Title

Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming

Authors

Delaney-Busch, Nathaniel
Morgan, Emily
Lau, Ellen
et al.

Publication Date

2019-06-01

DOI

10.1016/j.cognition.2019.01.001

Peer reviewed

HHS Public Access

Author manuscript

Cognition. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

Cognition. 2019 June ; 187: 10–20. doi:10.1016/j.cognition.2019.01.001.

Neural Evidence for Bayesian Trial-by-Trial Adaptation on the N400 during Semantic Priming

Nathaniel Delaney-Busch[1], **Emily Morgan**[1,2], **Ellen Lau**[3], and **Gina R. Kuperberg**[1,4]

[1]Department of Psychology, Tufts University, USA

[2]Department of Linguistics, University of California, Davis, USA

[3]Department of Linguistics, University of Maryland, USA

[4]Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA

Abstract

When semantic information is activated by a context prior to new bottom-up input (i.e. when a word is predicted), semantic processing of that incoming word is typically facilitated, attenuating the amplitude of the N400 event related potential (ERP) – a direct neural measure of semantic processing. N400 modulation is observed even when the context is a single semantically related "prime" word. This so-called "N400 semantic priming effect" is sensitive to the probability of encountering a related prime-target pair within an experimental block, suggesting that participants may be adapting the strength of their predictions to the predictive validity of their broader experimental environment. We formalize this adaptation using a Bayesian learning model that estimates and updates the probability of encountering a related versus an unrelated prime-target pair on each successive trial. We found that our model's trial-by-trial estimates of target word probability accounted for significant variance in trial-by-trial N400 amplitude. These findings suggest that Bayesian principles contribute to how comprehenders adapt their semantic predictions to the statistical structure of their broader environment, with implications for the functional significance of the N400 component and the predictive nature of language processing.

Keywords

adaptation; language comprehension; N400; prediction; precision; expected uncertainty; unexpected uncertainty

1. Introduction

It has long been established that more predictable words are processed faster than less predictable words (e.g. Ehrlich & Rayner, 1981; Fischler & Bloom, 1979; see Staub, 2015

Corresponding Author:Emily Morgan, Department of Linguistics, 469 Kerr Hall, University of California Davis, One Shields Ave, Davis, CA 95616.

Supplementary Materials

Project data and scripts are available on Open Science Framework at: https://osf.io/dm2hr/

for a recent review). Rather than being all-or-nothing or strategic in nature, these effects of contextual predictability are graded, probabilistic and implicit (Luke & Christianson, 2016; Smith & Levy 2013; see Kuperberg & Jaeger, 2016 for a review). Probabilistic prediction can aid language processing by alleviating the resource bottleneck that could otherwise occur at word onset (because some of the "work" of comprehension can be accomplished ahead of time, given the information provided in the context). Such benefits, however, require that prediction is based on probabilistic knowledge that approximates the statistical structure of the input. This presents a challenge for communication in the real world where our linguistic and non-linguistic environments often change. Each person we talk to and every book we read has its own unique set of syntactic and semantic preferences. Thus, in order for language comprehension to remain efficient, we must be able to *adapt* to these different environments so that our predictions continue to mirror their statistical structures. In the present study, we explore the close relationship between probabilistic prediction and adaptation in the brain by modeling a classic effect of adaptation on lexico-semantic processing: the influence of the predictive validity of the experimental environment on the N400 semantic priming effect.

The fundamental link between prediction and adaptation has been widely discussed in cognitive science, dating back to early models of animal learning (Pearce & Hall, 1980; Rescorla & Wagner, 1972). One way of formalizing this link is within a probabilistic generative framework (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; see Perfors, Tenenbaum, Griffiths, & Xu, 2011, for an excellent introduction). Here, the agent's overarching goal is to infer an underlying latent cause that best explains the statistics of its environmental input. As the agent receives more input (evidence), she is able to incrementally update her probabilistic beliefs using Bayes' rule — a process known as *belief updating*.

In the domain of language, this type of probabilistic framework has most commonly been used to model incremental syntactic parsing (e.g. Levy, 2008), as well as to describe sentence comprehension more generally (Kuperberg, 2016; Kuperberg & Jaeger, 2016). In addition, it has recently been used to explain how we *adapt* to the broader set of statistical contingencies that are associated with, and define, any given situational context (e.g. Fine, Qian, Jaeger, & Jacobs, 2010; Jaeger & Snider, 2013; Kleinschmidt & Jaeger, 2015; Myslin & Levy, 2016), where it is referred to as "rational" adaptation (see Anderson, 1990).[1]

Neural indices of online processing have shown similar effects of predictability as behavioral measures, suggesting that probabilistic prediction is instantiated in the brain during language comprehension. A well-established effect of contextual probability on language processing is on the N400 — an event-related potential (ERP) that peaks between 300–500ms following the onset of an incoming word, and that is thought to reflect the ease of semantically processing that word (Federmeier, 2007; Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). The N400 is highly sensitive to the semantic probability of incoming

[1]In the present study, we use the term "rational" descriptively to refer to the use of Bayes' rule to update beliefs. As in any other Bayesian model, we can infer rationality only with respect to our assumptions about participants' priors, likelihoods and hypothesis spaces (see Tauber, Navarro, Perfors & Steyvers, 2017 for discussion). We return to reconsidering these assumptions in interpreting our data in the Discussion.

words (DeLong, Urbach, & Kutas, 2005; Wlotko & Federmeier, 2012): its amplitude is less negative ("smaller") to words that are semantically more (versus less) predictable. This is the case regardless of whether the context is a sentence stem (e.g. Kutas & Hillyard, 1984), a larger discourse or text (e.g. van Berkum, Zwitserlood, Hagoort, Brown, 2003), or a single 'prime' word (Bentin, McCarthy & Wood, 1985; Rugg, 1985).

There is also evidence that the amplitude of the N400 adapts to the statistics of its broader environment. A classic illustration of this is the effect of relatedness proportion on N400 modulation during a semantic priming paradigm (Brown, Hagoort, & Chwilla, 2000; Holcomb, 1988; Lau, Holcomb, & Kuperberg, 2013). Behaviorally, the Relatedness Proportion effect on semantic priming was first described in the late 1970s by Tweedy, Lapinski, & Schvaneveldt (1977), and it has since been reported in numerous studies (reviewed by Neely, 1991). It refers to the finding that the semantic priming effect is larger in blocks that contain a higher (versus a lower) proportion of related (versus unrelated) prime-target pairs. The effect has long been linked to predictive mechanisms (Hutchison, 2007; Keefe & Neely, 1990; Neely & Keefe, 1989; Neely, Keefe, & Ross, 1989): in higher relatedness proportion blocks, participants are more likely to use the prime to generate stronger lexico-semantic predictions of the target.

Following these behavioral studies, as well as previous ERP experiments (Brown et al., 2000; Holcomb, 1988), we recently carried out an ERP study examining the effect of Relatedness Proportion on the N400 semantic priming effect (Lau, Holcomb & Kuperberg, 2013). We measured ERPs as the same participants viewed the same core set of prime-target pairs, which were counterbalanced across two blocks. These blocks differed in the proportion of semantically related and unrelated word-pairs. In Block 1 (the lower relatedness proportion block), only 10% of the prime-target pairs were semantically related, and in Block 2 (the higher relatedness proportion block), 50% of the prime-target pairs were semantically related. Short breaks were given within both blocks as well as between blocks, and participants were not explicitly told that there would be any change between the blocks. We showed that the magnitude of the N400 semantic priming effect was significantly larger in Block 2 (the higher relatedness proportion block) than in Block 1 (the lower relatedness proportion block). In follow-up studies using MEG and fMRI, we also showed that the higher relatedness proportion block was associated with enhanced modulation of neuroanatomical regions sensitive to both lexico-semantic processing and learning (Lau, Weber, Gramfort, Hamalainen & Kuperberg, 2016; Weber, Lau, Stillerman & Kuperberg, 2016).

These findings provide strong evidence that participants were able to implicitly *adapt* to the changes in the predictive validity across the two blocks (see Tweedy & Lapinski, 1981, for an early discussion of adaptation in relation to this effect). What remains unclear, however, is the time course and the computational principles underlying such adaptation in relation to prediction. In this investigation, we sought to address this question by building a computational model based on principles of rational (Bayesian) adaptation. This model computed and updated the probability of encountering target words on individual trials throughout Block 2 (the higher proportion block), with the assumption that participants had already seen Block 1 (the lower proportion block). We then use linear mixed effects

regression to ask whether the trial-by-trial outputs of our computational model in each participant could explain changes in the trial-by-trial modulation of the actual N400 data collected in each participant throughout Block 2 in the dataset collected by Lau, Holcomb & Kuperberg, 2013.

In the remainder of this paper, we describe the theory and mathematical computation of our model. We then give a brief overview of the experimental methods previously described in detail by Lau, Holcomb & Kuperberg (2013). We evaluate our model's trial-by-trial output in each participant against the empirical trial-by-trial ERP data in each participant, and we then discuss our findings in the context of the broader literature on prediction, adaptation, and language processing.

2. Theory

2.1 Development of a rational probabilistic model of trial-by-trial adaptation

Our rational adaptor model considers how a comprehender makes probabilistic predictions during a semantic priming paradigm as she adapts to a higher relatedness proportion block (Block 2), following a lower relatedness proportion block (Block 1). By *probabilistic prediction*, we simply refer to the existence of a probability distribution over possible target words after seeing a prime on each trial.

To compute these probabilistic predictions on each trial, we assume that the agent is potentially able to draw upon two different types of long-term stored knowledge: her knowledge about semantic associations between words, and her knowledge about the frequency of words when encountered in isolation. These types of knowledge are, of course, not the only factors that influence the amplitude of the N400 amplitude; rather they are the two factors that we assume are most relevant to understanding how the N400 is modulated as the agent adapts during a semantic priming paradigm. We assume that, on each trial, the degree to which the agent uses each of these sources of long-term knowledge, in combination with the prime, to generate probabilistic predictions about the target, is weighted by the degree to which she believes that she will encounter a related or an unrelated target. These latter beliefs are updated, based on Bayes' rule, on successive trials as she progresses through Block 2. As a result, the model outputs a final estimate of the probability of encountering a target word on each individual trial in each participant within Block 2. These final probabilities are then negative log transformed to yield the information-theoretical measure *surprisal* on each trial in each participant.

