may even be thought of as 'irrational'. While these results yielded no evidence for rational or probabilistic prediction, they do suggest that participants became increasingly better at predicting target words from prime words.

1. How 'rational' or probabilistic is semantic prediction?

Delaney-Busch and colleagues report an exploratory re-analysis of data from Lau et al. (2013). Their aim was to demonstrate that prediction is an adaptive, probabilistic process based on rational Bayesian principles, whereby people capitalize on statistical regularities in their environment to guide behavior (e.g., Anderson, 1990). In this framework, 'rational' means to adapt predictive behavior to the probability of prediction success, engaging more in prediction when predictions are repeatedly fulfilled rather than disconfirmed.[1] Lau et al. (2013) reported some initial support for this hypothesis from the well-known semantic priming effect (e.g., Brown, Hagoort, & Chwilla, 2000; Holcomb, 1988), the N400 reduction for target words that are related to a preceding prime word compared to unrelated target words. Relatedness was operationalized as forward association strength from the South Florida Association Norms (Nelson, McEvoy, & Schreiber, 2004), the probability of someone producing the target word when asked to produce words associated with the prime. Lau et al. found a stronger relatedness effect on the N400 in a stimulus block with 50% related trials ('higher relatedness proportion' Block 2) than in a preceding block with only 10% related trials ('lower relatedness proportion' Block 1), similar to previous reports (e.g., Brown et al., 2000). Lau et al. concluded that "predictions about upcoming stimuli make a substantial contribution to the N400 effect".

Delaney-Busch et al. re-analyzed the Lau et al. data[2] using a rational adaptor model. This model computes and updates the probability of encountering target words on individual trials throughout Block 2 (higher relatedness proportion), given a prior probability based on Block 1 (lower relatedness proportion), and combines this probability with the forward association strength from prime to target and the lexical

* Corresponding author at: Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.
E-mail address: mante.nieuwland@mpi.nl.

[1] Although not assumed within this framework, one potential argument for rationality is that predictions are costly (e.g., require metabolic resources) or hinder comprehension when disconfirmed.

[2] Not mentioned by Delaney-Busch et al., this re-analysis only included 31 of the 32 participants in Lau et al. (2013).

https://doi.org/10.1016/j.cognition.2021.104848
Received 18 March 2020; Received in revised form 8 May 2021; Accepted 11 July 2021

frequency of the target (SUBTLEX corpus; Brysbaert & New, 2009). The negative log-transform of this word probability yields a measure of 'surprisal' (e.g., Hale, 2001; Levy, 2008; see also Aurnhammer & Frank, 2019):

Surprisal = $-\log_2$ [μ*p(word|prime) + (1-μ)*p(word|average context)]

wherein parameter μ is a point estimate of the probability with which a rational adapter expects a related trial at that point in time (mean value of the associated beta distribution), p(word|prime) is forward association strength, and p(word|average context) is lexical frequency. In this model, increasing the proportion of related trials will strengthen the expectation of a related trial (μ), thereby increasing the effect of forward association strength and decreasing the effect of lexical frequency. In effect, increasing the proportion of related trials yields lower surprisal for related trials. For unrelated words, p(word|prime) equals zero,[3] which yields higher surprisal overall than for related words.

Delaney-Busch and colleagues present a series of analyses, some of which have to do with justifying parameter settings of the model. Their key finding was that lower surprisal in Block 2 was associated with smaller N400 amplitude, over and above the effect of relatedness and other variables. They concluded that the brain rationally adapts its probabilistic semantic predictions to the broad statistical structure of its environment.

1.1. A critique of Delaney-Busch et al

The results and conclusions of Delaney-Busch et al. may seem intuitively plausible, but one important aspect of their hypothesis remains unaddressed. While they test the effect of surprisal for related and unrelated words together, the key issue really is *whether this effect occurs for related words alone*. At least two arguments should be considered.

