Language and vision in conceptual processing: Multilevel analysis and statistical power

Pablo Bernabeu[1], Dermot Lynott[2], and Louise Connell[2]

[1]Department of Psychology, Lancaster University, UK

[2]Department of Psychology, Maynooth University, Ireland

Author Note

Pablo Bernabeu https://orcid.org/0000-0003-1083-2460

Dermot Lynott https://orcid.org/0000-0001-7338-0567

Louise Connell https://orcid.org/0000-0002-5291-5267

This manuscript is a draft and includes some appendices. Correspondence can be addressed to Pablo Bernabeu on pcbernabeu@gmail.com. All materials are available at http://doi.org/10.17605/OSF.IO/UERYQ.

Abstract

Research over the past two decades has suggested that conceptual processing depends on both language-based and vision-based information. We tested this interplay at three levels of the data: individuals, words and tasks. To this aim, we drew on three existing data sets that were composed of large samples, and implemented the paradigms of semantic priming, semantic decision and lexical decision. After extending these data sets with language-based and vision-based measures, we performed the analysis using mixed-effects models that included a comprehensive array of fixed effects—including covariates—and random effects. Overall, language-based information was found to be more important than vision-based information. Furthermore, in the semantic priming study—whose task required distinguishing between words and nonwords—, both language-based and vision-based information were more influential when words were presented faster. In addition, higher-vocabulary participants presented a greater effect of language-based similarity than lower-vocabulary participants. In contrast, this pattern did not hold for vision-based information, which was less relevant to the task. The 'relevance advantage' in higher-vocabulary participants also appeared in the semantic decision and the lexical decision studies. Furthermore, we compare the effects of two visual information measures, and discuss the role of measurement instruments in this research topic. Lastly, we estimated the sample size required to reliably examine each effect of interest. We found that 300 participants suffice to examine language-based information contained in words, whereas more than 1,000 participants are necessary to examine vision-based information as well as the interactions between both the former variables and vocabulary size, gender and presentation speed.

Keywords: conceptual processing, semantic priming, semantic decision, lexical decision, language, vision, vocabulary size, statistical power

Language and vision in conceptual processing: Multilevel analysis and statistical power

Over the past decades, research in psycholinguistics has suggested that conceptual processing depends on both language and embodiment systems (Barsalou et al., 2008; Connell & Lynott, 2013; Louwerse, 2011). A comprehensive approach to both these systems requires a direct comparison of them. Studies tackling such a comparison have found that the systems are selectively engaged, following contextual demands (Connell & Lynott, 2013; Louwerse & Connell, 2011; Ostarek & Huettig, 2017; Petilli et al., 2021). For instance, the role of perceptual simulation appears to be augmented in slower responses (Louwerse & Connell, 2011), when words are presented more slowly (Lam et al., 2015), and in tasks prompting deeper semantic processing (Connell & Lynott, 2013; Louwerse & Connell, 2011; Ostarek & Huettig, 2017; Petilli et al., 2021).

In spite of the amount of evidence favouring the interplay between language and embodiment, there are at least 2 reasons to continue testing the interplay theory. Firstly, the coexistence of several systems in a scientific theory must be thoroughly justified, due to the value of simplicity, as epitomised by Occam's razor (Gallese & Lakoff, 2005; Tillman et al., 2015). This point is particularly pressing because the language system has consistently produced larger effect sizes than the embodiment system (Banks et al., 2021; Kiela & Bottou, 2014; Lam et al., 2015; Louwerse et al., 2015; Pecher et al., 1998; Petilli et al., 2021). Consequently, in this research question, we examine whether the interplay between linguistic and visual processing holds in the expected directions and at multiple levels of the experimental structure.

Secondly, mixed evidence has appeared regarding some aspects of the interplay. For instance, whereas some studies have suggested that the language system is activated before the embodiment system (Lam et al., 2015; Louwerse & Connell, 2011), a recent study found the opposite pattern in a lexical decision task (Petilli et al., 2021). Similarly, some evidence has suggested that high-vocabulary participants are more sensitive to linguistic features (Yap et al., 2017), whereas other evidence has suggested the opposite (Yap et al.,

2009, 2012). Another case of mixed evidence regards gender: whereas some evidence has suggested that female participants draw on the language system more prominently than males (Hutchinson & Louwerse, 2013), other research has suggested that this difference is negligible in the general population (Wallentin, 2020). Presumably, the scarcity of statistical power that has ailed some studies in cognitive psychology and neuroscience (Marek et al., 2022; Vasishth & Gelman, 2021) may have affected some studies in the present topic area as well (see Lynott et al., 2014; Montero-Melis et al., 2022). Therefore, in this study, we re-examine longstanding questions using larger-than-average data sets, and calculate the sample size required to reliably detect a range of effects.

Language and vision

To retest the interplay theory, we revisit previous studies that contained larger-than-average samples of participants. Each of our three studies is organised around one *hub* study. In Study 1, the hub is Hutchison et al. (2013), which implements a semantic priming paradigm. In Study 2, the hub is Pexman et al. (2017), which implements a lexical decision paradigm. In Study 3, the hub is Balota et al. (2007), which implements a lexical decision paradigm. Upon this basis, we added word-level variables from other studies to tap into language (Mandera et al., 2017; Wingfield & Connell, 2022) and vision (Lynott et al., 2020; Petilli et al., 2021). Additionally, we included several covariates—or nuisance variables—to allow a rigorous analysis of the effects of interest (Sassenhagen & Alday, 2016). These covariates comprised participant-specific variables (e.g., attentional control), lexical variables (e.g., word frequency) and word concreteness.

Levels of analysis

Experimental data in psycholinguistics often consists of various levels, such as participants, words and tasks. A simultaneous examination of these levels should enhance our understanding of the evidence collected (Ostarek & Bottini, 2021)—for instance, by shedding light on the distribution of explanatory power within and across levels. This multilevel approach is complementary to another approach in the field that seeks to test the causal contribution of different sources of information to conceptual processing—e.g.,

language (Ponari, Norbury, Rotaru, et al., 2018), perception (Stasenko et al., 2014) and action (Speed et al., 2017).

The three levels examined in this study are described below.

Individual level

Variation The role of individual differences in domains such as language, perception, mental imagery and physical experience (e.g., Davies et al., 2017; Dils & Boroditsky, 2010; Fetterman et al., 2018; Holt & Beilock, 2006; Mak & Willems, 2019; Miceli et al., 2022; Pexman & Yap, 2018; Vukovic & Williams, 2015; Yap et al., 2012, 2017).[1] Furthermore, in some topic areas in which individual differences have not featured as prominently, the data have revealed a non-trivial role of individual differences (Kos et al., 2012; Montero-Melis, 2021).

Vocabulary size

Vocabulary size refers to the number of words a person can recognise out of a discrete set. Higher-vocabulary participants are expected to respond faster overall (Pexman & Yap, 2018). Some previous studies have found that the effect of vocabulary size was moderated by variables related to general intelligence, such as processing speed (Ratcliff et al., 2010; Yap et al., 2012). Owing to those findings, and the recommendations from other studies (James et al., 2018; Pexman & Yap, 2018; Sassenhagen & Alday, 2016), the present studies included covariates of vocabulary size, where available. These covariates were measures of intelligence that were not vocabulary-based, namely, attentional control in Study 1 and information uptake in Study 2 (such a variable was not available for Study 3).

Word level

*** to be edited *** Semantic or lexical information in words (e.g., De Deyne et al., 2021; Lam et al., 2015; Lund et al., 1995; Lund & Burgess, 1996; Lynott et al., 2020;

[1] According to Lamiell (2019), 'individual differences' is a misnomer in that the methods used to analyse individual differences (e.g, regression) are not participant-specific. We think that this observation is mitigated in the context of mixed-effects models (as used in our current study), which incorporate by-participant random intercepts and random slopes.

Mandera et al., 2017; Petilli et al., 2021; Pexman et al., 2017; Santos et al., 2011; Wingfield & Connell, 2022)

- Variables in this study: language-based and vision-based information;
- Covariates: lexical variables and word concreteness;
  - Lexical variables: The lexical covariates were selected in each study out of the same 5 variables, which have been identified as important in previous research (Wingfield & Connell, 2022; see General statistical analysis for details).
  - Word concreteness: Adjusting for word concreteness was deemed important due to the pervasive effect of this variable across lexical-semantic tasks (Brysbaert et al., 2014; Connell & Lynott, 2012; Pexman & Yap, 2018). Furthermore, it has been recently suggested that the effect of word concreteness does not stem from perceptual simulation, but instead from modality-independent properties of words (Bottini et al., 2021).

Studies have operationalised the *language system* using measures that capture the relationships among words without explicitly drawing on any sensory or affective modalities. Two main types of linguistic measures exist: those based on text corpora—dubbed *word co-occurrence* measures (Bullinaria & Levy, 2007; Petilli et al., 2021; Wingfield & Connell, 2022)—and those based on associations collected from human participants—dubbed *word association* measures (De Deyne et al., 2016, 2019). Word co-occurrence is more purely linguistic, as word association captures more of the sensory and affective meaning of words (De Deyne et al., 2021). In Studies 1 and 2 below, word co-occurrence measures are used to represent the language system at the word level. In Study 3, however, such a measure could not be implemented. The measurement of priming across consecutive trials was not possible due to the high prevalence of nonword trials throughout the experiment.

The embodiment system has been represented by measures that explicitly draw on perceptual, motor or affective modalities (Fernandino et al., 2022). The perceptual

modalities used in studies have often matched the 5 prototypical senses—vision, hearing, touch, taste, smell (Bernabeu et al., 2017, 2021; Louwerse & Connell, 2011)—and, less often, interoception (Connell et al., 2018). Yet, out of these senses, vision has been most frequently used in research (e.g., Bottini et al., 2021; De Deyne et al., 2021; Pearson & Kosslyn, 2015; Petilli et al., 2021; Yee et al., 2012). The prevalence of vision in research could be due to the prevalence of this sense in the brain (Reilly et al., 2020) and in language (Lynott et al., 2020; Winter et al., 2018). Furthermore, for the purpose of examining individual differences along with word-level variables, vision is easier to tap into than other senses. For instance, the perception of visual motion can be measured using a standard computer screen (Rajananda et al., 2018). Owing to these reasons, the three studies reported below implement the visual modality to represent the embodiment system at the word level.