This final model output on each trial is then tested against human trial-by-trial N400 data using linear mixed effects regression models. Below we step through the principles of the computational model and justify our assumptions at a conceptual level. Computational details are given in the following Calculation section.

2.2 Probabilistic predictions based on Forward Association Strength, Frequency and beliefs about trial type

In order to carry out a semantic priming task, we assume that participants can draw upon their stored knowledge about semantic associations between individual words to generate

predictions about a target on the basis of a prime. To index this semantic associative knowledge, we used the Forward Association Strength (FAS) of the prime, as estimated using the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 2004). These norms are derived from responses of a large number of participants who are given a "prime" word as a prompt and asked to produce the first associated word that comes to mind. The FAS is the proportion of participants who produced a particular target given the prime (see discussion of interpretation by Nelson, McEvoy, & Dennis, 2000), and it thus yields a probabilistic estimate of how likely a given target is produced having seen a given prime. Previous studies have shown that the FAS of a prime monotonically predicts the amplitude of the N400 produced by target words in semantic priming studies (Luka & Van Petten, 2014).

We also assume that, on any given trial, participants do not *only* base their predictions about a target on its prime's FAS, but that they also take into account their belief about whether they will encounter a related or an unrelated target following a prime (that is, whether the trial will be a related or an unrelated word-pair). This equates to their belief about whether the prime's FAS will be an *informative* predictor of the target. To take two extremes, if a participant believes with 100% probability that she is about to encounter a related word-pair, then she will be 100% confident that the prime will predict the target, and so she will base her probabilistic predictions about the target on the prime's FAS. If, on the other hand, the participant believes with 100% probability that she will encounter an unrelated word-pair, then she might ignore the prime (and the FAS of any potential target word given the prime) altogether. In this case, her predictions about the target will be based only on her stored distributional knowledge about the probability of seeing a target word in an average or random context. This is identical to word frequency: high frequency words are more probable than low frequency words, given an average context (Norris, 2006). Word frequency is also known to influence the magnitude of the N400 evoked by words presented in isolation, with more frequent (more probable) words eliciting a smaller N400 amplitude, with a logarithmic relationship (reviewed by Laszlo & Federmeier, 2014).

In a semantic priming paradigm, in which related and unrelated word-pair trials are presented in random order, the participant never knows in advance whether or not a target will be related or unrelated to the prime (whether the prime will be informative). However, at any point in the experiment, she may have some probabilistic estimate of how likely she will encounter a related versus an unrelated word-pair trial. Our model assumes that she uses this estimate as a blending factor that *weights* the relative influence of FAS versus word frequency knowledge to estimate the final probability of encountering any particular target. For example, if she is 100% confident that an upcoming the prime-target pair will be related, then, after encountering the prime word, "salt", she would estimate the probability of encountering "pepper" to be 0.7 – its FAS. If, however, given the wider contextual environment, she believes that the probability of encountering a related prime-target pair is only 0.1, then she might estimate the probability of encountering "pepper" following "salt" to be 0.07 (0.1*0.7) plus some very small probability of encountering it simply by chance as an unrelated word, as determined by its frequency. (Note that the probability of encountering "pepper" by chance as an unrelated target will be orders of magnitude lower than the probability of encountering it as a related target in this example.)

Importantly, we assume that the participant's estimate of the probability of encountering a related versus an unrelated word-pair is not static, but rather that it can change across the course of an experiment, as will her confidence in this estimate. A rational adapter framework provides a way to formalize this learning process. At any given point in the experiment, the participant has an initial prior belief about the probability of encountering a related versus an unrelated prime-target pair, with some degree of confidence in this belief. We assume that, at the very beginning of Block 2, this prior belief is based on the relatedness proportion within Block 1. Then, after encountering each prime-target pair within Block 2, the participant updates her belief about the relatedness proportion, using Bayes' rule, with the information learned from that trial. This new posterior distribution is then used to inform her belief about whether, on the next trial, she will encounter a target that is related or an unrelated to the prime. In this way, her beliefs about encountering a related versus an unrelated trial adapt incrementally over the course of the Block 2.

This dynamically changing belief about trial type (related or unrelated) then weights the relative influence of FAS and word frequency, so that, for each trial in Block 2, the model computes a final estimate of the probability of encountering the target word. Finally, this raw probability is log-transformed using the formula $-\log_2$[probability], which converts it into the information theoretic measure *surprisal* (Shannon & Weaver, 1949). We chose to carry out this final log transform on the basis of some empirical evidence that surprisal may be a better predictor than raw probability of behavioral measures of language processing difficulty, particularly at low estimates of probability (Hale, 2001; Levy, 2008; Smith & Levy 2013). In the ERP literature, there is also some evidence that surprisal predicts the amplitude of the N400 (Frank, Otten, Galli, & Vigliocco, 2015; Frank & Willems, 2017), although it is unclear whether it is a better predictor than raw probability (see analysis by Yan, Kuperberg & Jaeger, 2017 of data shared by Nieuwland et al., 2018, as well as response by Nieuwland et al., 2018).

Finally, we took the trial-by-trial output values yielded by our rational adaptor model in each participant, and we used linear mixed effects regression to ask whether these values accounted for trial-by-trial changes in the amplitude of the N400 evoked by target words measured over the course of Block 2 in each participant.

3. Calculation

3.1 Experimental Design

The experiment by Lau, Holcomb & Kuperberg (2013) crossed Relatedness (semantically related versus semantically unrelated word-pairs) and Relatedness Proportion (higher relatedness proportion versus lower relatedness proportion block). The related word-pairs had an FAS of 0.5 or higher (mean FAS: 0.65) as estimated using the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 2004), and the unrelated word-pairs were created by randomly redistributing the primes across the target items and checking to confirm that this did not accidentally result in any associated pairs. The Relatedness Proportion manipulation was achieved by adding different numbers of related or unrelated filler word-pairs to the two blocks. In the lower relatedness proportion block (Block 1 for all participants), 10% of the word-pairs (40/400) were related, and in the higher

relatedness proportion block (Block 2), 50% of the word-pairs (200/400) were related. A core set of 40 controlled and counterbalanced target items was rotated across each of the four conditions, counterbalanced such that no participant saw the same prime or target word twice. Order of related and unrelated trials were randomized individually for each participant, i.e. each participant viewed trials in a different order.

The principles described in the Theory section above led to the development of a Rational Adapter model.

The whole model takes the form:

$$\text{Model output} = -\log_2[\mu * p(\text{word}|\text{prime}) + (1-\mu) * p(\text{word}|\text{average context})]$$

where μ is a point estimate of the probability with which a rational adapter expects a related trial at that point in time.

3.2 Estimating and updating the probability of receiving a related versus an unrelated prime-target pair

To describe a participant's belief about the probability of seeing a related versus an unrelated prime-target pair, we assumed a beta-binomial model over expected trial types. Throughout this paper, we frequently use the parameterization of the beta distribution in terms of a mean μ and a precision ν.[2] At any point in the experiment, the expectation μ of this distribution is used to estimate the probability with which a participant expects to receive a related trial. To set a prior on participants' beliefs, we assumed that participants entered Block 2 believing that the parameters of the experiment would be the same as in Block 1, with a 10% chance of receiving a related trial, hence μ = 0.1.

In addition to setting an initial value for the mean parameter, μ (i.e. participants' beliefs about the probability of seeing a related versus an unrelated prime-target pair at the beginning of Block 2), we also needed to set an initial value for the precision parameter, v (participants' confidence in this belief; as discussed below, this effectively determines *how quickly* participants adapt to the new experimental environment). This precision parameter v can be thought of as the "sample size" of the prior, or the weight given to the prior observations in *pseudocounts*. For example, if ν = 20, then participants give the same weight to 20 trials of new data as to their prior beliefs (see Figure 1 for a depiction of how different values of ν influence the rate of adaptation). As a best-guess approximation, we set ν = 50 at the beginning of Block 2. This value was chosen to be non-trivially different from 0 (assuming that participants did retain some expectations from Block 1) but much less than 400 (the total number of trials observed in Block 1).[3]

[2]This precision parameter is also known as the concentration parameter of the beta distribution. In this paper, we chose to use the term *precision* because this is the term that is most used in the neuroscience literature on predictive processing (see Clark, 2013, for a review). In a Gaussian distribution, *precision* is the reciprocal of variance. Intuitively, it refers to the narrowness of the distribution. In terms of the more common pseudocount parameterization, $\mu = \alpha/(\alpha+\beta)$ and $v=\alpha+\beta$.

[3]To prevent inflation of the type I error rate due to experimenter degrees of freedom, we conducted all our hypothesis tests using this plausible *a priori* value of ν, rather than allowing the selection of the value to be influenced by the process of analysis. After the relevant hypothesis tests had been completed, we then empirically derived the optimal prior for this particular experiment to empirically derive the apparent rate of adaptation in this experimental setting, see Results, Finding the optimal prior certainty.

With μ = 0.1 and ν = 50, the beta prior at the beginning of Block 2 can alternatively be expressed in the pseudocount parameterization as beta(5, 45), i.e. 5 pseudocounts of related trials and 45 of unrelated trials. After each new prime-target pair in Block 2, this beta distribution is updated through Bayes' rule. Thus, μ incrementally changes after each trial. For example, after encountering 5 related and 5 unrelated prime-target pairs in Block 2, a participant's beliefs would be modeled as beta(10, 50). In the present experiment, as evidence accumulates, participants' certainty about the probability of encountering a related trial (μ) increases on average, and the rate of change across trials for μ decreases on average. Given the prior we chose and the statistics of Block 2, the estimated probability of encountering a related trial begins at μ = 0.1, and it asymptotes to μ = 0.5.