According to the rational semantic prediction hypothesis, participants predict more strongly when they encounter a greater proportion of related words (i.e. the increasing μ in Block 2), and this increase in prediction further facilitates access to the meaning of related words. The most direct test of the rational semantic prediction hypothesis is therefore whether related words elicit smaller (less negative) N400 responses in Block 2 as a function of surprisal. In contrast, whether or not participants strengthen their predictions is likely of little consequence for N400 responses to unrelated words. This is because N400 responses to unexpected words do not depend on whether the context afforded a strong prediction (e.g., Rommers & Federmeier, 2018; for a review, see Kutas & Federmeier, 2011; Van Petten & Luka, 2012); failed predictions are known to neither facilitate nor hinder access to word meaning. The importance of examining effects for related and unrelated words separately was already recognized by Lau et al., who performed pairwise follow-ups on the interaction between relatedness and block. Related words elicited smaller N400s in the higher relatedness proportion block compared to the lower proportion block, but unrelated words elicited similar N400 responses in the two blocks.

The second argument for looking at related words alone is that the effect of surprisal on unrelated words is tantamount to an effect of lexical frequency. This follows from the surprisal formula, because for unrelated words, p(word|prime) equals zero, such that surprisal equals (1-μ)*p(word|average context). According to Delaney-Busch et al., this explains why surprisal fluctuates more strongly for unrelated trials than for related trials.[4] Whereas surprisal of related words depends strongly on forward association strength and only varies between 0.5 and 1, surprisal of unrelated words depends strongly on frequency. Although in theory μ also impacts the effect of lexical frequency, this impact is negligible because frequency varies across many orders of magnitude. For this reason, lexical frequency is the primary determinant of surprisal for unrelated words, as further demonstrated graphically in Fig. 1.

Perhaps needless to say, the correlation between lexical frequency and N400 amplitude is already well-established (e.g., Kutas & Federmeier, 2011, for a review). Crucially, Delaney-Busch et al. present claims about rational semantic prediction, but the supporting evidence is strongly driven by this correlation that is essentially of no theoretical interest.

For these reasons, there is an a priori theoretical argument for testing the effects of surprisal separately for related and unrelated words. Such tests also circumvent the problem with high multicollinearity between relatedness and surprisal in Delaney-Busch et al. (e.g., correlation of fixed effects = −0.963; in each of the reported tests, the Variance Inflation Factor for surprisal was over 13; see also Zuur, Ieno, & Elphick, 2010). In their model with relatedness and surprisal, higher surprisal elicits larger (more negative) N400s (b = −2.21, S.E. = 0.80, $t = 2.76$, $p = 0.006$), but this estimate is about twice greater than the estimate from a model where surprisal is the only predictor. Furthermore, in all their reported analyses, the sign of the estimates for unrelated trials is incorrect. While Delaney-Busch and colleagues acknowledge this issue, they claim that the associated *p*-values ("marginal significance") can be interpreted without a problem, which is unnecessarily hazardous.

To provide a direct test of the original hypothesis about strengthening predictions for related words, I tested the effect of surprisal for related and unrelated items separately. An effect of surprisal for unrelated words is of little interest because it corresponds to a lexical frequency effect. Crucially, the litmus test of the rational semantic prediction hypothesis is an effect of surprisal for related words alone, with N400 responses that gradually become smaller (more positive) depending on prediction success.

2. Re-analysis results

For related words alone, Delaney-Busch's main analysis (p.14) yielded a non-significant effect in the expected direction, b = −3.28, S.E. = 3.89, $t = 0.84$, $p = 0.40$. The model with a precision parameter optimized to fit the rational adaptor model ($\nu = 77$) yielded a similar result, b = −3.24, S.E. = 3.66, $t = 0.88$, $p = 0.38$. Furthermore, from all three analyses with which Delaney-Busch et al. showed an effect of surprisal over and above that of other variables (relatedness, forward association strength, word probability), none yielded a statistically significant effect for related words alone.