Task level

*** to be edited ***

- *Variable in this study:* stimulus-onset asynchrony (SOA), present in Study 1 only.[2]).

In addition, the different paradigms applied across the three studies can be compared, albeit cautiously as the studies differ in other influential ways, such as the number of participants and stimulus words. The three tasks examined in the present studies are likely to reflect different degrees of semantic depth, in the following order (for

[2] The names of all variables were slightly adjusted for this text to facilitate their understanding—for instance, by replacing underscores with spaces (see code scripts at https://osf.io/ueryq). One specific case deserves further comment. We use the formula of the 'stimulus-onset asynchrony' (SOA) in this paper, instead of the 'interstimulus interval' (ISI), as the SOA has been more commonly used in previous papers (e.g., Hutchison et al., 2013; Pecher et al., 1998; Petilli et al., 2021; Yap et al., 2017). The difference between these formulas is that the ISI does not count the presentation of the prime word Di Lollo et al. (2004). In the current study, the presentation of the prime word lasts 150 ms. Thus, the 200-ms SOA is equivalent to an ISI of 50 ms, and the 1,200-ms SOA corresponds to an ISI of 1,050 ms (Hutchison et al., 2013). The use of either formula in the analysis would not change our results, as we recoded the levels of the factor as -0.5 and +0.5, and then z-scored those, following the advice of Brauer and Curtin (2018). In our analyses (https://osf.io/ueryq), we used the ISI formula as it was the one present in the data set of Hutchison et al. (2013) (retrieved from https://www.montana.edu/attmemlab/documents/all%20ldt%20subs_all%20trials3.xlsx

background, see Connell & Lynott, 2013; Lam et al., 2015; Ostarek & Huettig, 2017; Versace et al., 2021; Wingfield & Connell, 2022).

1. **Semantic decision** (Study 2) likely elicits the deepest semantic processing, as the instructions of this task ask for a concreteness judgement. In this task, participants are asked to classify words as abstract or concrete.

2. **Semantic priming** (Study 1). The task in the semantic priming paradigm is often lexical decision, as in Study 1 herein. The fundamental characteristic of this paradigm is that, in each trial, a prime word is presented for a very short period before the target word. The prime word is not relevant to the task. Nonetheless, participants process both the prime word and the target word, and this combination allows researchers to analyse responses based on the relationship between those words. Therefore, this paradigm is more deeply semantic than the lexical decision paradigm. Indeed, slower responses in semantic priming studies—reflecting difficult lexical decisions—have been linked to larger priming effects (Balota et al., 2008; Hoedemaker & Gordon, 2014; Yap et al., 2013), suggesting a degree of semantic association that does not appear in the standard lexical decision paradigm.

3. **Lexical decision** (Study 3) is likely the semantically-shallowest task (see Balota & Lorch, 1986; Becker et al., 1997; Connell & Lynott, 2013; de Wit & Kinoshita, 2015; Joordens & Becker, 1997; Ostarek & Huettig, 2017).

Interactions across levels

The three levels are strongly intertwined, and studies have probed into more than one level at once: for instance, word level and individual level (Aujla, 2021; Lim et al., 2020; Pexman & Yap, 2018; Yap et al., 2009), or word level and task level (Al-Azary et al., 2022; Connell & Lynott, 2013; Ostarek & Huettig, 2019; Petilli et al., 2021). Thus, hypotheses are available regarding interactions across levels, as addressed below.

Vocabulary size and linguistic features

Three hypotheses exist regarding the interaction between vocabulary size and lexical-semantic features. First, higher-vocabulary participants—compared to lower-vocabulary ones—might be more sensitive to linguistic features, thanks to a larger number of semantic associations (Connell, 2019; Landauer et al., 1998; Louwerse et al., 2015; Paivio, 1990; Pylyshyn, 1973). For instance, Yap et al. (2017) observed a larger semantic priming effect in higher-vocabulary participants from (Hutchinson & Louwerse, 2013). The second hypothesis, in contrast, states that higher-vocabulary participants would be *less* sensitive to linguistic features, thanks to a more automatic process (Perfetti & Hart, 2002). For instance, Yap et al. (2009) observed a smaller semantic priming effect in higher-vocabulary participants. Similarly, Yap et al. (2012) found that higher-vocabulary participants in a lexical decision task (Balota et al., 2007) were less sensitive to a cluster of lexical and semantic features (i.e., word frequency, semantic neighborhood density and number of senses). Last, the third hypothesis proposes that higher-vocabulary participants might present a selective sensitivity, tailored to the task. For instance, Pexman and Yap (2018) analysed the semantic decision task of Pexman et al. (2017), in which participants assessed the concreteness of words. Pexman and Yap found that higher-vocabulary participants—compared to lower-vocabulary ones—were more sensitive to word concreteness and less sensitive to word frequency and age of acquisition, which were less relevant to the task (also see Lim et al., 2020). The three hypotheses herein were tested in the present study by revisiting some of the aforementioned studies (Balota et al., 2007; Hutchinson & Louwerse, 2013; Pexman et al., 2017).

Vocabulary size and visual strength

Research demonstrating the interplay between linguistic and embodied information (Connell, 2019; Louwerse et al., 2015; Paivio, 1990) affords the hypothesis that lower-vocabulary participants—compared to higher-vocabulary ones—might profit more from sensory, motor, affective and social associations of the words.

Gender and linguistic features

In addition, the language system is expected to be more important in female than male participants (Burman et al., 2008; Hutchinson & Louwerse, 2013; Jung et al., 2019; Ullman et al., 2008), provided that this interaction effect is not too small to be detected (Wallentin, 2020).

In another study, Schmidtke et al. (2018) collected human ratings on the association between words that make up compounds. Some compounds were found to be relatively transparent—e.g., *doorbell*—, and others relatively opaque—e.g., *deadline*. Furthermore, Schmidtke et al. (2018) found an interaction between semantic transparency and participants' reading experience (the latter measure encompassing vocabulary size and exposure to printed materials): namely, more experienced readers processed transparent compounds more quickly than opaque ones, whereas less experienced readers processed opaque compounds more quickly than transparent ones. This interaction suggests _____. Thus, _____.

Power analysis

Statistical power depends on the following factors: (1) sample size—comprising the number of participants, items, trials, etc.—, (2) effect size, (3) measurement variability and (4) number of comparisons being performed. Out of these, sample size is the factor that can best be controlled by researchers (Kumle et al., 2021). The three studies we present below, containing larger-than-average sample sizes, offer an opportunity to conduct an a-priori power analysis to help determine the sample size of future studies (Albers & Lakens, 2018).

Motivations

Insufficient statistical power lowers the reliability of effect sizes, and increases the likelihood of false positive results—i.e., Type I error—and of false negative results—i.e., Type II error (Gelman & Carlin, 2014; Loken & Gelman, 2017; Tversky & Kahneman, 1971; von der Malsburg & Angele, 2017). For instance, Vasishth and Gelman (2021) illustrate how, in low-powered studies, effect sizes associated with significant results tend to

be overestimated (also see Vasishth, Mertzen, et al., 2018).

Over the past decade, replication studies and power analyses have uncovered insufficient sample sizes in psychology (Brysbaert, 2019; Lynott et al., 2014; Montero-Melis et al., 2017, 2022; Rodríguez-Ferreiro et al., 2020; Vasishth, Mertzen, et al., 2018). In the neighbouring field of neuroscience, Marek et al. (2022) recently estimated the sample size that is required to reliably study the mapping between individual differences, such as fluid intelligence, and brain structures. Marek et al. found that the current median of 25 participants contributing to one of these studies contrasted with the thousands of participants—around 10,000—that would be needed for a well-powered study (also see Button et al., 2013).

More topic-specific power analyses are necessary due to three reasons. Firstly, power analyses provide greater certainty on the reasons behind non-replications (e.g., Open Science Collaboration, 2015), and behind non-significant results. Indeed, insufficient power is one of the possible reasons, alongside procedural errors or subtler methodological differences, and publication bias (Anderson et al., 2016; Barsalou, 2019; Corker et al., 2014; Gilbert et al., 2016; Williams, 2014). Currently, if we consider—for instance—research on individual differences, we will find several non-significant results, both in behavioural studies (e.g., Hedge et al., 2018; Rodríguez-Ferreiro et al., 2020; for a Bayes factor analysis, see Rouder & Haaf, 2019) and in neuroscientific studies (e.g., Diaz et al., 2021). A greater availability of power analyses in specific topics will at least shed light on the role of power in the results. Secondly, power analyses facilitate the identification of sensible sample sizes for future studies. Thirdly, power analyses help maximise the use of research funding in the long term by fostering studies that are more replicable (see Vasishth & Gelman, 2021).

Methods

The analytical method was largely similar across the three studies. Below, we present the commonalities in the statistical analyses and in the power analyses.

General method for statistical analysis

The statistical analysis was designed to examine the unique contribution of each effect of interest. In all three studies, the dependent variable—response time (RT)—was z-scored around each participant's mean to curb the influence of each participant's baseline speed (Balota et al., 2007; Kumar et al., 2020; Lim et al., 2020; Pexman et al., 2017; Pexman & Yap, 2018; Yap et al., 2012, 2017). This was important because the size of experimental effects is known to increase with longer RTs (Faust et al., 1999). Next, binary predictors were recoded into continuous variables (Brauer & Curtin, 2018). Specifically, participants' gender was recoded as follows: male = -0.5; female = 0.5; NA = 0. The SOAs in Study 1 were recoded as follows: 150 ms = -0.5; 1,200 ms = 0.5. Finally, following Brauer and Curtin (2018), all predictors were z-scored (i.e., $M \approx 0$, $SD \approx 1$).