3.3 A mixture model to estimate the specific probability of encountering a given target following a given prime for each trial in Block 2

At each point in the experiment, we used a mixture model to estimate the final probability of encountering the specific target word given the prime and the agent's beliefs about the statistical structure of the environment. The mixture has three inputs: a) the expectation of encountering a related target, μ, b) the Forward Association Strength (FAS) from the prime (Nelson et al., 2004), and c) target frequency, estimated from the SUBTLEX corpus (Brysbaert & New, 2009) and converted into a proportion of the corpus total in order to yield the same units as FAS (probability).

Just after encountering each prime, the probability of the target is computed as the weighted sum of its probability as a related target (FAS) and its probability as an unrelated target (frequency), weighted by the expectation of encountering a related target, μ:

$$p(\text{word}) = \mu * \text{FAS} + (1 - \mu) * \text{Frequency}$$

Finally, we log-transform this final estimate of raw word probability ($-\log_2[p(\text{word})]$) to convert it into the information theoretic measure surprisal.

We compute this value individually for all 80 critical targets in Block 2 for each participant, taking into account each participant's idiosyncratic history of related and unrelated trials seen up until that point in the experiment.

4. Methods

4.1 Participants and ERP Data collection

Details about participants and ERP data collection have been previously described in detail by Lau, Holcomb & Kuperberg (2013), and are summarized below.

Participants were all right-handed native speakers of American English recruited from Tufts University. All gave written informed consent to participate. Data were originally collected from 33 participants (19 women; mean age = 20.5 years) and two were omitted due to artifacts. All participants saw the lower relatedness proportion block first (Block 1), followed by the higher relatedness proportion block (Block 2). To ensure that participants processed the words semantically while at the same time not drawing their explicit attention

to semantic relationships between primes and targets, they were instructed to press a button as quickly as possible when they saw a name of an animal. In each block, eighty of the unrelated filler word-pairs included an animal word. Within each block, participants were given short breaks after every 100 trials such that each block was divided into four runs. A similar break was given in between the two blocks. Participants were not explicitly told that there would be any differences between any of the runs.

Stimuli were presented on a computer monitor in yellow 20-point uppercase Arial font on a black background. The prime was visible for 500ms, followed by 100ms of blank screen (total SOA 600ms). The target was then presented for 900ms, followed by 100ms of blank screen. EEG data were collected from twenty-nine tin electrodes, held in place on the scalp by an elastic cap, in a modified 10–20 configuration (Electro-Cap International, Inc., Eaton, OH). The EEG signal was referenced online to the left mastoid, amplified by an Isolated Biolectric Amplifier System Model HandW-32/BA (SA Instrumentation Co., San Diego, CA) with a bandpass of 0.01 to 40 Hz, and digitized at a 200 Hz sampling rate.

4.2 Preprocessing and extraction of individual trial ERP data

The EEG signal was time-locked to target words and segmented. Trials with ocular and muscular artifact were removed as described by Lau, Holcomb & Kuperberg (2013). A 100-ms pre-stimulus baseline was subtracted from all waveforms prior to statistical analysis.

Lau, Holcomb & Kuperberg (2013) reported the results of analyses that averaged the N400 over related and unrelated targets and compared these averages between Block 1 and Block 2. For the purposes of the present study, we extracted the single trial ERP data collected during Block 2.

In each of the 32 participants, we extracted the N400 component evoked by each of the 80 targets per participant in Block 2 — the 40 related and 40 unrelated targets that were counterbalanced across conditions and across the two blocks, as described above.[4] The N400 was operationalized as the averaged voltage between 300–500ms evoked by each target, averaged across three parietal channels (CP1, CP2, and Pz). These were the channels where the block-level N400 effect appeared maximal in the analysis reported by Lau, Holcomb & Kuperberg (2013). Extreme outliers in N400 amplitude were removed (3 standard deviations or more from the mean). Altogether, after the removal of both artifact and extreme outliers, 18.3% of related trials and 17.7% of unrelated trials were removed from analysis.

4.3 Hypothesis Testing

We ran our rational adaptor model for each participant, based on the specific sequence of trials he/she saw in Block 2. This yielded trial-by-trial model outputs for each participant for each individual target item in Block 2. These values were entered as the predictor variable into a linear mixed effects regression model in the R statistical software program version

[4]In principle, our single-trial modeling approach could be used to model every item in the experiment (not just these counterbalanced target items). However, because the remaining items were coded as fillers in the original experiment, we were unable to extract their ERPs for single trial analysis.

3.2.4 (R Core Team, 2016). The trial-by-trial amplitude of the N400 evoked by each target word in each participant in Block 2 was the outcome variable. Additional control predictors were included as necessary for particular hypothesis tests, as described below. The maximal random effects structure across (crossed) subjects and items for the independent variable of interest (rational adaptor model output) was used (Barr, Levy, Scheepers, & Tily, 2013). Regression models were fit using restricted maximum likelihood with the lme4 package version 1.1–11 (Bates, Mächler, Bolker, & Walker, 2015). All continuous predictors were z-transformed, with p-values calculated using the Satterthwaite approximation to degrees of freedom in the lmerTest package version 2.0–30 (Kuznetsova, Brockhoff, & Christensen, 2015).

5. Results

5.1 Visualization of trial-by-trial ERP data and model predictions

In order to visualize the changes in N400 amplitude over target items in Block 2, without assuming any particular parameters of the adaptation, we conducted a loess local regression over N400 amplitudes for related and unrelated words across the ordinal position of critical items in the experiment. The N400 amplitudes evoked by related and unrelated critical targets in Block 2 are shown in Figure 2. As can be seen, the amplitude of the N400 evoked by related and unrelated targets were initially similar, but then diverged as participants were exposed to more and more trials within Block 2 and adapted to its statistical structure. We also noticed that N400 amplitudes for these two conditions converged again at the very end of Block 2 (see Discussion).

For comparison, we also visualize the trial-by-trial output of our rational adaptor model, computed for each individual participant based on the specific sequence of trials they saw in Block 2, see Figure 3. We first observe that, across participants, the model's estimates of target probability are more consistent for related than for unrelated trials. This is because the model's estimates of the probability of encountering targets in related trials are largely driven by FAS, which is relatively consistent across all related trials (between 0.5–1 for all related targets). In contrast, its estimates of the probability of encountering targets in unrelated trials are largely driven by frequency, which can vary across many orders of magnitude. Thus, each participant's idiosyncratic ordering of critical trials causes larger fluctuations in model outputs for unrelated than for related trials. Comparing Figure 3 to Figure 2, we see that, although our computational model predicts a large disparity between related and unrelated trials from the start (unlike the initial similarity seen in the N400 data), it 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. Our model does not predict the convergence at the end of Block 2 (see Discussion).

5.2 Trial-by-trial model output explains trial by trial variance in N400 amplitudes

We first asked whether our model's trial-by-trial output explains trial-by-trial variance in N400 amplitudes within Block 2. Recall that our model makes individualized predictions for each trial in each participant, based on the word-level characteristics of the trial and the participant's idiosyncratic history of related and unrelated trials seen up until that point in

the experiment. We can thus test our model predictions against single-trial N400 amplitudes in each participant.

We first z-transformed the trial-by-trial output values of our model and used these values as the predictor in a linear mixed effects regression analysis with the N400 amplitude on each trial in Block 2 as the outcome variable. We included the maximal random effects structure: by-subjects intercepts and random slopes of model output, and by-target-word intercepts and random slopes of model output. As expected, the model's trial-by-trial output values significantly accounted for variance in trial-by-trial N400 amplitudes ($\beta = -1.14$, $t = -5.24$, $p < 0.001$): the higher the estimated probability of the target words, the larger (more negative) the N400 amplitude, with a 1 bit increase in model-predicted surprisal (-log probability) corresponding with a ~0.18μv increase in N400 amplitude, 95% CI [0.11, 0.24 μv/bit].

By design, the unrelated targets had lower probabilities than the related targets. As such, it is plausible that effect described above simply reflected the well-established categorical semantic priming effect (in this study, the main effect of Relatedness that was already reported by Lau, et al., 2013). To determine whether the inclusion of our model's trial-by-trial outputs accounted for variance in N400 amplitude *over and above* this categorical effect, we ran a second linear mixed effects regression model that included not only the model's output on each trial, but also a categorical Relatedness control predictor. We again included by-subject and by-target-word intercepts and random slopes of model output. This showed that our model's output on individual trials accounted for variance in N400 amplitude ($\beta = -2.21$, $t = -2.76$, $p = 0.006$), over and above the categorical effect of Relatedness. We caution that, given the multicollinearity between the model's trial-by-trial output and the Relatedness effect (the primary motivation for running this test in the first place), this β estimate is likely inflated. We therefore limit our conclusions here to establishing the significance of the effect, rather than its marginal magnitude.

5.3 The Rational Adapter Model outperforms its constituent elements alone

The analysis described above indicates that our computational model was able to explain trial-by-trial variance in N400 amplitudes. However, it may be that its additional explanatory power over and above the categorical effect of Relatedness, was simply due to the inclusion of items-level information (FAS and frequency, both known to predict N400 amplitude) within the model, rather than the specific way in which such information combined together with rational adaptation on each trial to yield final trial-by-trial outputs. To address this possibility, we ran an additional regression model that tested the effect of the rational adaptor model's output, but this time controlling not only for categorical Relatedness (as in the previous regression model), but also for trial-by-trial frequency and FAS — the item-specific constituent elements that went into our rational adaptor model. Specifically, log-transformed frequency and log-transformed FAS were included as control predictors. (Because the log transformation requires a non-zero probability, all unrelated targets were given a probability of 0.005 — i.e. half a percent—prior to log transforming.[5]) Again, by-subjects and by-target-word intercepts and random slopes of model output were included, and all continuous predictors were standardized.

We found that the model's trial-by-trial output significantly accounted for variance in N400 amplitudes ($\beta = -2.30$, $t = -2.11$, $p = 0.036$), above and beyond frequency, FAS, and Relatedness. This indicates that the increased fit described above was not simply due to the fact we included additional information about items-level features. Rather, it tells us that the particular way in which these items-level features combined within the model, including weighting by the Rational Adaptation component, was an important source of explanatory power.