Unlike the related items, the unrelated items showed a statistically significant surprisal effect in the expected direction in all of the reported analyses.[5] For example, the main model that only included surprisal yielded b = −2.13, S.E. = 0.83, $t = 2.56$, $p = 0.016$). However, as I

[3] Delaney-Busch et al. and Lau et al. use forward association strength in a way that is similar to cloze completion values, a measure of lexical predictability that is zero for unpredictable words. However, both measures do not capture broader semantic relationships between prime and target that can impact N400 amplitude (e.g., Van Petten, 2014; for discussion, see Nieuwland et al., 2020).

[4] Delaney-Busch et al. state that "each participant's idiosyncratic ordering of critical trials causes larger fluctuations in model outputs for unrelated than for related trials". However, in fact, there was no idiosyncratic trial order for each participant and the study used fixed-order trial lists, and a given unrelated target word always appeared in the same position in Block 2.

[5] Analyses for unrelated words that include both surprisal and either forward association strength or lexical frequency or both are problematic, however: models with surprisal and forward association strength are rank deficient because forward association strength for unrelated words is zero, whereas models with surprisal and lexical frequency suffer from multicollinearity because surprisal is almost perfectly related to lexical frequency.

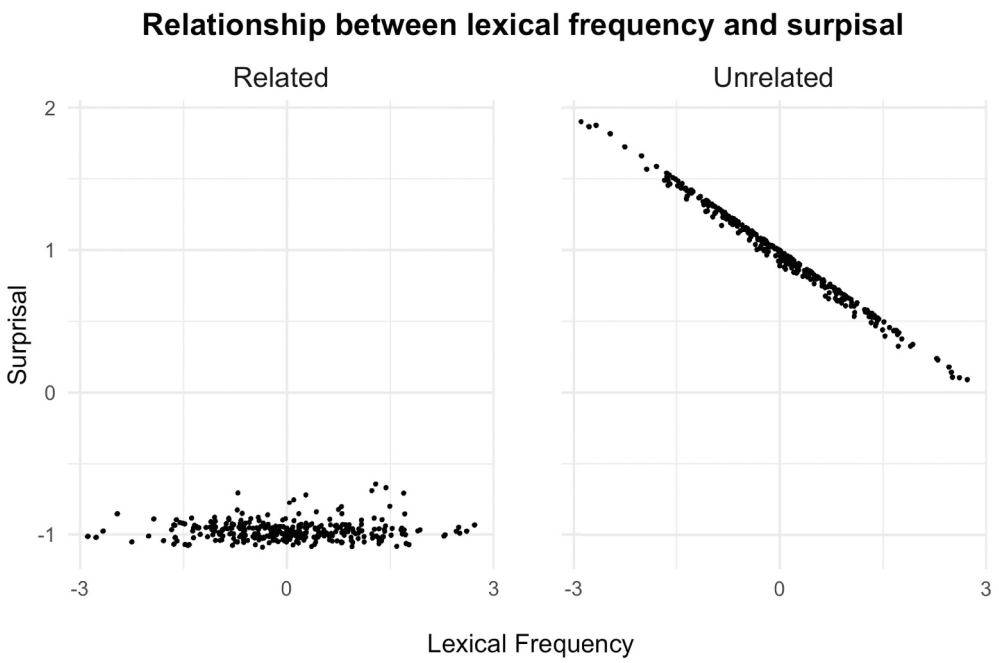

Fig. 1. Scatter plots depicting the relationship between (log-transformed) lexical frequency and surprisal in the Delaney-Busch et al. data, for related and unrelated words separately. Surprisal was computed with precision parameter $\nu = 50$ and z-transformed using data from related and unrelated trials together. For unrelated words, surprisal is highly correlated with lexical frequency.

explained in the introduction, this effect is primarily driven by lexical frequency and has little to do with strengthening predictions in the experiment.

To test the actual contribution of surprisal in a different way, while avoiding multicollinearity issues, I compared the surprisal model with alternative models (forward association strength model, relatedness model) in terms of statistical fit. Each model included an interaction with lexical frequency to capture the different contributions of frequency for related and unrelated words. The fits differed by less than one AIC/BIC point and therefore did not clearly favour one model over another (e.g., Raftery, 1995). Moreover, AIC/BIC model comparisons for related words separately also show that surprisal does not yield a better model fit than forward association strength. Model comparisons for unrelated words further confirmed that surprisal does not yield a better model fit than frequency.