Several covariates—or nuisance variables—were included in each study, to allow a rigorous analysis of the effects of interest (Sassenhagen & Alday, 2016). Unlike the effects of interest, these covariates were not critical to the research question. They comprised participant-specific variables (e.g., attentional control), lexical variables (e.g., word frequency) and word concreteness.

The lexical covariates were selected in every study out of the same 5 variables, which had been used as covariates in Wingfield and Connell (2022; also see Petilli et al., 2021). They comprised: number of letters (i.e., orthographic length), word frequency, number of syllables (both the latter from Balota et al., 2007), orthographic Levenshtein distance (Yarkoni et al., 2008) and phonological Levenshtein distance (Suárez et al., 2011). The selection among these candidates was performed because some of them were highly intercorrelated—i.e., $r > .70$ (Dormann et al., 2013; Harrison et al., 2018). The correlations and the selection models are available in Appendix A.

Word concreteness was included due to its correlation with visual strength, as shown in each study below. A recent study suggested that the psycholinguistic import of concreteness is fundamentally lexical, and does not involve perceptual simulation (Bottini et al., 2021).

The participant-specific covariates were measures akin to fluid intelligence, and were

included due to their relationship with vocabulary size, as previous studies have expressed the desirability of including such covariates (James et al., 2018; Pexman & Yap, 2018). As described in each study, these covariates were used in Studies 1 and 2, but not in Study 3, as such a variable was not available.

Random effects

The participants and the stimulus items were crossed in the three studies. That is, each participant was presented with a subset of the stimulus words. Conversely, each word was presented to a subset of participants. Therefore, linear mixed-effects models were implemented. These models included a maximal random-effects structure, with by-participant and by-item random intercepts, and the appropriate random slopes for all effects of interest (Barr et al., 2013). In the semantic priming study, the items were prime-target pairs, whereas in the semantic decision and lexical decision studies, the items were individual words. In the case of interactions, random slopes were included only when the interacting variables varied within the same unit (Brauer & Curtin, 2018)—e.g., an interaction of two variables varying within participants (only present in Study 1). Where required due to convergence warnings, random slopes for covariates were removed, as inspired by Remedy 11 from Brauer and Curtin (2018). In this regard, whereas Brauer and Curtin (2018) contemplate the removal of random slopes for covariates only when the covariates are not interacting with any effects of interest, we removed random slopes for covariates even if they interacted with effects of interest because these interactions were covariates themselves (covariates are indicated in the results tables).

To avoid inflating the Type I error rate (false positives), the random slopes for the effects of interest (indicated by non-shaded rows in the results tables below) were never removed (see Table 17 in Brauer & Curtin, 2018; for an example of this approach, see Diaz et al., 2021). This approach arguably provides a better protection against false positives (Barr et al., 2013; Brauer & Curtin, 2018; Singmann & Kellen, 2019) than the practice of removing random slopes when they do not significantly improve the fit (Baayen et al., 2008; Bates et al., 2015; e.g., Bernabeu et al., 2017; Pexman & Yap, 2018; but also see Matuschek et al., 2017).

Frequentist analysis

P values were calculated using the Kenward-Roger approximation for degrees of freedom (Luke, 2017), in the R package 'lmerTest', Version 3.1-3 (Kuznetsova et al., 2017). The latter package in turn used 'lme4', Version 1.1-26 (Bates et al., 2015; Bates et al., 2021). To facilitate the convergence of the models, the maximum number of iterations was set to 1 million. Diagnostics regarding convergence and normality are provided in Appendix B. Effects that were non-significant or very small are best interpreted by considering their 95% confidence intervals (Cumming, 2014), which are shown in the results tables as well as in plots.

Bayesian analysis

A Bayesian analysis was performed to complement the estimates that had been obtained in the frequentist analysis. Whereas the focus of the frequentist analysis had been hypothesis testing, using *p* values, the purpose of the Bayesian analysis was parameter estimation. Accordingly, we estimated the posterior distribution of every effect, without calculating Bayes factors (for other examples of this approach, see Milek et al., 2018; Pregla et al., 2021; Rodríguez-Ferreiro et al., 2020; for comparisons between estimation and hypothesis testing, see Cumming, 2014; Kruschke & Liddell, 2018; Rouder et al., 2018; Schmalz et al., 2021; Tendeiro & Kiers, 2019, in press; van Ravenzwaaij & Wagenmakers, 2021). In the estimation approach, the estimates are interpreted by considering the position of their credible intervals in relation to the expected effect size. That is, the closer an interval is to an effect size of 0, the smaller the effect of that predictor. For instance, an interval that is symmetrically centred on 0 indicates a very small effect, whereas—in comparison—an interval that does not include 0 at all indicates a far larger effect.

This analysis served two purposes: firstly, to ascertain the interpretation of the smaller effects—which were identified as unreliable in the power analyses—, and secondly, to complement the estimates obtained in the frequentist analysis. The latter purpose was pertinent because the frequentist models presented convergence warnings—even though it must be noted that a previous study found that frequentist and Bayesian estimates were similar despite convergence warnings appearing in the frequentist analysis

(Rodríguez-Ferreiro et al., 2020). Furthermore, the complementary analysis was pertinent because the frequentist models presented residual errors that deviated from normality—even though mixed-effects models are fairly robust to such a deviation (Knief & Forstmeier, 2021; Schielzeth et al., 2020). Owing to these precedents, we expected to find broadly similar estimates in the frequentist and the Bayesian analyses. Each frequentist model in the three studies has a Bayesian counterpart, with the exception of a sub-analysis performed in the first study (semantic priming), which included 'vision-based similarity' as a predictor.

The R package 'brms', Version 2.17.0, was used for this analysis (Bürkner, 2018; Bürkner et al., 2022).

Priors. The priors were established by inspecting the effect sizes obtained in previous studies as well as the effect sizes obtained in our frequentist analyses of Studies 1, 2 and 3.

The previous studies that were considered for determining the priors were selected because they had used experimental paradigms and analytical procedures similar to those used in the current studies. Specifically, the paradigms were (I) semantic priming with a lexical decision task—as in Study 1—, (II) semantic decision—as in Study 2—,and (III) lexical decision—as in Study 3. The analytical procedures consisted of the z-scoring of the dependent and the independent variables. We found two studies that matched these characteristics: Lim et al. (2020) (see Table 5 therein) and Pexman and Yap (2018) (see Tables 6 and 7 therein). These studies and the frequentist analyses reported below yielded effect sizes smaller than ±0.30. The bounds of this range were determined by the results from Pexman and Yap (2018), who found a word concreteness effect of $\beta = 0.41$ in the concrete-words analysis, and an effect of $\beta = 0.20$ in the abstract-words analysis. Since we did not separate abstract from concrete words, we averaged the former values, and set -0.30 as the lower bound, and 0.30 as the upper bound. This provided the basis for informative priors. Specifically, in these informative priors, 95% of values would fall within the range [-0.30, 0.30].

Next, we considered the direction of effects. In the results of Lim et al. (2020) and

Pexman and Yap (2018), and in our frequentist results, some effects consistently presented a negative direction—i.e., an inhibitory effect on RT—, whereas some other effects were consistently positive. We only incorporated the direction of effects into the priors in cases of large effects that had presented consistent directions in previous studies and in our frequentist analyses. These criteria were matched by the following variables: word frequency—with a negative direction, as higher word frequency leads to shorter RTs (Brysbaert et al., 2018; Brysbaert et al., 2016; Lim et al., 2020; Mendes & Undorf, 2021; Pexman & Yap, 2018)—, number of letters and number of syllables—both with positive directions (Barton et al., 2014; Beyersmann et al., 2020; Pexman & Yap, 2018)—, and orthographic Levenshtein distance—with a positive direction (Cerni et al., 2016; Dijkstra et al., 2019; Kim et al., 2018; Yarkoni et al., 2008). We did not incorporate information about the direction of the word concreteness effect as this effect can follow different directions in abstract and concrete words (Pexman & Yap, 2018), and we analysed both sets of words together. Last, it is noteworthy that some previous studies have integrated effect direction in some priors (e.g., Stone et al., 2021), but most have not (e.g., Pregla et al., 2021; Rodríguez-Ferreiro et al., 2020; Stone et al., 2020). In conclusion, the four predictors that had directional priors were covariates (also known as nuisance variables). All the other predictors had priors centred on 0.

These priors were used on the fixed effects and on the corresponding standard deviation parameters. The correlation among the random effects had a weakly-informative prior, LKJ(2) (Lewandowski et al., 2009), which assumes that high correlations among the random effects are rare (also used in Rodríguez-Ferreiro et al., 2020; Stone et al., 2021; Stone et al., 2020; Vasishth, Nicenboim, et al., 2018).

Prior distributions and prior predictive checks. We aimed to perform prior sensitivity analyses of our results. Prior sensitivity analyses are checks that assess the influence of different priors on the results (Lee & Wagenmakers, 2014; Schoot et al., 2021; Stone et al., 2020). The range of priors established for this purpose varied in their standard deviations. Their means were the same, most being centred on 0, as explained above.

The priors we settled upon—shown in Figure 1—comprised an 'informative' prior

($SD = 0.1$), a 'weakly-informative' one ($SD = 0.2$) and a 'diffuse' one ($SD = 0.3$). These priors resembled others from previous psycholinguistic studies (Pregla et al., 2021; Stone et al., 2021; Stone et al., 2020). For instance, Stone et al. (2020) used the following priors: $Normal(0, 0.1)$, $Normal(0, 0.3)$ and $Normal(0, 1)$. The range of standard deviations we used—i.e., 0.1, 0.2 and 0.3—was narrower than those of previous studies because our dependent variable and our predictors were z-scored, resulting in small estimates and small SDs (see Lim et al., 2020; Pexman & Yap, 2018).