5.4 Finding the optimal prior precision

As discussed in the Calculation section, in addition to setting a prior over participants' beliefs about the expectation of seeing a related versus an unrelated prime-target pair (μ), we also needed to set a prior over participants' *confidence* in this belief. This was given by the "sample size" or pseudocount of the prior — its precision, ν. This number specifies the number of trials within Block 2 that participants would need to have encountered before beginning to give more weight to the statistics of Block 2 trials (50% related pairs) than the statistics of Block 1 trials (10% related pairs, as reflected in their prior beliefs). It therefore determines the rate of adaptation over trials (see Figure 1, which shows the adaptation of μ for different prior pseudocounts ν). Our model assumes that participants had some prior expectation that the environment was non-stationary—that is, that the statistical structure of Block 2 might differ from Block 1, and we set this prior precision at $\nu = 50$. This 50 pseudocount prior, however, was merely a ballpark figure, chosen to be non-trivially different from 0 (assuming that participants did retain some expectations from the previous block), but much less than 400 (the total number of trials in Block 1). We therefore next sought to ensure that our results were not idiosyncratically dependent on having made a lucky guess.

It is an empirical question what prior ν best accounts for the variance in N400 amplitude over the course of Block 2. To estimate this value, we re-calculated our rational adaptor model for every possible integer-valued ν from 1 to 800 pseudocounts (i.e. we re-calculated the model's target probability estimates for all trials for all participants, as above.) This 1–800 range encompassed the range from holding almost no beliefs from Block 1 to near-complete resistance to new statistical information from Block 2. We then ran 800 separate regression models with the outputs of each computational model and categorical Relatedness as predictors, the amplitude of the N400 as the dependent variable, and a maximal random effects structure (as described above). After fitting each of these 800 regression models, we extracted the log-likelihood of each in order to identify the ν that maximized model fit. As our aim was only to describe the present dataset, and we don't believe that the exact speed of adaptation here should necessarily generalize to other experiments (meaning that the optimal ν should be interpreted as a descriptive statistic characterizing the present data set only), we

[5]Log-transformed frequency and FAS predictors (rather than raw probabilities) were used for the most direct comparison with our model output, which also log-transforms its final predicted probability. Moreover, frequency is standardly log-transformed for use as a predictor of processing difficulty (see references in Theory). However, as FAS is not standardly log-transformed, we also ran a version of this model using raw FAS, as well as log frequency and (categorical) Relatedness, as predictors. Results were qualitatively similar, and all patterns of statistical significance were identical.

preferred this approach of maximizing model fit over all of the data to an approach that maximized cross-validated prediction error over smaller subsets of the data.

These data are shown in Figure 4. This shows that there was a single maximum log-likelihood with a beta(69.3, 7.7) prior, or $\nu = 77$ pseudocounts. However, all pseudocounts between 70 and 85 yielded very similar model fits, and performance degrades smoothly on either side. This indicates that, on average, participants in this study began giving more weight to the data in Block 2 than in Block 1 after around 77 trials into Block 2. In contrast, a much lower or much higher precision ν did not account for the N400 data well because it would lead to adaptation that was too fast or too slow respectively. We note that some models had poor fits because they did not converge. (They appear to follow a second curve, suggesting that they failed to converge in a similar way.) None of the models that failed to converge were within the 70–85 range capturing the maximum log-likelihood.

5.5 Log-transformed word probability (Surprisal) significantly accounts for variance in N400 amplitudes above and beyond raw estimates of word probability

In our rational adaptor model, we log-transformed our final estimate of probability to convert these estimates into surprisal values for each target. As discussed in the Theory section, this decision was based on previous empirical evidence from behavioral studies that surprisal may be a preferable measure of word processing difficulty than raw probability, particularly for very low estimates of probability (Hale, 2001; Levy, 2008; Smith and Levy, 2013). We then explicitly tested the hypothesis that log-transformed probability (surprisal) is a better predictor than raw probability of the N400 amplitude in the present semantic priming dataset. Importantly, we addressed this question only after completing all the other models described above (rather than trying many different assumptions and selecting the ones that yielded the most publishable p-values).

To test this hypothesis, we directly compared surprisal and raw probability (as computed by our model in both cases) as predictors of N400 amplitudes. It would not be particularly fair to directly pit raw probability against log-transformed probability (surprisal) using a computational model with a prior precision that maximized the word surprisal effect (computed as $\nu = 77$ pseudocounts, as described above), and it would also not be particularly coherent to continue using the arbitrarily-guessed prior precision of $\nu = 50$ pseudocounts. We therefore decided to derive the precision that optimized the word probability effect, so that we could test the marginal contribution of word surprisal given this prior degree of certainty. This can be viewed as the most conservative prior with which to carry out this test.

We determined the optimal prior certainty for (standardized) word probability by fitting 800 linear regression models with word probability and Relatedness as fixed predictors, and we then identified the value of ν that maximized log-likelihood, as described above. This approach yielded a precision of $\nu = 58$ pseudocounts. We reran our rational adaptor model using $\nu = 58$ and extracted both raw probability and log-transformed trial-by-trial outputs. We then carried out another regression model to test the hypothesis that these log-transformed probability (surprisal) values could significantly account for variance in N400 amplitude, above and beyond what could already be accounted for with raw word probability

estimates. Again, we included the control categorical Relatedness variable in this regression model, and we again used the maximal random effects structure.

We found that surprisal on each trial did indeed significantly account for variance in N400 amplitudes ($\beta = -2.09$, $t = -2.57$, $p = 0.011$) above and beyond estimates of raw word probability on each trial (in addition to the categorical Relatedness control variable). In contrast, estimates of raw word probability on each trial did not account for any significant variance in N400 amplitude that was not already accounted for by surprisal and the control variable, Relatedness ($\beta = 0.38$, $t = 0.67$, $p = 0.48$), despite the fact that its optimal prior had been assumed.

Discussion

It is well established that the magnitude of the behavioral semantic priming effect is sensitive to the predictive validity of the broader experimental environment (Neely, 1991; Tweedy et al., 1977). This effect of predictive validity also influences the modulation of the N400 ERP component — a direct neural index of semantic processing (Kutas & Federmeier, 2011): when the proportion of related word-pairs within an experimental block increases, the N400 priming effect increases (Brown et al., 2000; Holcomb, 1988; Lau et al., 2013). In this study, we show that a quantitative Bayesian model was able to predict how the amplitude of the N400 evoked by individual target words changed as participants adapted, trial by trial, to a new, higher predictive validity environment (Block 2, in which 50% of trials were related word-pairs), following a lower predictive validity environment (Block 1, in which only 10% of trials were related word-pairs).

Several previous studies of semantic priming have shown that the amplitude of the N400 evoked by a target word is influenced by the FAS of its prime (Luka & Van Petten, 2014; van Vliet et al., 2016) and by the proportion of related word-pairs in its broader experimental environment (Brown et al., 2000; Holcomb, 1988; Lau et al., 2013). We also know that when words are presented in isolation of any context, the amplitude of the N400 is influenced by their frequency (Laszlo & Federmeier, 2014). What is novel about our computational model is that it explicitly specifies how these factors quantitatively combine to compute the final probability of encountering a given target word. We found that this final estimate of probability (log-transformed) accounted for variance in the amplitude of the N400 evoked by targets beyond the static (average) categorial effect of Relatedness (within Block 2), and beyond the independent effects of items-level information like frequency and FAS. In other words, it was the particular way our model combined these two types of information and updated their weights, trial-by-trial, that accounted for additional variance.

Our model incorporates several core principles of probabilistic prediction, rational adaptation and their relationship. First, it assumes that participants weight the degree to which they use the predictability of a local context (here, the FAS of the prime) by their belief about whether that local context will be *informative* of the upcoming input. In our model, the informativeness (or predictive validity) of the prime was operationalized as the agent's belief that she would encounter a related (as opposed to an unrelated) target, and it was modeled by the mean parameter of the beta distribution.

Second, our model assumes that participants' belief about the informativeness/predictive validity of a local context can be updated, based on new inputs, according to Bayes' rule. A core principle of Bayesian inference is that the degree to which the agent updates her prior belief depends on her certainty in that belief, with greater uncertainty leading to more updating based on the current input (as opposed to the prior history). In our model, participants' uncertainty about the predictive validity of the prime was represented by the precision parameter of the beta distribution, which describes participants' (expected) uncertainty about the statistical structure of the current environment (Yu & Dayan, 2005).[6] Intuitively, the lower the precision, the wider the beta distribution and the less confident the agent is about the prime's predictive validity. Importantly, we initially set the precision parameter to a value that was lower than the number of trials that had actually been observed over Block 1 ($\nu = 50$ as opposed to $\nu = 400$). Therefore, a key assumption of our model was that, at the beginning of Block 2, participants had some uncertainty about the statistical structure of the environment. This uncertainty is what allowed them to successfully adapt to Block 2.

Third, our model assumes that, as participants accumulate more data, they become increasingly confident about statistical structure of the environment (and hence the predictive validity of the prime), and so the rate of adaptation decreases. This once again illustrates a core principle of Bayesian inference: uncertainty decreases as more data are observed (so long as the statistical structure of the environment is assumed to be stable). In our model, this increase in confidence was reflected by the trial-by-trial increase in the value of the precision parameter of the beta distribution (the number of pseudocounts). Because this precision parameter set the degree to which participants weighted their prior history versus the current input during belief updating, its increase on successive trials meant that, on average, each successive trial carried less weight. This is illustrated in Figure 2: the pattern of N400 amplitude across trials suggests that adaptation proceeded more rapidly at the beginning of Block 2 and then slowed as the block continued (see Figure 3, which illustrates our model's estimate of the adaptation effect in each participant; see also Figure 1).