To summarize, these new results from Block 2 shed doubt on the conclusion that participants in Lau et al. adapted their semantic predictions to the environment in a rational, Bayesian fashion. According to the rational semantic prediction hypothesis, one should observe trial-by-trial adaptation in semantic prediction on related trials, such that associated N400 responses became gradually smaller (less negative) as a function of prediction success. Although the effect went in the expected direction, related trials alone did not yield a statistically significant effect of surprisal. Instead, it appears that the key finding from Delaney-Busch et al. was primarily driven by N400 responses to unrelated trials. However, this pattern is merely an effect of a word property (lexical frequency), and has nothing to do with adaptation in prediction. Moreover, once the multicollinearity problems in DeLaney-Busch et al. are dealt with, the N400 data from Lau et al. is as well, hence more parsimoniously explained without invoking Bayesian adaptation, contra the Delaney-Busch et al. conclusion.

Of course, these new results also raise new questions. If not rational, probabilistic adaptation of prediction, then what gives rise to the relatedness proportion block effects in Lau et al.? One potential answer is that adaptation occurred gradually throughout the entire experiment *irrespective of relatedness proportion*. Lau and colleagues briefly discussed this potential pattern (p. 495), but did not test for it. Delaney and colleagues discussed the importance of Block 1 in setting the prior for Block 2, but did not report effects from Block 1. To address this issue, I therefore re-analyzed data from both blocks together[6] to examine effects of within-experiment or with-block trial position.

2.1. Analysis of Block 1 and 2

Visual inspection of N400 responses for related and unrelated trials during the experiment already suggests that the divergence between these two conditions started before onset of Block 2 (Fig. 2a). Statistical analysis showed that the effect of forward association strength gradually increased throughout the experiment (see Fig. 2b), such that the N400 effect associated with 1 standard deviation in forward association strength differed by about 1.31 μV (S.E. = 0.47) between the beginning and end of the experiment ($t = 2.79$, $p = 0.006$).

To compare the effect of block directly, I tested the 3-way interaction analysis between forward association strength, block and within-block trial position. As shown in Fig. 2c, the effect of forward association strength gradually increased throughout the blocks (b = 0.97, S.E. = 0.48, $t = 2.01$, $p = 0.05$), but the estimate for the 3-way interaction term was approximately 6 times smaller (b = 0.17, S.E. = 0.89, $t = 0.19$, $p = 0.85$), yielding no evidence that the interaction between forward association strength and trial position differed between the two blocks.

2.2. Linear versus nonlinear effects of trial position

During peer-review, a co-author of Delaney-Busch et al. raised the concern that the relationship between trial position and N400 activity might not be linear. This contrasts with surprisal, which drops across trials in Block 2 in a non-linear fashion (i.e., it drops steeply at the beginning of the block but does not change much towards the end, see DB19's Fig. 3). Delaney-Busch et al. argue that their adaptor model

[6] The experiment was divided into 8 runs separated by breaks, of which the first 4 runs were always lower relatedness proportion blocks. The block manipulation was not mentioned to participants in the instruction.