The adequacy of each of these priors was assessed with prior predictive checks, in which we compared the observed data to data predicted by the priors (Schoot et al., 2021). In these checks, we also tested the adequacy of two model-wide distributions: the traditional Gaussian distribution (default in most analyses) and an exponentially modified Gaussian—dubbed 'ex-Gaussian'—distribution (Matzke & Wagenmakers, 2009). The ex-Gaussian distribution was considered because the residual errors of the frequentist models were not normally distributed (Lo & Andrews, 2015), and because this distribution was found to be more appropriate than the Gaussian one in a related, previous study (see supplementary materials of Rodríguez-Ferreiro et al., 2020). The ex-Gaussian distribution had an identity link function, which preserves the interpretability of the coefficients, as opposed to a transformation applied directly to the dependent variable (Lo & Andrews, 2015).

Prior predictive checks revealed that the priors were adequate, and that the ex-Gaussian distribution was more appropriate than the Gaussian one, converging with Rodríguez-Ferreiro et al. (2020) (see the corresponding plots in Appendix C). Therefore, the ex-Gaussian distribution was used in the final models.

Prior sensitivity analysis. In the main analyses, the informative, weakly-informative and diffuse priors were used in separate models. In other words, in each model, all priors had the same degree of informativeness (as done in Pregla et al., 2021; Rodríguez-Ferreiro et al., 2020; Stone et al., 2021; Stone et al., 2020). In this way, a prior sensitivity analysis was performed to acknowledge the likely influence of the priors on the posterior distributions—that is, on the results (Lee & Wagenmakers, 2014; Schoot et al.,

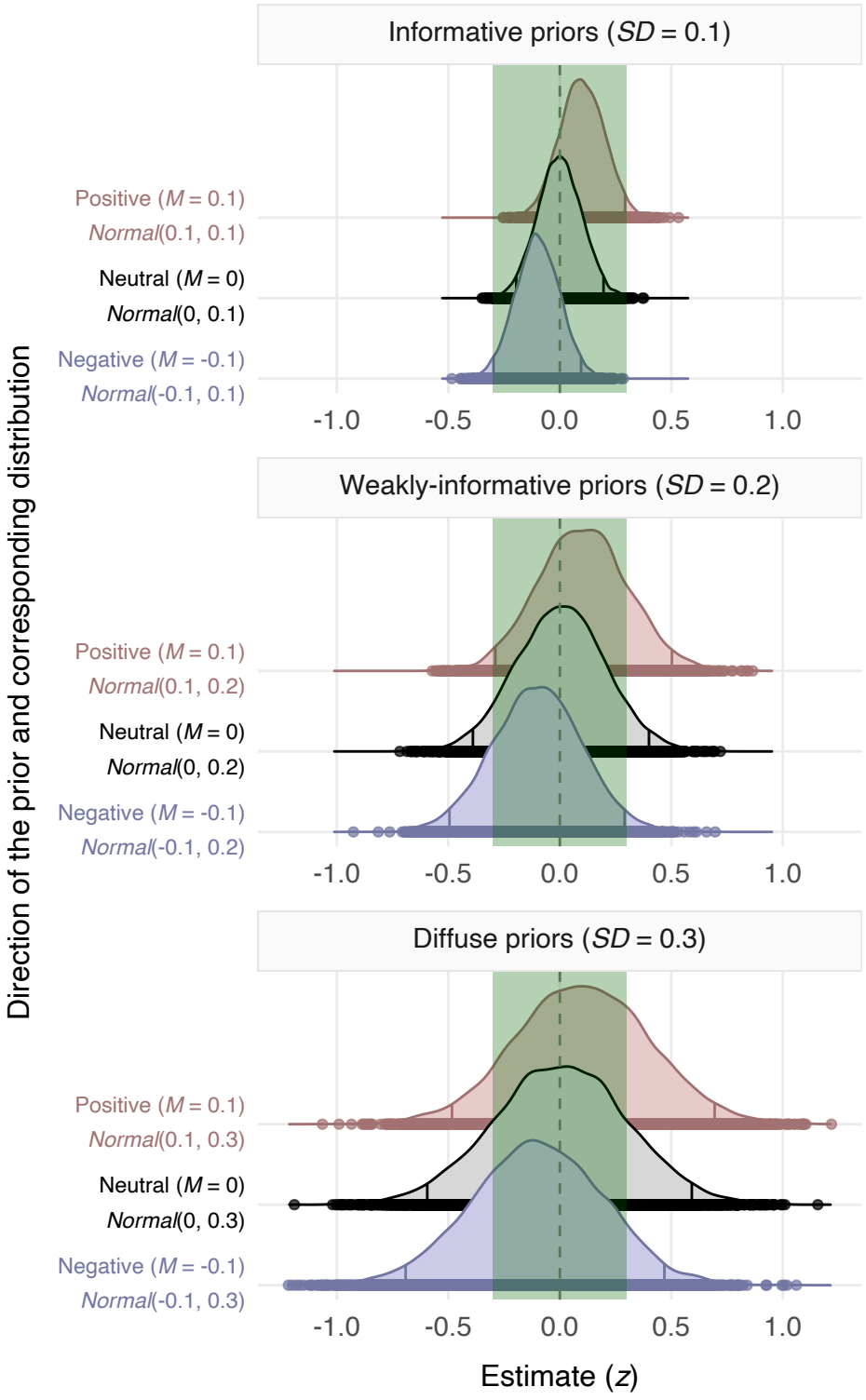

Figure 1

Priors used in the three studies. The green vertical rectangle shows the range of plausible effect sizes based on previous studies and on our frequentist analyses.

2021; Stone et al., 2020).

Posterior distributions. Posterior predictive checks were performed to assess the fit between the observed data and new data predicted by the posterior (Schoot et al., 2021). These checks are available in Appendix C.

Convergence. Where convergence was not reached in a model, as indicated by $\widehat{R} > 1.01$ (Schoot et al., 2021; Vehtari et al., 2021), the number of iterations was increased. Furthermore, where necessary, the random slopes for covariates were removed (Brauer & Curtin, 2018). The resulting random effects in these models were largely the same as those present in the frequentist models. The only exception regarded the lexical decision models. In the frequentist model for lexical decision, the random slopes for covariates were removed due to convergence warnings, whereas in the Bayesian analysis, these random slopes did not have to be removed as the models converged.

The Bayesian models in the semantic decision study could not be made to converge, and the results were not valid. Therefore, those estimates are not shown in the main text but in Appendix E.

General method for statistical power analysis

Power curves based on Monte Carlo simulations were performed for most of the effects of interest, using the R package 'simr', Version 1.0.5 (Green & MacLeod, 2016). Obtaining power curves for a range of effects in each study allows for a comprehensive assessment of the plausibility of the power estimated for each effect.

Monte Carlo simulations are performed by running the statistical model a large number of times, under slight, random variations of the dependent variable (Green & MacLeod, 2016; for a comparable approach, see Loken & Gelman, 2017). The power to detect each effect of interest is calculated by dividing the number of times that the effect is significant by the total number of simulations run. For instance, if an effect is significant on 85 simulations out of 100, the power for that effect is 85% (Kumle et al., 2021). The sample sizes tested in the semantic priming study ranged from 50 to 800 participants, whereas those tested in the semantic decision and lexical decision studies ranged from 50 to

2,000 participants. These sample sizes were unequally spaced to limit the computational requirements. They comprised the following: 50, 100, 200, 300, 400, 500, 600, 700, 800, 1,200, 1,600 and 2,000 participants.[3] The variance of the results decreases as more simulations are run. In each of our three studies, 200 simulations (as in Brysbaert & Stevens, 2018) were run for each effect of interest and for each sample size under consideration. Thus, for a power curve examining the power for an effect across 12 sample sizes, 2,400 simulations were run. In each study, the item-level sample size—i.e., the number of words—was not modified. Thus, each of the power curves we present assume the same number of words that existed in each of our studies (these numbers are detailed in each study below). *P* values were calculated using the Satterthwaite approximation for degrees of freedom (Luke, 2017).

It is difficult to determine an effect size for each effect examined in a power analysis, as the amount and the scope of relevant research are usually finite and biased (Albers & Lakens, 2018; Gelman & Carlin, 2014; Kumle et al., 2021). Power analyses sometimes use the original effect sizes from previous studies (e.g., Pacini & Barnard, 2021; Villalonga et al., 2021). In contrast, some authors have opted to reduce the previous effect sizes to account for factors that can influence the effect size of the planned study. First, publication bias and low-powered studies cause published effect sizes to be inflated (Brysbaert, 2019; Loken & Gelman, 2017; Open Science Collaboration, 2015; Vasishth, Mertzen, et al., 2018; Vasishth & Gelman, 2021). Second, there might be differences between the studies used in the power analysis and the study to be conducted. Some of these differences could be foreseeable—for instance, if they are due to a limitation in the literature available for the power analysis—, whereas other differences might arise from unexpected circumstances arising during the project or from random variation (Barsalou, 2019). Reducing the effect size in the power analysis leads to an increase of the sample size of the planned study (Brysbaert & Stevens, 2018; Green & MacLeod, 2016; Hoenig & Heisey, 2001). The reduced effect size—sometimes dubbed the smallest effect size of interest—is often set with a certain arbitrariness. For instance, Fleur et al. (2020) applied a reduction of 1/8 (i.e.,

[3] For the semantic priming study, the remaining sample sizes up to 2,000 participants have not finished running yet. Upon finishing, they will be reported in this manuscript.