6.1 Implications of our findings for understanding the roles of probabilistic prediction and adaptation during language comprehension

Probabilistic prediction and adaptation are closely linked and highly relevant to communicating in the real world. Probabilistic prediction leads to more efficient language processing, but only if such predictions are based on the probabilistic statistical structure of the communicative environment (Kuperberg & Jaeger, 2016, section 1). Because our real-world environment is *non-stationary* — that is, the statistical structures of our linguistic (and non-linguistic) inputs are constantly changing in systematic ways, depending on who we are talking to or what we are reading, we must *adapt* to different environments so that probabilistic prediction remains efficient. Indeed, there is plenty of evidence that we are able

[6]This type of expected uncertainty about the statistical structure of the environment is known as *estimation uncertainty* (Payzan-LeNestour & Bossaerts, 2011). Another type of expected uncertainty is *outcome uncertainty* or *risk*, which describes uncertainty resulting from the inherently stochastic nature of an outcome. In the current model, outcome uncertainty would be influenced by the mean parameter of the beta distribution and was maximal at 0.5.

to adapt to changing environments by adjusting our predictions at multiple levels of linguistic representation, including phonetic (Kraljic & Samuel, 2006; Norris, McQueen, & Cutler, 2003; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004), lexical (Creel, Aslin, & Tanenhaus, 2008), syntactic (Chang, Dell, & Bock, 2006; Hanulikova, van Alphen, van Goch, & Weber, 2012; Kamide, 2012), and pragmatic (Grodner & Sedivy, 2011; Nieuwland, Ditman, & Kuperberg, 2010). The present model highlights the close computational links between probabilistic prediction and adaptation (see also Chang et al., 2006; Dell & Chang, 2014). It is also consistent with previous work suggesting that adaptation is, at least in part, based on rational Bayesian principles (e.g. Jaeger & Snider, 2013; Kleinschmidt & Jaeger, 2015; Myslin & Levy, 2016). Our results extend this previous work to show that these principles of rational adaptation are evident in the brain, influencing the N400, which indexes the earliest stages of accessing meaning from incoming words.

6.1.1. Adapting probabilistic prediction to different statistical environments during language comprehension—

Although in the present study, we modeled the effects of adaptation on prediction in a simple semantic priming paradigm, we suggest that our findings are relevant for understanding the relationships between probabilistic prediction and adaptation during higher-level language comprehension. Recent evidence suggests that, just as the ratio of related to unrelated word-pair trials within a block influences the magnitude of the behavioral and N400 semantic priming effect, the ratio of predictable to unpredictable sentences in an experimental environment influences predictability effects during sentence comprehension. For example, the effect of lexical probability on reading times is increased when there is a higher proportion of predictable sentences in the environment (although this was a between-group effect; Brothers, Swaab & Traxler, 2017, Experiment 2). And a recent ERP study suggests that the proportion of predictable spoken sentences in the environment can also influence the magnitude of the N400 expectancy effect within participants, although, unlike in the present study, the change between blocks was accompanied by a more overt signal — a change in speaker identity (Brothers, Hoversten, Dave, Traxler & Swaab, under review), which, as discussed below, may have provided a cue that the environment had changed.

It will be therefore be important to determine whether the principles incorporated in our computational model shed light on these sentence-level adaptation effects. Of course, in extending this model, we emphasize that there are critical differences between prediction during semantic priming and sentence comprehension. First, in a semantic priming paradigm, the “context” is a single word (the prime), and, to generate predictions about the target, participants are likely to draw upon their knowledge about simple semantic associations between words, estimated in our model by the prime’s Forward Association Strength. In contrast, during sentence comprehension, the context constitutes the full set of words (and non-verbal information) that has been encountered prior to a given incoming word. Comprehenders are therefore likely to draw upon multiple different types of linguistic and non-linguistic knowledge to generate estimates of the probability of the upcoming words, typically estimated using the cloze procedure (Taylor, 1953).[7] Moreover, different words, or combinations of words, within a context may be associated with different predictive validities (or reliabilities), which may be weighted, possibly in a Bayes optimal

fashion (cf Knill & Saunders, 2003; Ernst & Banks, 2002), in generating predictions (see Kuperberg, 2016, page 610 for discussion).

Another factor to consider when extending this type of model to sentence comprehension is whether it is appropriate to log-transform the final estimate of probability to compute the information theoretic measure, surprisal. In the present model, we included this final log-transform step because previous work had suggested that surprisal can be a better predictor of processing difficulty than raw probabilities, particularly for very low probability words (Hale, 2001; Levy, 2008; Smith & Levy 2013), and we estimated the unrelated targets in the present study to have very low probabilities. Indeed, we subsequently verified that the amplitude of the N400 was better predicted by the log-transformation of the model's estimated probability of each target word (its surprisal) than by its raw probability. However, it is an open question whether surprisal is a better predictor of N400 amplitude than raw probability during sentence and discourse comprehension. It can be challenging to estimate the true probabilities of low probability items using ngram corpus-based methods (Ong & Kliegl 2011) and cloze procedures, and so it will be important for future studies to investigate the link between neural (and behavioral) responses and probability in more detail, combining large-scale cloze studies with state-of-the-art language models.

Finally, it is also important to bear in mind that, during higher-level language comprehension, there may be metabolic costs incurred in *generating* predictions based on higher-level context. Such costs are less likely to be incurred during a simple semantic priming paradigm, and, indeed, our mixture model assumed no such costs, weighting the use of FAS and frequency purely by participants' estimates of the prime's predictive validity (the mean of the beta distribution). Thus, in the present model, participants' confidence about prime's predictive validity (represented by the precision of the beta distribution) influenced the rate of adaptation across trials, but it had no direct effect on the amplitude of the N400 evoked on any given trial. During higher-level language comprehension, however, comprehenders may rationally allocate their limited resources in generating predictions based on their confidence about the predictive validity of the local context. For example, in situations of high uncertainty about the informativeness of a given local context, comprehends may limit the influence of top-down contextual prediction, relying more on bottom-up stimulus features. This would be keeping with frameworks that highlight a role of precision in hierarchical message passing during predictive coding (e.g. Feldman & Friston, 2010; see Clark, 2013, for a review).

While there are obviously many open questions, this type of model illustrates some of the core computational principles that are important to consider in models of prediction and adaptation during higher-level language comprehension. More practically, it also highlights the need to carefully describe not only any experimental manipulation of interest, but also details about surrounding stimuli including fillers when sharing research findings. For example, based on cloze norms, one might estimate the probability of a particular word in a

[7]We emphasize that FAS between individual pairs of words is *not* thought to play a major role in sentence or discourse comprehension (for behavioral evidence, see Foss & Ross, 1983; Morris, 1994, Experiment 2; Traxler & Foss, 2000; for ERP evidence, see Camblin, Gordon, & Swaab, 2007; Coulson, Federmeier, Van Petten, & Kutas, 2005; Van Petten, 1993).

highly constraining sentence context to be 0.9. However, if only 50% of sentences in an experiment end with an expected word, then, by the end of the experiment, participants may estimate the actual probability of encountering a predictable word to be significantly lower (in our model, expectations would asymptote to a probability of 0.45).

6.1.2. Inferring when to adapt during language comprehension—Another major set of outstanding questions is how the brain determines when and how quickly to adapt in any given situation. Our model was highly simplified in that, by initially setting the precision parameter to a number that was much lower than the number of trials actually observed during Block 1, we assumed that participants believed that Block 2 would be different from Block 1, leading them to down-weight the importance of information that they had gained over the course of Block 1, and adapt to Block 2. During real world language processing, however, the brain must *infer* when and how quickly to adapt to different communicative environments (see Qian, Jaeger & Aslin, 2012, and Kleinschmidt & Jaeger, 2015 for discussion). This can be challenging because the agent must be able to distinguish between inputs that are unpredicted as a result of a true systematic change in the environment (so-called unexpected surprise) from inputs that are unpredicted because of the inherent stochasticity of the current environment and uncertainty about its statistical structure (expected surprise; see Yu & Dayan, 2005 and Qian, Jaeger & Aslin, 2012 for discussion). Correctly inferring how quickly to adapt to a systematically changing environments is crucial for efficient language processing: if the brain adapts too slowly or too quickly, then its probabilistic predictions will, on average, be inaccurate.

Future work may be able to capture something about how the agent infers how quickly to adapt by incorporating hyperparameters into models of adaptation that specify beliefs about environmental non-stationarity. These hyperparameters might specify expectations about the rate of continuous environmental change (volatility, e.g. Behrens, Woolrch, Walton, Rushworth, 2007) or the frequency of discrete change points (e.g. Gallistel, Mark, King & Latham, 2001). Thus, in addition to learning the current environmental statistics, this type of hierarchical model would also be learning how likely the environment is to change, with different levels of the model influencing one another. The ability to incorporate these hyperparameters is a strength of the Bayesian modeling approach we take here.

Another major challenge for the brain is that adaptation can potentially interfere inappropriately with long-term knowledge, particularly when changes in the local environment are only temporary (e.g. when listening to an atypical speaker). This need to adapt locally without losing the benefit of one's previous or longer-term knowledge is known as the stability-plasticity dilemma. One proposed solution to this dilemma is that comprehenders are able to keep track of multiple sets of beliefs about environmental statistics (i.e. multiple models; see Kleinschmidt & Jaeger, 2015; Qian, Jaeger & Aslin, 2012, 2016). Thus, unexpected surprise may not simply lead participants to adapt their current model to a new environment; it may instead lead them to *switch* to a different pre-stored model, or switch to learn a new model entirely (see Qian, Jaeger & Aslin, 2012, 2016; Gallistel, Krishan, Liu, Miller & Latham, 2014). In real-world communicative situations, evidence that comprehenders should switch models can also come from external cues, such as a new face or a new voice indicating that one is now communicating with a different

speaker. Again, this type of model switching can be implemented naturally within a Bayesian modeling framework. The rational adaptor model described here forms the basis for adaptation to a single environment, which is a necessary building block towards being able to store and retrieve multiple environmental models.