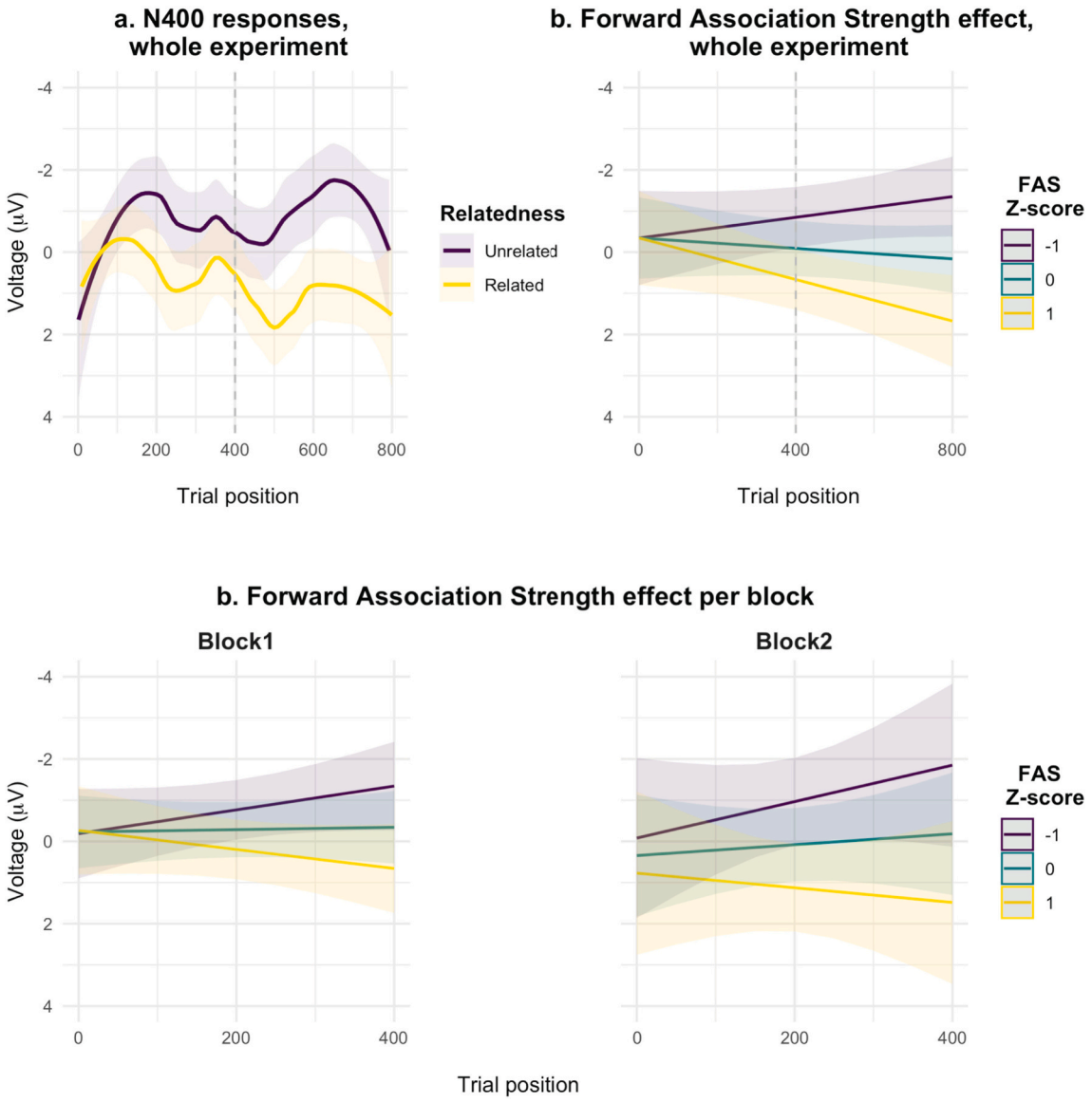

Fig. 2. N400 effects as a function of trial position in Lau et al. (2013). Fig. 2a plots N400 responses for related and unrelated trials, and is equivalent to a plot from Delaney-Busch et al.'s online code that was not in the published article. The regression is fitted with 'Local Polynomial Regression Fitting' and t-value based confidence bounds. Fig. 2b plots fitted responses (marginal effects) from the statistical model that tested whether the effect of forward association strength changed during the experiment. For demonstration purposes, the plot shows the effects of trial position at three levels of z-transformed forward association strength, namely the mean and one standard deviation below and above the mean. Fig. 2c plots fitted responses from the model that tested whether the effects of forward association strength increased in the same manner during Block 1 and 2.

"correctly predicts an increase in the divergence between related and unrelated trials over the course of Block 2, particularly within the first approximately 100 trials.[7]" However, this claim about the N400 cherry-picks a small part of the time series data, whereas data patterns that did not fit their conclusion were left unexplained. It is not clear that the data is best captured by a nonlinear pattern wherein the N400 rapidly decreases for related trials at the start of Block 2.