12.5%), whereas Kumle et al. (2021) applied a 15% reduction. In the present study, a reduction of 20% was applied to every effect examined in the power analysis. By comparison with the power analyses reviewed in this paragraph, the present reduction will lead to a relatively-conservative estimate of required sample sizes. Yet, it is intrinsically difficult to determine how conservative a power analysis is, once we consider the insufficient power of most studies in psychology, and the resulting overestimation of effect sizes.

Both the primary analysis and the power analysis were performed in R (R Core Team, 2021). Version 4.0.2 was used for the frequentist analysis, Version 4.1.0 was used for the Bayesian analysis, and Version 4.1.2 was used for fast operations such as plotting. All the statistical and the power analyses were run on the High-End Computing facility at Lancaster University.[4]

Study 1: Semantic priming

The core data set in this study was that of the Semantic Priming Project (Hutchison et al., 2013; also see Yap et al., 2017). Out of the tasks delivered in the experiment, we used the lexical decision one because it was most relevant to our forthcoming research. In the lexical decision task, participants judged whether strings of letters constituted real words (e.g., 'building') or nonwords (e.g., 'gop'). Crucially, in the experimental manipulation that characterises semantic priming, a prime word was presented before the target word in each trial. Since prime words facilitate the comprehension of target words, the participants' responses to the targets can be analysed as a function of the semantic relationship between primes and targets (Hoedemaker & Gordon, 2014).

In some studies, the association between prime and target words has been examined in terms of related versus unrelated pairs (Pecher et al., 1998; Trumpp et al., 2013), and in terms of first- and second-order relationships (Hutchison et al., 2013). In contrast to these categorical associations, other studies have measured the association between the prime and the target words using language-based similarity estimates (Günther et al., 2016a, 2016b; Hutchison et al., 2008; Jones et al., 2006; Lam et al., 2015; Lund et al., 1995; Lund

[4] Information about this facility is available at https://answers.lancaster.ac.uk/display/ISS/High+End+Computing+%28HEC%29+help

& Burgess, 1996; Mandera et al., 2017; McDonald & Brew, 2002; Padó & Lapata, 2007; Petilli et al., 2021; Wingfield & Connell, 2022). In one of these studies, Mandera et al. (2017) found that the latter computational measures outperformed human-based associations at explaining the priming effect.

Priming associations beyond the linguistic realm have also been examined, with early studies finding perceptual priming effects (Flores d'Arcais et al., 1985; Schreuder et al., 1984). Yet, the earliest findings were soon reframed by Pecher et al. (1998), who conducted a follow-up with an improved design, and observed the perceptual priming effect only when the word processing task was preceded by a visually-intensive task (Pecher et al., 1998). This moderating condition, replicated by Yee et al. (2012), is consistent with other conceptual-processing studies using a non-priming paradigm (Ostarek & Huettig, 2017). However, a considerable number of studies have observed perceptual priming even in the absence of a preparatory perceptual task. A set of these studies used the Conceptual Modality Switch paradigm, in which the primes and the targets are presented in separate, consecutive trials (Bernabeu et al., 2017; Collins et al., 2011; Hald et al., 2011, 2013; Louwerse & Connell, 2011; Lynott & Connell, 2009; Pecher et al., 2003; Trumpp et al., 2013). The other set of studies, using the more standard priming intervals, are described below.

Lam et al. (2015) conducted a semantic priming experiment containing a lexical decision task. Participants were instructed to respond whether the prime word and the target word in each trial were both real words or pseudowords. The semantic-priming manipulation consisted of the following types of associations between the prime and the target words: (1) semantic association (e.g., bolt → screwdriver), (2) action association (e.g., housekey → screwdriver), (3) visual association (e.g., soldering iron → screwdriver), and (4) no association (e.g., charger → screwdriver). In addition, four conditions were present that consisted of SOAs of 500, 650, 800 and 1,400 ms. In the results, Lam et al. firstly observed priming effects of the semantic-association type with all SOAs. Secondly, the authors observed motor-association priming effects with the SOAs of 500, 650 and 1,400 ms. Lastly, they observed visual-association priming effects only with the SOA

of 1,400 ms. Overall, semantic-association priming—corresponding to language-based association—was more consistent than the priming based on visual and action associations. This greater influence of the language system converges with other studies on semantic priming (Lam et al., 2015; Pecher et al., 1998; Petilli et al., 2021) and with studies using other paradigms (Banks et al., 2021; Kiela & Bottou, 2014; Louwerse et al., 2015).

Similarly, the results of Lam et al. (2015) regarding the time course of language-based and vision-based priming were consistent with a wealth of literature observing that perceptual systems, such as vision, are activated later than the language system (Barsalou et al., 2008; Connell & Lynott, 2013; Louwerse & Connell, 2011; Santos et al., 2011). For instance, studies using electroencephalography have found that perceptual priming effects emerged in the first 300 ms. Following the 300-ms stage, the perceptual priming increased in some studies (Amsel et al., 2014; Bernabeu et al., 2017), whereas other studies it stabilised (Kiefer et al., 2022), and in a third set of studies, the effect fluctuated (Amsel, 2011). Taken together, the most prevalent pattern is of a gradual accumulation of information throughout word processing (also see Hauk, 2016), which is consistent with the accummulation of information required for the integration of context in sentences (Hald et al., 2006). This progression invites two hypotheses regarding SOA in the current study: (1) that language-based information will be more prevalent than vision-based information in both the short and the long SOA, and (2) that vision-based information will be more prevalent in the long SOA than in the short one. Consistent with the first hypothesis, Petilli et al. (2021) observed that language-based information was more influential in both the short and the long SOA, within the data set of Hutchison et al. (2013). The second hypothesis, in contrast, is challenged by other findings. First, Hutchison (2003) concluded that the vision-based priming effect was negligible. Second, other studies have observed the effect only when the word processing task was preceded by a visually-intensive task (Pecher et al., 1998; Yee et al., 2012). Third, Petilli et al. observed vision-based priming with the short SOA (150 ms) but not with the long one (1,200 ms).

The findings of Petilli et al. (2021) were based on a novel analysis of the data set of Hutchison et al. (2013). The strengths of Petilli et al.'s study centred on the simultaneous

analysis of language- and vision-based similarity using predictors that were continuous (see Cohen, 1983; Günther et al., 2016a; Mandera et al., 2017) and not based on human ratings (cf. Hutchison et al., 2008, 2013; Lam et al., 2015; Pecher et al., 1998). The authors used such measures to pursue a comprehensive analysis that was not affected by the problem of circularity between the independent and the dependent variables. The circularity problem obtains even though the participants who contributed to the ratings—or 'norms'—are always different from those who participate in the main experiment.

Petilli et al. operationalised language-based similarity based on text-based co-occurrence, producing a continuous measure in contrast to the categorical factors used earlier. Next, Petilli et al. created a visual-similarity measure by retrieving ImageNet images for each word, and training vector representations on those images using neural networks. Importantly, the authors compared the resulting visual-similarity measure to previous scores based on human ratings (Pecher et al., 1998), and found a comparable pattern. Using these materials, Petilli et al. examined language-based and vision-based priming in two tasks—lexical decision and naming—and with both a short and a long SOA. In lexical decision, the largest effect observed by the authors was that of language-based priming with the short SOA (150 ms). The second largest effect was that of language-based priming with the long SOA (1,200 ms). Next, the weakest effect that was significant was that of vision-based priming with the short SOA. Last, there was no effect of vision-based priming with the long SOA. Petilli et al. (2021) explained the absence of vision-based priming with the long SOA in lexical decision by contending that visual activation had likely decayed before participants processed the target words (also see Yee et al., 2011), owing to the limited semantic processing required for lexical decision (also see Balota & Lorch, 1986; Becker et al., 1997; Connell & Lynott, 2013; de Wit & Kinoshita, 2015; Joordens & Becker, 1997; Ostarek & Huettig, 2017). Therefore, the authors suggested that perceptual simulation does *not* outlast language-based processing in lexical decision, in contrast to the longer latency found in other tasks (Barsalou et al., 2008; Louwerse & Connell, 2011).

In the naming task of Petilli et al. (2021), the largest effect was that of

language-based priming with the long SOA. The second largest effect was that of language-based priming with the short SOA. Last, there was no effect of vision-based priming with either SOA. This finding contrasted with Connell and Lynott (2014), who found facilitatory effects of visual strength in both lexical decision and naming. Petilli et al. explained the lack of vision-based priming in the naming task by alluding to the lower semantic depth of this task—compared to lexical decision—, and the mixture of visual and auditory processing in the naming task (Connell & Lynott, 2014).

In the present study, we revisited the interplay between linguistic and perceptual simulation by conceptually replicating Petilli et al. (2021). Specifically, we used the same primary data set (Hutchison et al., 2013), and a language-based similarity measure that was very similar to that used by Petilli et al. (2021; both our measure and theirs originated from Mandera et al., 2017). In contrast, our predictors in the domain of vision differed. Whereas Petilli et al. used a human-independent measure based on images from the Internet (see description above), we drew on visual strength—from the modality ratings of Lynott et al. (2020)—and calculated the difference in visual strength between the prime and the target word in each trial.[5]

Methods

Effects of interest

- *Z*-scored vocabulary size [`z_vocabulary_size`; calculated from average of `vocaba`, `vocabb` and `vocabc` in Hutchison et al. (2013)]. The test used by Hutchison et al. (2013) comprised a synonym test, an antonym test, and an analogy test, all three extracted from the Woodcock–Johnson III diagnostic reading battery (Woodcock et al., 2001). We operationalised the vocabulary measure as the mean score across the three tasks per participant.

- *Z*-scored, recoded participants' gender [`z_recoded_participant_gender`; calculated from `gender` in Hutchison et al. (2013)]

[5] These measures are compared at the end of the Results section.