Relatedly, at a neural level, it will be important to determine whether ERP components other than the N400 more specifically track inferences about environmental change. For example, we have previously hypothesized that a family of late positivities (anteriorly distributed when evoked by highly informative incoming words, and posteriorly distributed —the P600 — when evoked by words that are semantically or syntactically anomalous) may reflect the detection of *unexpected* surprise, triggering either rapid adaptation of the current model, or model switching (Kuperberg, 2013; Kuperberg & Jaeger, 2016, section 4).

Finally, we emphasize that, in addition to rational principles of Bayesian updating, both prediction and adaptation are likely to depend on many other factors that influence utility, including task demands. Indeed, as shown in Figure 2, towards the end of the block, the magnitude of the N400 effect appeared to become smaller. This may simply reflect a general fatigue effect or anticipating the end of the experiment, leading participants to invest less in the task. The best models of adaptation would need to be robust to these additional processes or else account for them.

7. Conclusion

In conclusion, our quantitative model of trial-by-trial adaptation on the N400 ERP component provides evidence that (1) the brain combines immediate contextual constraints with global probabilistic constraints to influence semantic processing of incoming words, (2) the brain has some prior expectation that the broad statistical structure of its environment might change and is able to rationally adapt its probabilistic semantic predictions of incoming words in response to this new environment.

Of course, there remains much work to be done to determine exactly how these principles of probabilistic prediction and adaptation are instantiated at the algorithmic and implementation/neural levels. There is evidence for close links between probabilistic principles and some connectionist models of language processing (McClelland, Mirman, Bolger & Khaitan, 2014; Rabovsky, et al., 2018) and language adaptation (Chang et al., 2006; see Jaeger & Snider, 2013, for discussion). There is also evidence that the population activity of neurons can represent uncertainty that underlies probabilistic computation (Fiser, Berkes, Orban & Lengyel, 2010; Orban, Berkes, Fiser & Lengyel, 2016), although we know little about how this plays out during language comprehension. This study provides a mathematical description of the links between probabilistic prediction and adaptation. By showing that these principles influence modulation on the N400 — a direct neural measure of semantic processing — our findings pave the way towards bridging the experimental ERP, computational modeling and neuroscience literatures, thereby providing new insights into how our brains infer meaning from language.

Acknowledgements

The authors thank Eric Fields and Sorabh Kothari for their assistance with data collection, and Trevor Brothers for his helpful comments on the manuscript. This work was funded by the National Institute of Mental Health (R01MH071635 to GRK), National Institute of Child Health and Human Development (F32HD063221 to EFL and R01HD082527 to GRK), National Science Foundation (SPRF-FR 1715072 to EM), and the Sidney R. Baer Jr. Foundation (to GRK).

References

Anderson JR (1990). The Adaptive Character of Thought. Hillsdale, NJ: Erlbaum.

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. doi: 10.1016/j.jml.2012.11.001

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

Behrens TEJ, Woolrich MW, Walton ME, & Rushworth MFS (2007). Learning the value of information in an uncertain world. Nature Neuroscience, 10(9), 1214–1221. doi:10.1038/Nn1954 [PubMed: 17676057]

Bentin S, McCarthy G, & Wood CC (1985). Event-related potentials, lexical decision and semantic priming. Electroencephalography and Clinical Neurophysiology, 60, 343–355. [PubMed: 2579801]

Brothers T, Hoversten L, Dave S, Traxler MJ, & Swaab T (under review). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes.

Brothers T, Swaab TY, & Traxler MJ (2017). Goals and strategies influence lexical prediction during sentence comprehension. Journal of Memory and Language, 93, 203–216. doi:10.1016/j.jml. 2016.10.002

Brown CM, Hagoort P, & Chwilla DJ (2000). An event-related brain potential analysis of visual word priming effects. Brain and Language, 72(2), 158–190. doi:10.1006/brln.1999.2284 [PubMed: 10722786]

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. Behavioral Research Methods, 41, 977–990.

Camblin CC, Gordon PC, & Swaab TY (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. Journal of Memory and Language, 56(1), 103–128. [PubMed: 17218992]

Chang F, Dell GS, & Bock JK (2006). Becoming syntactic. Psychological Review, 113(2), 234–272. doi:10.1037/0033-295x.113.2.234 [PubMed: 16637761]

Clark A (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. doi:10.1017/S0140525X12000477 [PubMed: 23663408]

Coulson S, Federmeier KD, Van Petten C, & Kutas M (2005). Right hemisphere sensitivity to word- and sentence-level context: evidence from event-related brain potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1), 129–147. doi:10.1037/0278-7393.31.1.129

Creel SC, Aslin RN, & Tanenhaus MK (2008). Heeding the voice of experience: the role of talker variation in lexical access. Cognition, 106(2), 633–664. doi:10.1016/j.cognition.2007.03.013 [PubMed: 17507006]

Dell GS, & Chang F (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1634), 20120394. doi:10.1098/rstb.2012.0394

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. doi: 10.1038/nn1504 [PubMed: 16007080]

Ehrlich SF, & Rayner K (1981). Contextual effects on word perception and eye movements during reading. Journal of Verbal Learning and Verbal Behavior, 20(6), 641–655. doi:10.1016/s0022-5371(81)90220-6

Ernst MO, & Banks MS (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433. doi:10.1038/415429a [PubMed: 11807554]

Federmeier KD (2007). Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology, 44(4), 491–505. doi:10.1111/j.1469-8986.2007.00531.x [PubMed: 17521377]

Feldman H, & Friston KJ (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4, 215. doi:10.3389/fnhum.2010.00215 [PubMed: 21160551]

Fine AB, Qian T, Jaeger TF, & Jacobs RA (2010). Is there syntactic adaptation in language comprehension? Paper presented at the Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics (CMCL '10), Uppsala, Sweden.

Fischler IS, & Bloom PA (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. Journal of Verbal Learning and Verbal Behavior, 5, 1–20. doi:10.1016/S0022-5371(79)90534-6

Fiser J, Berkes P, Orbán G, & Lengyel M (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in Cognitive Sciences, 14(3), 119–130. [PubMed: 20153683]

Foss DJ, & Ross JR (1983). Great expectations: Context effects during sentence processing In Flores d'Arcais GB & Jarvella RJ (Eds.), The Process of Language Understanding (pp. 169–191). Chichester: John Wiley & Sons.

Frank SL, Otten LJ, Galli G, & Vigliocco G (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and Language, 140, 1–11. doi:10.1016/j.bandl.2014.10.006 [PubMed: 25461915]

Frank SL, & Willems RM (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Language, Cognition and Neuroscience, 1–12. doi:10.1080/23273798.2017.1323109

Gallistel CR, Krishan M, Liu Y, Miller R, & Latham PE (2014). The perception of probability. Psychological Review, 121(1), 96. [PubMed: 24490790]

Gallistel CR, Mark TA, King AP, & Latham PE (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. Journal of Experimental Psychology: Animal Behavior Processes, 27(4), 354. [PubMed: 11676086]

Griffiths TL, Chater N, Kemp C, Perfors A, & Tenenbaum JB (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14(8), 357–364. doi:10.1016/j.tics.2010.05.004 [PubMed: 20576465]

Grodner D, & Sedivy J (2011). The effect of speaker-specific information on pragmatic inferences In Gibson E & Pearlmutter NJ (Eds.), The Processing and Acquisition of Reference (Vol. 2327, pp. 239–272). Cambridge, MA: MIT Press.

Hale J (2001). A probabilistic Earley parser as a psycholinguistic model Paper presented at the Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01), Pittsburgh, PA.

Hanulikova A, van Alphen PM, van Goch MM, & Weber A (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. J Cogn Neurosci, 24(4), 878–887. doi:10.1162/jocn_a_00103 [PubMed: 21812565]

Holcomb PJ (1988). Automatic and attentional processing: an event-related brain potential analysis of semantic priming. Brain and Language, 35(1), 66–85. [PubMed: 3179703]

Hutchison KA (2007). Attentional control and the relatedness proportion effect in semantic priming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(4), 645–662. doi:10.1037/0278-7393.33.4.645

Jaeger TF, & Snider NE (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. Cognition, 127(1), 57–83. doi:10.1016/j.cognition.2012.10.013 [PubMed: 23354056]

Kamide Y (2012). Learning individual talkers' structural preferences. Cognition, 124(1), 66–71. doi:10.1016/j.cognition.2012.03.001 [PubMed: 22498776]

Keefe DE, & Neely JH (1990). Semantic priming in the pronunciation task: The role of prospective prime-generated expectancies. Memory & Cognition, 18(3), 289–298. doi:10.3758/bf03213882 [PubMed: 2355858]

Kleinschmidt DF, & Jaeger FT (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychol Rev, 122(2), 148–203. doi:10.1037/a0038695 [PubMed: 25844873]

Knill DC, & Saunders JA (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Research, 43(24), 2539–2558. doi:10.1016/s0042-6989(03)00458-9 [PubMed: 13129541]

Kraljic T, & Samuel AG (2006). Generalization in perceptual learning for speech. Psychonomic Bulletin & Review, 13(2), 262–268. doi:10.3758/bf03193841

Kuperberg GR (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension In Miller B, Cutting L, & McCardle P (Eds.), Unraveling Reading Comprehension: Behavioral, Neurobiological, and Genetic Components (pp. 176–192). Baltimore, MD: Paul Brookes Publishing.

Kuperberg GR (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. Language, Cognition and Neuroscience, 31(5), 602–616.

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

Kutas M, & Federmeier KD (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. doi:10.1146/annurev.psych.093008.131123

Kutas M, & Hillyard SA (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307(5947), 161–163. doi:10.1038/307161a0 [PubMed: 6690995]

Kuznetsova A, Brockhoff PB, & Christensen RHB (2015). Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0–33. http://cran.r-project.org/package=lmerTest.

Laszlo S, & Federmeier KD (2014). Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. Language, Cognition and Neuroscience, 29(5), 642–661.