To address this issue, I performed a linearity diagnostic check (Long, 2019) for the analysis on the three-way interaction between relatedness, block and trial position, including relatedness as categorical predictor (for simplicity). This is not an exhaustive time-series analysis (for example, using generalized additive mixed-effects modelling, e.g., Baayen, Vasishth, Kliegl, & Bates, 2017; but see also Thul, Conklin, & Barr, 2021, for a discussion and critique of its application), but it can nevertheless yield initial insights regarding patterns and sources of potential nonlinearity (Hainmueller, Mummolo, & Xu, 2019).

The check is visualized in Fig. 3, which superimposes linear regression lines and locally weighted regression lines (Cleveland & Devlin, 1988) for related and unrelated words in each block. Deviations from linearity can be visually inspected by comparing the regression lines, stronger overlap indicates stronger evidence for linearity. Fig. 3 clearly shows 2 relevant patterns: if there is a nonlinear effect of trial position to begin with, this effect is (a) very similar for the two blocks and (b) primarily driven by the unrelated words. Therefore, the results yet again fail to support the main conclusion of Delaney-Busch et al., which rests on a nonlinear, rapid decrease in the N400 - specifically - for related words in the beginning of Block 2.

[7] As pointed out by a reviewer, one problem in Delaney-Busch et al. is that, due to the large number of filler trials, there are only few data points from these early trials, with low power as a result.

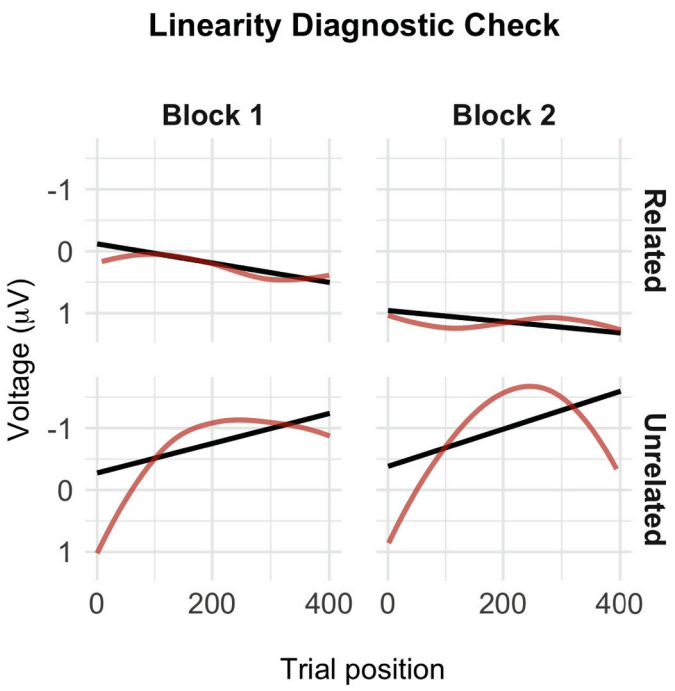

Fig. 3. Linearity diagnostic check. Voltage for related and unrelated words as a function of trial position within a block, plotted with linear regression lines (black) and locally weighted regression lines (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Discussion

The current re-analyses reveal patterns that are inconsistent with Delaney-Busch et al.'s main conclusion. Delaney-Busch et al.'s surprisal effect in Block 2 primarily reflects the effect of frequency for unrelated words, and related words by themselves yielded insufficient evidence for such an effect. Re-analyses of data from both blocks shed further doubt and suggested an alternative interpretation. If adaptation of semantic predictions is rational, such that participants adapt their semantic predictions to the probability of prediction success, they should initially predict target words very weakly or inconsistently throughout Block 1 (lower relatedness), and then start predicting target words more strongly or consistently once they enter Block 2 (higher relatedness), and vice-versa had the experiment counterbalanced block order and relatedness proportion. However, the current re-analyses show that the observed ERP effect of relatedness evolved in a similar fashion in Block 1 and 2.