- Z-scored language-based similarity between prime and target words [`z_cosine_similarity`]. This measure was calculated using a semantic space from Mandera et al. (2017), which the authors found to be the second-best predictor (R^2 = .465) of the semantic priming effect in the lexical decision task of Hutchison et al. (2013) (we could not use the best semantic space, R^2 = .471, owing to computational limitations). The second-best semantic space (see first row in Table 5 of Mandera et al., 2017) was based on lemmas from a subtitle corpus, processed in a Continuous Bag Of Words model, and the space had 300 dimensions and a window size of 6.
- Z-scored vision-based information in words [`z_visual_rating_diff`; calculated from `Visual.mean` in Lynott et al. (2020)]
- Z-scored, recoded SOA [`z_recoded_interstimulus_interval`; calculated from `isi` in Hutchison et al. (2013)]

The final data set contained 496 participants, 5,943 prime-target pairs, and 345,666 RTs. On average, there were 697 prime-target pairs per participant (SD = 33.34), and conversely, 58 participants per prime-target pair (SD = 4.25).

Figure 2 shows the zero-order correlations among the predictors and the dependent variable.

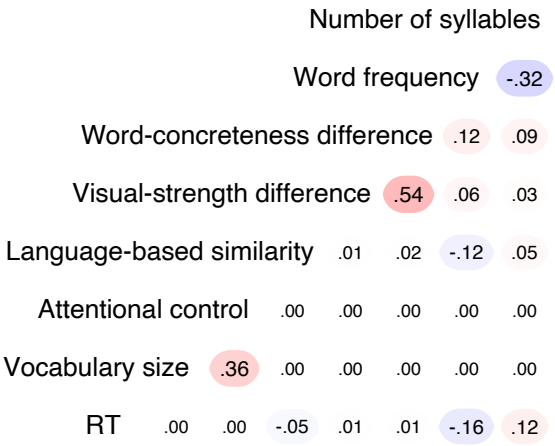

Figure 2
Zero-order correlations in the semantic priming study.

Covariates

The following covariates were included in the model to allow a rigorous analysis of the effects of interest.

- Lexical (see Appendix A): z-scored word frequency and orthographic Levenshtein distance (Balota et al., 2007)
- Semantic: z-scored word concreteness (Brysbaert et al., 2014), used as a covariate of visual rating.
- Individual differences: z-scored attentional control (Hutchison et al., 2013). This covariate is related to vocabulary size (Ratcliff et al., 2010), and previous studies have expressed the desirability of including such covariates (James et al., 2018; Pexman & Yap, 2018). Attentional control was operationalised as the average score across three tasks of Hutchison et al. (2013)—namely, operation span, Stroop and antisaccade.

Diagnostics for the frequentist model

The model presented convergence warnings. To avoid removing important random slopes, which could increase the Type I error (Brauer & Curtin, 2018; Singmann & Kellen, 2019), we examined the model after refitting it using seven optimization algorithms through the 'allFit' function of the R package 'lme4' (Bates et al., 2021). The results showed that all optimizers produced virtually identical means for all effects, suggesting that the convergence warnings were not consequential (Bates et al., 2021; see Appendix B).

The residual errors were not normally distributed, and attempts to mitigate this deviation proved unsuccessful (see Appendix B). However, this is not likely to have posed a major problem, as mixed-effects models are fairly robust to deviations from normality (Knief & Forstmeier, 2021; Schielzeth et al., 2020).

The model did not present multicollinearity problems, all variance inflation factors (VIF) being smaller than 2 (Dormann et al., 2013; Harrison et al., 2018).

Diagnostics for the Bayesian model

Three Bayesian models were run that were respectively characterised by informative, weakly-informative and diffuse priors (note that the weakly-informative prior model has not yet finished running). In each model, 16 chains were used. In each chain, 1,500 warmup iterations were run, followed by 4,500 post-warmup iterations. Thus, a total of 72,000 post-warmup draws were produced over all the chains.

The maximum $\widehat{R}$ value for the fixed effects in both models was 1.00, suggesting that these parameters had converged (Schoot et al., 2021; Vehtari et al., 2021). In contrast, the maximum $\widehat{R}$ value for the random effects was 1.13, exceeding the 1.01 threshold (Vehtari et al., 2021). Therefore, models with more iterations will be run, and once completed, they will be reported in this manuscript.

The posterior predictive checks were sound (see Appendix C). Furthermore, in the prior sensitivity analysis, the results were virtually identical with the three priors that were considered (to recall the priors, see Figure 1 above; to view the results in detail, see Appendix E).

Results

Table 1 presents the results of the frequentist model. The fixed effects explained 4.22% of the variance, and the random effects 11.01% (Nakagawa et al., 2017). It is to be expected that random effects explain more variance, as they involve a far larger number of coefficients for each effect. For instance, the by-item random slopes (specifically, by prime-target pair) for a single individual-level variable (e.g., vocabulary size) involve as many coefficients as the number of items. Conversely, the by-participant random slopes for a single item-level variable (e.g., language-based similarity) involve as many coefficients as the number of participants. In contrast, each fixed effect involves 1 coefficient only.

Table 1
Frequentist model for the semantic priming study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.00	0.00	[0.00, 0.01]	1.59	.112
Individual differences					
Attentional control	0.00	0.00	[0.00, 0.00]	-0.56	.577
Vocabulary size [a]	0.00	0.00	[0.00, 0.00]	0.02	.987
Gender [a]	0.00	0.00	[0.00, 0.00]	-0.03	.979
Target-word lexical covariates					
Word frequency	-0.16	0.00	[-0.16, -0.15]	-49.40	<.001
Number of syllables	0.07	0.00	[0.07, 0.08]	22.81	<.001
Prime-target semantic relationship					
Word-concreteness difference	0.01	0.00	[0.01, 0.02]	3.48	.001
Language-based similarity [b]	-0.08	0.00	[-0.08, -0.07]	-22.44	<.001
Visual-strength difference [b]	0.01	0.00	[0.01, 0.02]	4.18	<.001
Task condition					
Stimulus-onset asynchrony (SOA) [b]	0.06	0.01	[0.04, 0.07]	7.47	<.001
Interactions					
Word-concreteness difference × Vocabulary size	0.00	0.00	[0.00, 0.01]	1.31	.189
Word-concreteness difference × SOA	0.00	0.00	[0.00, 0.01]	2.57	.010
Word-concreteness difference × Gender	0.00	0.00	[-0.01, 0.00]	-0.97	.332
Language-based similarity × Attentional control	-0.01	0.00	[-0.01, 0.00]	-2.46	.014
Visual-strength difference × Attentional control	0.00	0.00	[0.00, 0.00]	0.24	.810
Language-based similarity × Vocabulary size	-0.01	0.00	[-0.01, 0.00]	-2.34	.020
Visual-strength difference × Vocabulary size	0.00	0.00	[-0.01, 0.00]	-1.37	.172
Language-based similarity × Gender	0.00	0.00	[-0.01, 0.00]	-0.79	.433
Visual-strength difference × Gender	0.00	0.00	[0.00, 0.01]	1.46	.144
Language-based similarity × SOA [b]	0.01	0.00	[0.00, 0.01]	3.22	.001
Visual-strength difference × SOA [b]	0.00	0.00	[-0.01, 0.00]	-2.25	.025

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Figure 3 displays the frequentist estimates alongside the Bayesian estimates. The latter are from the informative prior model. The estimates of the diffuse prior model were virtually identical to these (see Appendix E).

Figure 4-a shows the interaction between vocabulary size and language-based similarity, whereby higher-vocabulary participants presented a greater benefit from the language-based similarity between prime and target words. That is, the greater the similarity between prime and target words, the greater the advantage for participants with higher vocabularies. This interaction replicates the results of Yap et al. (2017), who analysed the same data set but using a categorical measure of similarity instead. Indeed, this replication is noteworthy as it holds in spite of some methodological differences between the studies. First, Yap et al. (2017) operationalised the priming effect as a categorical difference between related and unrelated prime-target pairs. In contrast, the present study applied a continuous measure of relatedness—i.e., cosine similarity—, which is more precise and may thus afford more statistical power (Mandera et al., 2017; Petilli et al., 2021). Second, the analysis conducted by Yap et al. (2017) was correlational, whereas the present analysis used mixed-effects models that included several covariates to measure the effects of interest as rigorously as possible.

Figure 4-b presents the interaction between vocabulary size and visual-strength difference.[6] Albeit a non-significant interaction, it is noteworthy that the effect of visual-strength difference is larger in lower-vocabulary participants.

Methodologically, the interaction between vocabulary size and language-based similarity underscores the consistency that exists between human ratings and computational approximations to meaning (e.g., Charbonnier & Wartena, 2019, 2020; Günther et al., 2016b; Louwerse et al., 2015; Mandera et al., 2017; Petilli et al., 2021; Solovyev, 2021; Wingfield & Connell, 2022). In contrast to this consistency, some studies have found human ratings to have more explanatory power than computational measures (De Deyne et al., 2016, 2019; Gagné et al., 2016; Schmidtke et al., 2018), and other research has suggested that some computational measures are not intrinsically contentful,

[6] All interaction plots across the three studies are based on the frequentist models.