Lau EF, Holcomb PJ, & Kuperberg GR (2013). Dissociating N400 effects of prediction from association in single-word contexts. J Cogn Neurosci, 25(3), 484–502. doi:10.1162/jocn_a_00328 [PubMed: 23163410]

Lau EF, Weber K, Gramfort A, Hamalainen MS, & Kuperberg GR (2016). Spatiotemporal signatures of lexico-semantic prediction. Cerebral Cortex, 26(4), 1377–1387. doi:10.1093/cercor/bhu219 [PubMed: 25316341]

Levy R (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 1126–1177. doi:10.1016/j.cognition.2007.05.006 [PubMed: 17662975]

Luka BJ, & Van Petten C (2014). Gradients versus dichotomies: how strength of semantic context influences event-related potentials and lexical decision times. Cognitive, Affective, & Behavioral Neuroscience, 14(3), 1086–1103. doi:10.3758/s13415-013-0223-1

Luke SG, & Christianson K (2016). Limits on lexical prediction during reading. Cognitive Psychology, 88, 22–60. doi:10.1016/j.cogpsych.2016.06.002 [PubMed: 27376659]

McClelland JL, Mirman D, Bolger DJ, & Khaitan P (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. Cognitive Science, 38(6), 1139–1189. doi:10.1111/cogs.12146 [PubMed: 25098813]

Morris RK (1994). Lexical and message-level sentence context effects on fixation times in reading. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 92–103.

Myslin M, & Levy R (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. Cognition, 147, 29–56. doi:10.1016/j.cognition.2015.10.021 [PubMed: 26605963]

Neely JH (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories In Besner D & Humphreys GW (Eds.), Basic Processes in Reading and Visual Word Recognition (pp. 264–333). Hillsdale, NJ: Erlbaum.

Neely JH, & Keefe DE (1989). Semantic context effects on visual word processing: A hybrid prospective-retrospective processing theory In Bower GH (Ed.), Psychology of Learning and Motivation: Advances in research and theory (Vol. 24, pp. 207–248). New York: Academic Press.

Neely JH, Keefe DE, & Ross K (1989). Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(6), 1003–1019. doi: 10.1037/0278-7393.15.6.1003

Nelson DL, McEvoy CL, & Dennis S (2000). What is free association and what does it measure? Memory & Cognition, 28(6), 887–899. doi:10.3758/bf03209337 [PubMed: 11105515]

Nelson DL, McEvoy CL, & Schreiber TA (2004). The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers, 36(3), 402–407.

Nieuwland MS, Ditman T, & Kuperberg GR (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. Journal of Memory and Language, 63(3), 324–346. doi:10.1016/j.jml.2010.06.005 [PubMed: 20936088]

Nieuwland MS, Politzer-Ahles S, Heyselaar E, Segaert K, Darley E, Kazanina N, … Huettig F (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. Elife, 7. doi:10.7554/eLife.33468

Norris D (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. Psychological Review, 113(2), 327–357. doi:10.1037/0033-295X.113.2.327 [PubMed: 16637764]

Norris D, McQueen JM, & Cutler A (2003). Perceptual learning in speech. Cognitive Psychology, 47(2), 204–238. doi:10.1016/S0010-0285(03)00006-9 [PubMed: 12948518]

Ong J, & Kliegl R (2008). Conditional co-occurrence probability acts like frequency in predicting fixation durations. Journal of Eye Movement Research, 2(1), 1–7. doi:10.16910/jemr.2.1.3

Orbán G, Berkes P, Fiser J, & Lengyel M (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. Neuron, 92(2), 530–543. [PubMed: 27764674]

Payzan-LeNestour E, & Bossaerts P (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Computational Biology, 7(1), e1001048. doi:10.1371/journal.pcbi.1001048 [PubMed: 21283774]

Pearce JM, & Hall G (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87(6), 532–552. doi: 10.1037//0033-295x.87.6.532 [PubMed: 7443916]

Perfors A, Tenenbaum JB, Griffiths TL, & Xu F (2011). A tutorial introduction to Bayesian models of cognitive development. Cognition, 120(3), 302–321. doi:10.1016/j.cognition.2010.11.015 [PubMed: 21269608]

Qian T, Jaeger TF, & Aslin RN (2012). Learning to represent a multi-context environment: more than detecting changes. Front Psychol, 3, 228. doi:10.3389/fpsyg.2012.00228 [PubMed: 22833727]

Qian T, Jaeger TF, & Aslin RN (2016). Incremental implicit learning of bundles of statistical patterns. Cognition, 157, 156–173. [PubMed: 27639552]

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: ISBN 3-900051-07-0.

Rabovsky M, Hansen SS, & McClelland JL (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nature Human Behaviour, 2(9), 693–705. doi:10.1038/s41562-018-0406-4

Rescorla RA, & Wagner AR (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement In Prokasy WE & Black AH (Eds.), Classical conditioning II: Current research and theory (pp. 64–99). New York: Appleton- Century-Crofts.

Rugg MD (1985). The effects of semantic priming and word repetition on event-related potentials. Psychophysiology, 22, 642–647. [PubMed: 4089090]

Shannon CE, & Weaver W (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Smith NJ, & Levy R (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302–319. doi:10.1016/j.cognition.2013.02.013 [PubMed: 23747651]

Staub A (2015). The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. Language and Linguistics Compass, 9(8), 311–327. doi:10.1111/lnc3.12151

Tauber S, Navarro D, Perfors A, & Steyvers M (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. Psychol Rev, 124(4), 410–441. doi:10.1037/rev0000052 [PubMed: 28358549]

Taylor W (1953). 'Cloze' procedure: A new tool for measuring readability. Journal Q, 30, 415–433.

Traxler MJ, & Foss DJ (2000). Effects of sentence constraint on priming in natural language comprehension. Journal of Experimental Psychology: Learning, Memory and Cognition, 26(5), 1266–1282. doi:10.1037/0278-7393.26.5.1266

Tweedy JR, & Lapinski RH (1981). Facilitating word recognition: Evidence for strategic and automatic factors. The Quarterly Journal of Experimental Psychology Section A, 33(1), 51–59. doi:10.1080/14640748108400768

Tweedy JR, Lapinski RH, & Schvaneveldt RW (1977). Semantic-context effects on word recognition: Influence of varying the proportion of items presented in an appropriate context. Memory & Cognition, 5(1), 84–89. doi:10.3758/BF03209197 [PubMed: 21331872]

Van Berkum JJA, Zwitserlood P, Hagoort P, & Brown CM (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Cognitive Brain Research, 17(3), 701–718. [PubMed: 14561457]

Van Petten C (1993). A comparison of lexical and sentence-level context effects in event-related potentials. Special Issue: Event-related brain potentials in the study of language. Language and Cognitive Processes, 8, 485–531.

van Vliet M, Chumerin N, De Deyne S, Wiersema JR, Fias W, Storms G, & Van Hulle MM (2016). Single-trial ERP component analysis using a spatiotemporal LCMV beamformer. IEEE Trans Biomed Eng, 63(1), 55–66. doi:10.1109/TBME.2015.2468588 [PubMed: 26285053]

Vroomen J, van Linden S, Keetels M, de Gelder B, & Bertelson P (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. Speech Communication, 44(1–4), 55–61. doi:10.1016/j.specom.2004.03.009

Weber K, Lau EF, Stillerman B, & Kuperberg GR (2016). The Yin and the Yang of Prediction: an fMRI study of semantic predictive processing. PLoS One, 11(3). doi:ARTN e014863710.1371/journal.pone.0148637

Wlotko EW, & Federmeier KD (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. NeuroImage, 62(1), 356–366. doi:10.1016/j.neuroimage.2012.04.054 [PubMed: 22565202]

Yan S, Kuperberg GR, & Jaeger TF (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and Delong et al. (2005). bioRxiv. doi:10.1101/143750

Yu AJ, & Dayan P (2005). Uncertainty, neuromodulation, and attention. Neuron, 46(4), 681–692. doi:10.1016/j.neuron.2005.04.026 [PubMed: 15944135]

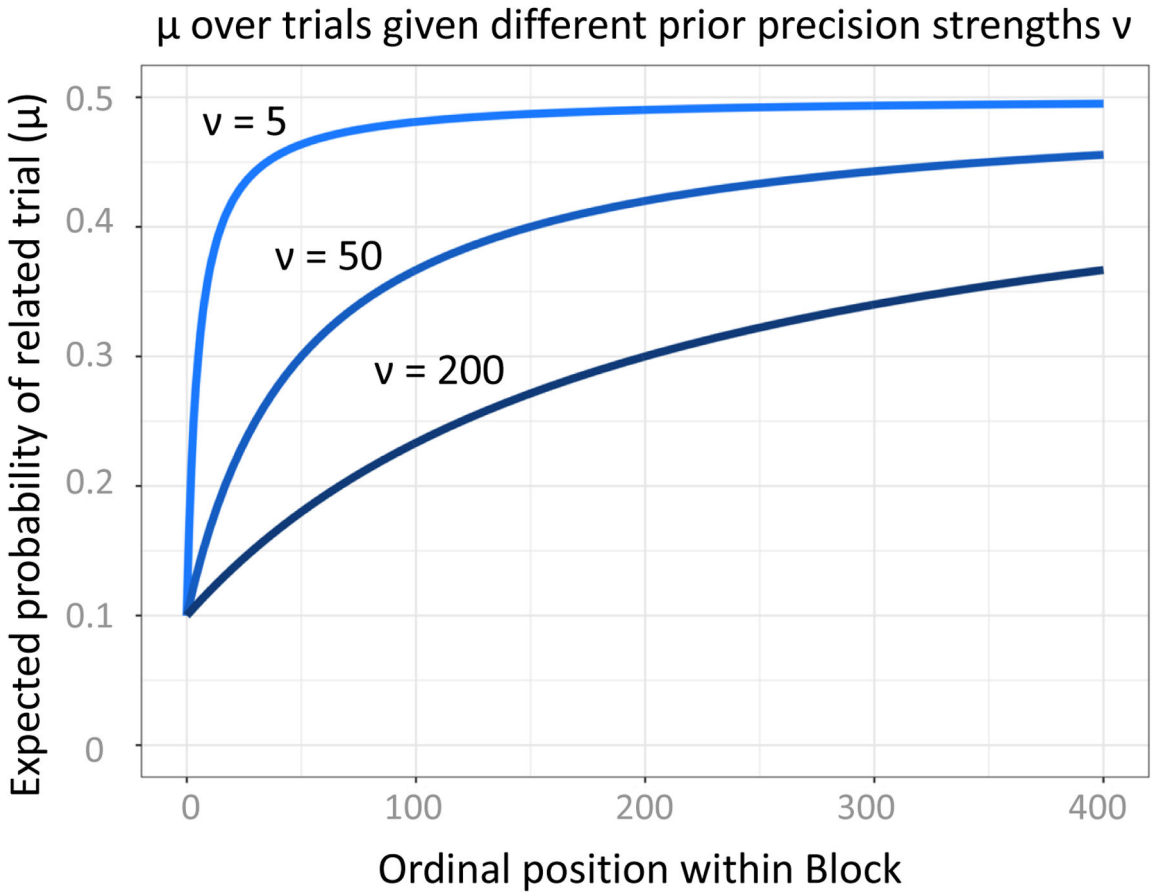

Figure 1.