While inconsistent with the conclusion of Delaney-Busch et al., these results could be compatible with the 'trial order' explanation mentioned by Lau et al. (2013), in which participants adapt their predictive strategies with the number rather than the proportion of related trials they encountered. In the framework of Delaney-Busch et al., such adaptation may be considered 'irrational' or non-probabilistic, because it is not driven by the changes in probability of prediction success. However, such adaptation could be considered rational in a colloquial sense: participants may have learned to predict target words from prime words and became better with practice.

A related and perhaps surprising finding is that forward association strength did not appear to have much of an effect in roughly the first 50 trials of the experiment (see Fig. 2). This may have to do with the fact that participants were instructed to perform an animacy-decision task, which required a button-press whenever a prime word was followed by an (unrelated) animal name. This task may have initially diverted attention away from the associative relationships in the stimuli, delaying participants' recognition of said relationships. Possibly, then, the beginning of the experiment does not yield prediction effects because participants did not adopt a predictive task strategy from the outset. If this is true, then it is to be expected that changing or removing the meta-linguistic judgment task impacts the gradually changing relatedness effect. When participants engage in predictive processing as soon as they start the experiment, changes throughout the experiment would be smaller and harder to observe.[8] This raises a more general question, namely whether the observed changes generalize to other experimental contexts and materials (see also Lau et al. and Delaney-Busch et al. for discussion). Learning to predict associated words could reflect an adaptation to the demands of a word-pair experiment. To my knowledge there is currently no consistent pattern from sentence comprehension studies that investigate effects of trial order on predictive processing.

Given the potential role of experimental task demands, one could also ask whether the changing effect is a genuine modulation of the N400 component or perhaps of another component like the P300 (e.g., (Nieuwland, 2019) for discussion). Even if the explicit task instruction about animal names did not mention word associations, frequent repetition of related trials may have caused participants to perceive implicit task demands. Associated task strategies may have also played a role in modulating ERP responses to the prime words during the experiment (see Fig. 7 in Lau et al.).

Questions beyond a task influence also remain. One question is whether N400 responses to related and unrelated words developed in a linear or nonlinear fashion during the experiment. Linearity diagnostic checks suggested that N400 responses to unrelated words followed a u-curve pattern within each block, but what this means remains uncertain. Another question concerns the relative contributions of associative versus 'merely' semantic relationships between prime and target. While these types of relationships may be confounded in the Lau et al. stimuli, they can be disentangled (e.g., 'cereal' and 'bread' are semantically but not associatively related), and previous research suggests that associative relationships contribute more strongly to 'semantic priming' effects on the N400 than non-associative semantic relationships (e.g. Ortu, Allan, & Donaldson, 2013; Rhodes & Donaldson, 2008; see also Van Petten, 2014). If the N400 effects like those reported here are due to changes in prediction, they would occur for associative relationships but not for non-associative semantic relationships.

Finally, a word of caution because the flaws in the Delaney-Busch et al. analyses will also apply to other studies using this paradigm. For example, Sharpe, Weber, and Kuperberg (2020) use the same paradigm and analyses on data from schizophrenia patients, who did not show effects of surprisal like control participants. Sharpe et al. claimed that "impairments in probabilistic prediction and Bayesian learning can explain reduced neural semantic priming in schizophrenia". However, the current analyses suggest a simpler explanation, namely that schizophrenia patients merely elicit weaker lexical frequency effects than control participants, a pattern that has already been established in previous research (e.g., Condray, Siegle, Keshavan, & Steinhauer, 2010).

4. Conclusion

I should emphasize that the patterns of results reported in this article do not strongly speak *against* Bayesian adaptation of semantic predictions writ large, but it also does not speak in favour of the conclusions proposed by Delaney-Busch et al. Follow-up research, possibly with alternative approaches, may provide the necessary evidence. Ideally, such research is not exploratory, like the analyses reported here and in Delaney-Busch et al., but confirmatory and planned with sufficient power and pre-registered analyses (e.g., Fleur, Flecken, Rommers, & Nieuwland, 2020; Nieuwland, Arkhipova, & Rodríguez-Gómez, 2020). For now, I conclude that participants may well adapt their semantic predictions during an experiment, but on the issue of whether

[8] I thank Rachel Ryskin for this suggestion.

participants in this particular experiment did so rationally and probabilistically, it seems that the jury is still out.