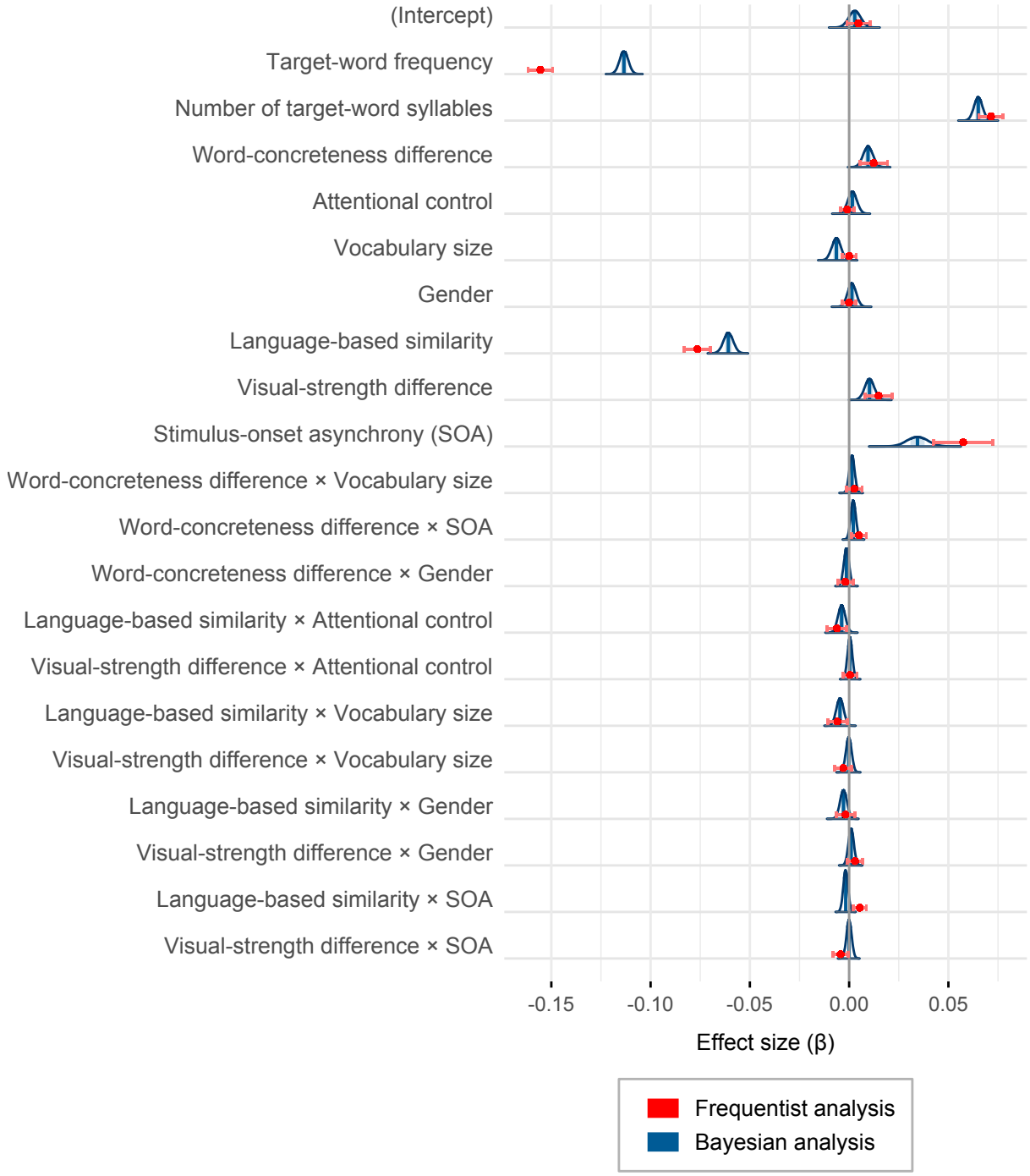

Figure 3

Estimates from the frequentist analysis (in red) and from the Bayesian analysis (in blue) for the semantic priming study. The frequentist means (represented by points) are flanked by 95% confidence intervals. The Bayesian means (represented by vertical lines) are flanked by 95% credible intervals, in light blue (in some cases, the interval is covered up by the bar of the mean).

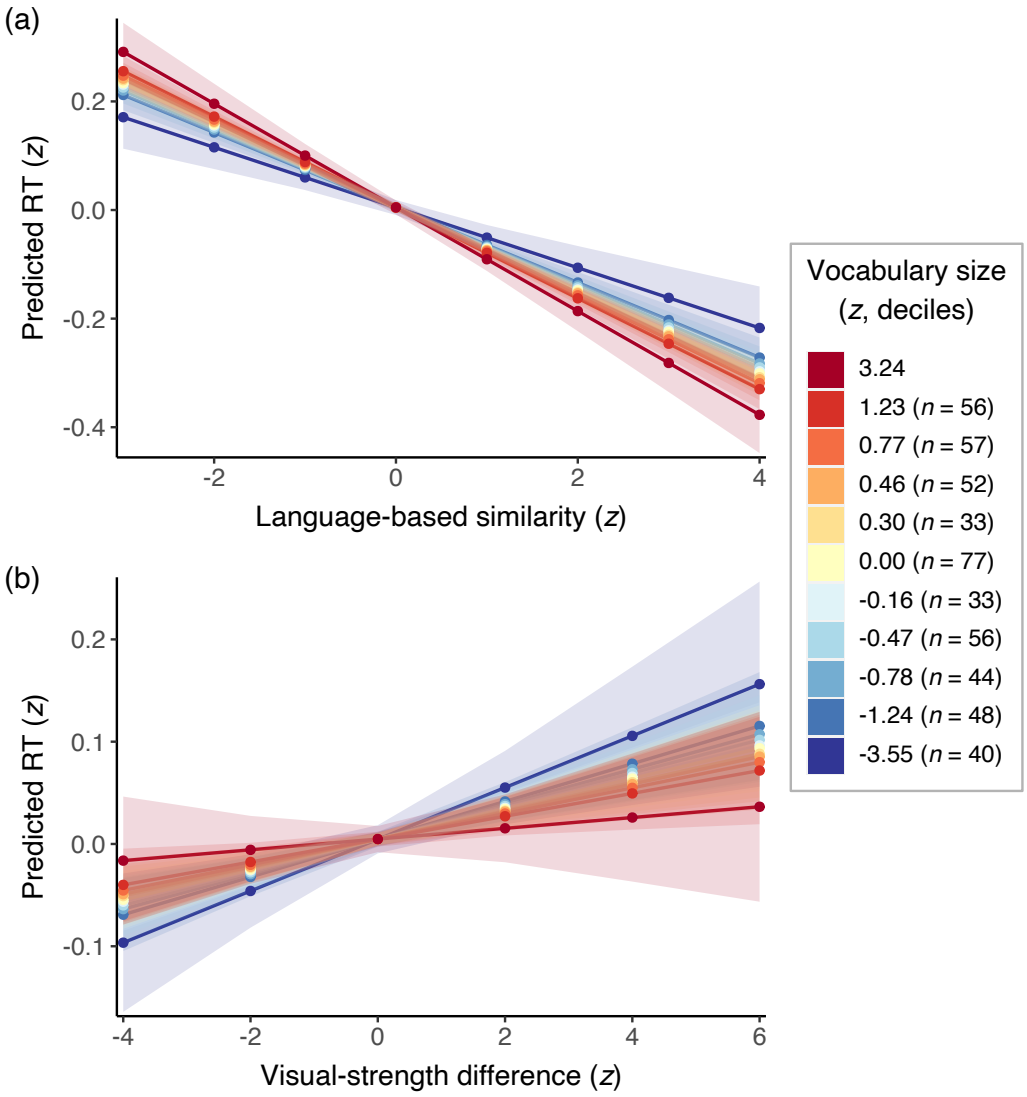

Figure 4

Interactions of vocabulary size with language-based similarity and visual-strength difference. Vocabulary size is constrained to deciles (ten sections) in this plot, whereas in the statistical analysis it contained more values within the current range. n *= number of participants contained between deciles.*

but arise from a missing-data problem (Snefjella & Blank, 2020). In sum, this evidence suggests that computational measures are artificial, and yet valid for the study of cognition.

Figure 5 shows that the effects of language-based similarity and visual strength were both larger with the short SOA. Yet there is a notable difference: whereas the effect of language-based similarity is present with both SOAs (i.e., 150 ms and 1,200 ms), the effect of visual strength is almost reserved to the the long SOA. This finding replicates Petilli et al. (2021), and stands in contrast to previous findings regarding the slower pace of the visual system in semantic priming (Lam et al., 2015) and in other paradigms (Connell & Lynott, 2013; Louwerse & Connell, 2011). **conclude**_____

Figure 6 shows the interactions of gender with language-based similarity and visual-strength difference, albeit non-significant.[7]

The importance of outliers. The interaction shown in Figure 4 was patent in all deciles of vocabulary size but most notably in the participants who lie more than 1 standard deviation away from the mean. Outliers in individual differences have played important roles in other areas of cognition as well, such as in the domains of aphantasia and hyperphantasia—traits characterised, respectively, by a diminished and an extraordinary ability to mentally visualise objects (Milton et al., 2021; Zeman et al., 2020). Furthermore, if we consider the difficulty of detecting effects involving individual differences (Diaz et al., 2021; Hedge et al., 2018; Rodríguez-Ferreiro et al., 2020), and the limited representativeness of most samples of participants (Henrich et al., 2010), it may be fruitful to study more varied samples, where possible.

Comparing two measures of vision-based information

Next, we reflected on the adequacy of visual-strength difference as a measurement instrument, as it had never been used before to operationalise the semantic priming effect. On the one hand, its effect on RT was negative, as we would expect. However, we were concerned about the low correlation between this variable and language-based similarity (r = 0.01). This concern was especially motivated by the correlation that Petilli et al. (2021)

[7] Plots of other interactions are available in Appendix D.