Sample beliefs μ (probability of a related trial) over the course of Block 2 of the experiment at different values of precision parameter ν. Higher precision (i.e. more certainty in prior beliefs) leads to slower adaptation to the new environment.

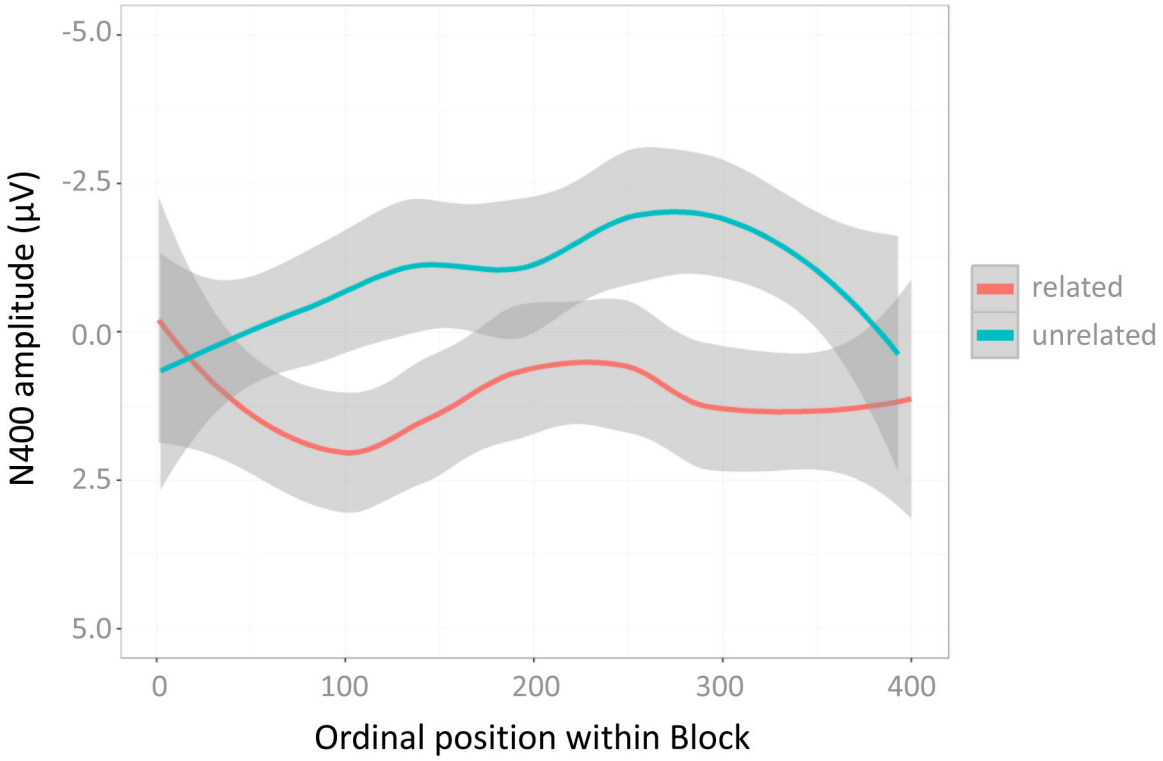

Figure 2.
N400 amplitudes over trials for related and unrelated trials, averaged over all participant data in Block 2.

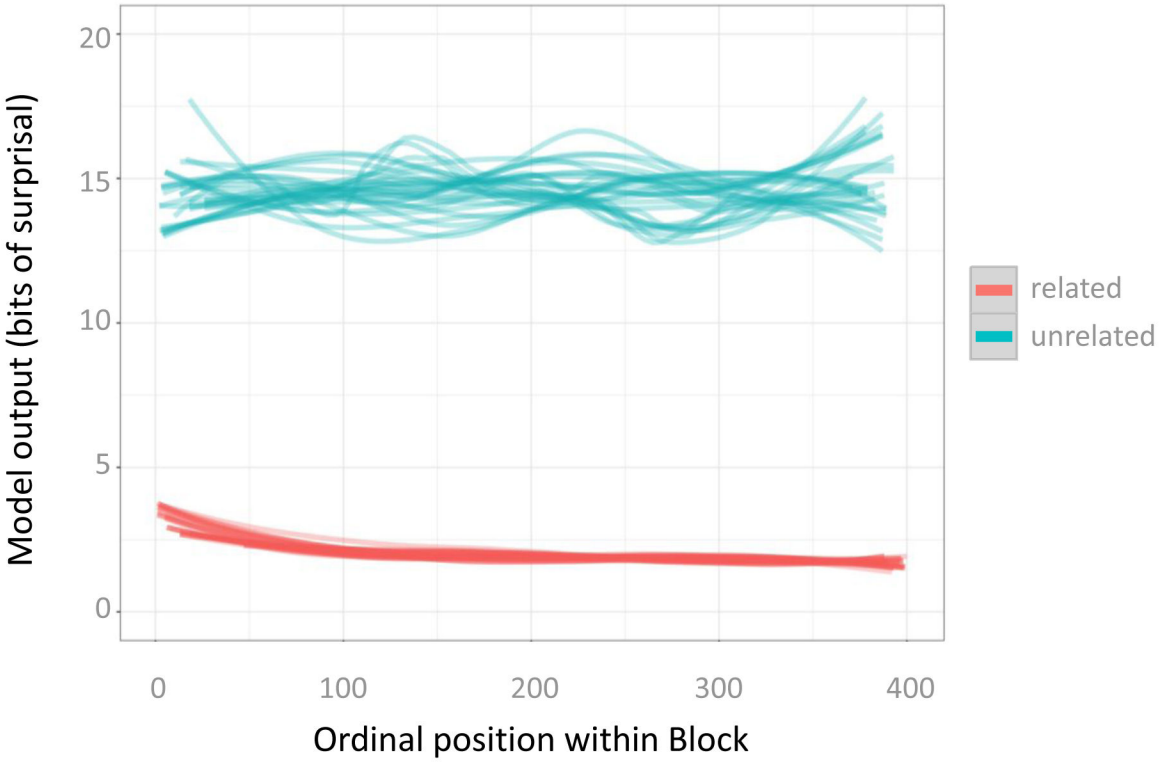

Figure 3.
The output of the rational adaptor model for each participant, calculated based on their idiosyncratic trial ordering in Block 2

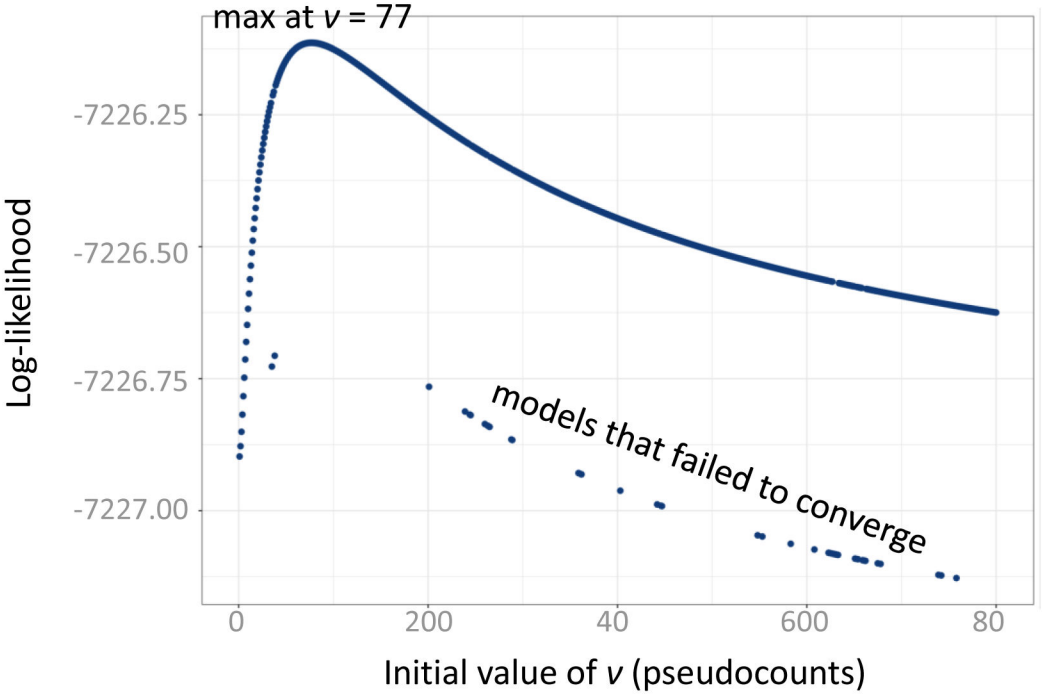

Figure 4.

Log-likelihoods of fitted regression models testing the rational adaptor model with different prior values of the precision parameter ν. Larger (i.e. less negative) log-likelihoods indicate better fit. A prior strength of $\nu = 77$ optimizes the fit of the rational adaptor model to the empirical N400 data.