Acknowledgements

I thank Delaney-Busch and colleagues for providing data and analysis script on OSF at https://osf.io/dm2hr/. Code for the current article is available on OSF at https://osf.io/wzvp6/ and as a supplement. I thank Phillip Alday, Trevor Brothers, Stefan Frank, Gina Kuperberg, Andrea E. Martin, Emily Morgan, Joost Rommers, Elise van Wonderen and an anonymous reviewer for comments on a previous draft of this manuscript. For the analyses and plots, I used Rmarkdown (Xie, Dervieux, & Riederer, 2020) and the following packages for R (R Core Team, 2018): "lme4" (Bates, Maechler, Bolker, & Walker, 2015), "performance" (Lüdecke, Makowski, Waggoner, & Patil, 2020), "ggplot2" (Wickham, 2016), "dplyr" (Wickham, François, Henry, & Müller, 2019), "sjPlot" (Lüdecke, 2020), "patchwork" (Pedersen, 2020), "emmeans" (Lenth, 2019), "interactions" (Long, 2019).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia, 134*, 107198.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Brown, C. M., Hagoort, P., & Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. *Brain and Language, 72*, 158–190.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*(403), 596–610.

Condray, R., Siegle, G. J., Keshavan, M. S., & Steinhauer, S. R. (2010). Effects of word frequency on semantic memory in schizophrenia: Electrophysiological evidence for a deficit in linguistic access. *International Journal of Psychophysiology, 75*(2), 141–156.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition, 187*, 10–20.

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition, 204*, 104335.

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis, 27*(2), 163–192.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Association for Computational Linguistics.

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language, 35*(1), 66–85.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502.

Lenth, R. (2019). *Emmeans: Estimated marginal means, aka least-squares means. R package version 1.4.2*. https://CRAN.R-project.org/package=emmeans.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Long, J. A. (2019). *Interactions: Comprehensive, user-friendly toolkit for probing interactions. R package version 1.1.0*. <URL: https://cran.r-project.org/package=interactions.

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science. R package version 2.8.6*. https://CRAN.R-project.org/package=sjPlot.

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2020). Assessment of regression models performance. In *CRAN*. Available from https://easystats.github.io/performance/.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407.

Nieuwland, M. S. (2019). Do "early" brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews, 96*, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019.

Nieuwland, M. S., Arkhipova, Y., & Rodríguez-Gómez, P. (2020). Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex, 133*, 1–36. https://doi.org/10.1016/j.cortex.2020.09.007.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 375*, 20180522. https://doi.org/10.1098/rstb.2018.0522.

Ortu, D., Allan, K., & Donaldson, D. I. (2013). Is the N400 effect a neurophysiological index of associative relationships? *Neuropsychologia, 51*(9), 1742–1748.

Pedersen, T. L. (2020). *Patchwork: The composer of plots*.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111–163.

Rhodes, S. M., & Donaldson, D. I. (2008). Association and not semantic relationships elicit the N400 effect: Electrophysiological evidence from an explicit language comprehension task. *Psychophysiology, 45*(1), 50–59.

Rommers, J., & Federmeier, K. D. (2018). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage, 183*, 263–272.

Sharpe, V., Weber, K., & Kuperberg, G. R. (2020). Impairments in probabilistic prediction and Bayesian learning can explain reduced neural semantic priming in schizophrenia. *Schizophrenia Bulletin., 46*(6), 1558–1566.

Thul, R., Conklin, K., & Barr, D. J. (2021). Using GAMMs to model trial-by-trial fluctuations in experimental data: More risks but hardly any benefit. *Journal of Memory and Language, 120*, 104247.

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology, 94*(3), 407–419.

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

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation. R package version 0.8.3*. https://CRAN.R-project.org/package=dplyr.

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook. Chapman and hall/CRC*. ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14.