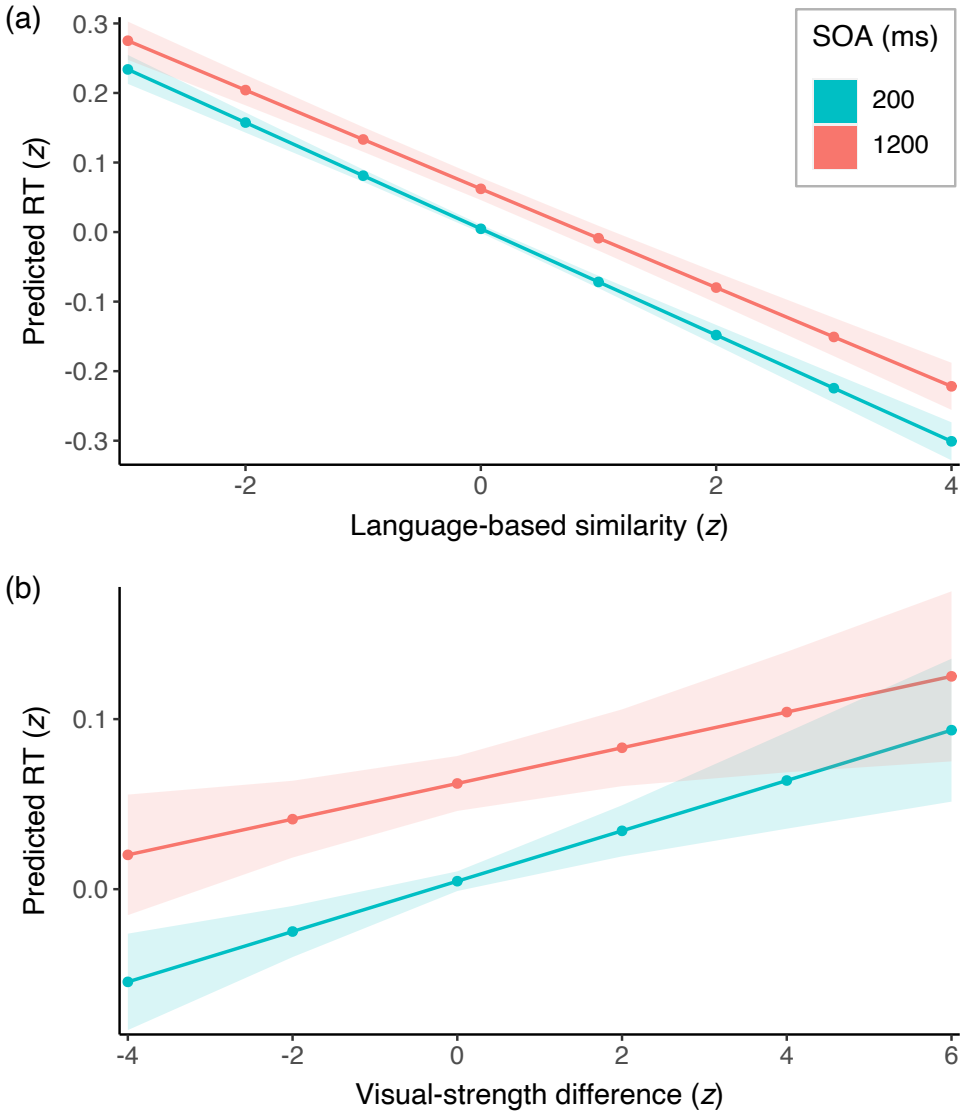

Figure 5

Interactions of stimulus-onset asynchrony (SOA) with language-based similarity and visual strength. SOA was analysed using z-scores, but for clarity, the variable is shown in its basic form here.

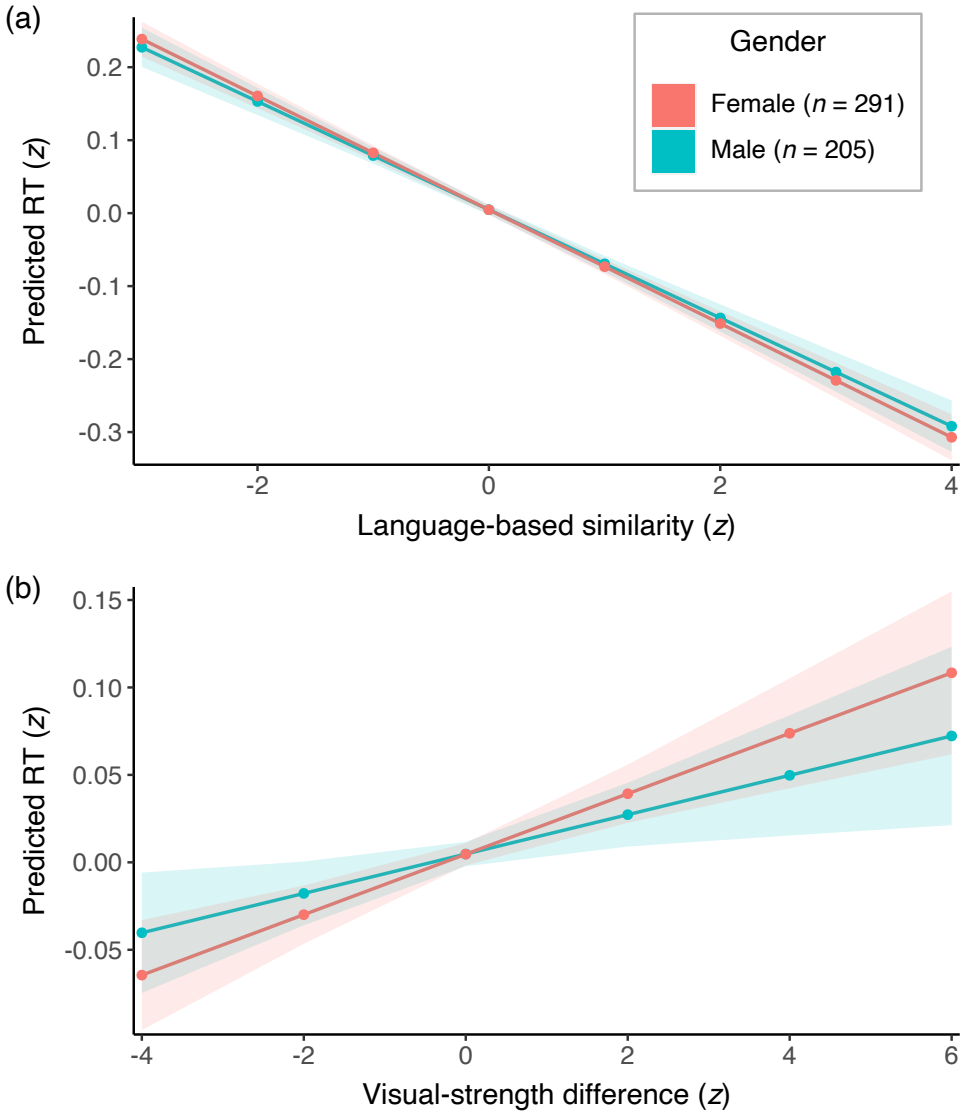

Figure 6

Interactions with gender in the semantic priming study. Gender was analysed using z-scores, but for clarity, the variable is shown in its basic form here.

had found between their 'vision-based similarity' measure and language-based similarity, which reached $r = .50$. Based on this information, we set out to run a model with the aim of comparing the performance of our measures of vision-based information.

We created a subset of our data set in which we ensured that all the trials contained data from all the relevant variables—i.e., from all the existing variables and from the newly-added visual similarity from Petilli et al. This subsetting resulted in the loss of 83% of trials, owing to the strict selection criteria applied by Petilli et al. in the creation of the visual-similarity measure (e.g., both the target and the prime word had to be associated to at least 100 pictures in ImageNet). As a result, our data set in this analysis contained 496 participants, 1,091 prime-target pairs, and 254,140 RTs. On average, there were 128 prime-target pairs per participant ($SD = 10.37$), and conversely, 58 participants per prime-target pair ($SD = 4.90$).

Figure 7 shows the zero-order correlations among the predictors and the dependent variable.

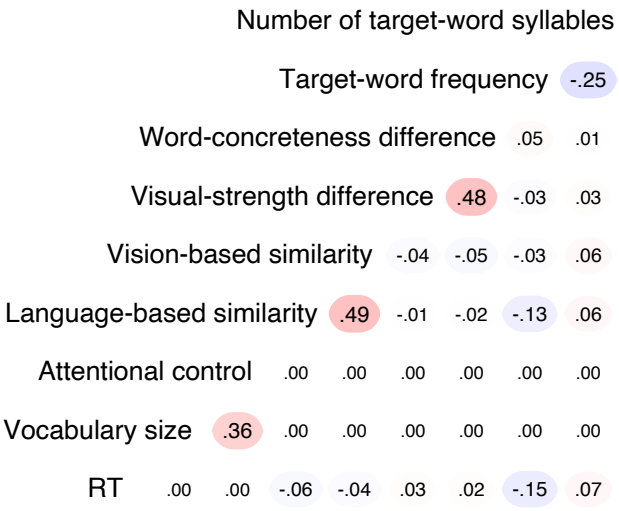

Figure 7
Zero-order correlations in the semantic priming data set that included visual similarity.

Diagnostics for the frequentist model

The model presented convergence warnings. To avoid removing important random slopes, which could increase the Type I error (Brauer & Curtin, 2018; Singmann & Kellen,

2019), we examined the model after refitting it using seven optimization algorithms through the 'allFit' function of the 'lme4' package (Bates et al., 2021). The results showed that all optimizers produced virtually identical means for all effects, suggesting that the convergence warnings were not consequential (Bates et al., 2021; see Appendix B).

The residual errors were not normally distributed, and attempts to mitigate this deviation proved unsuccessful (see Appendix B). However, this is not likely to have posed a major problem, as mixed-effects models are fairly robust to deviations from normality (Knief & Forstmeier, 2021; Schielzeth et al., 2020).

The model did not present multicollinearity problems, all VIFs being smaller than 2 (Dormann et al., 2013; Harrison et al., 2018).

Due to time constraints, this analysis did not include a Bayesian model.

Results. Table 2 presents the results of the frequentist model. Due to space, the covariates are shown in Table 3. The fixed effects explained 3.53% of the variance, and the random effects 18.47% (Nakagawa et al., 2017; for an explanation of this difference, see above). Figure 8 displays the estimates for the effects of interest.

Table 2

Effects of interest in the semantic priming model that included visual similarity.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.01	0.01	[-0.01, 0.02]	0.90	.370
Individual differences					
Vocabulary size [a]	0.00	0.00	[-0.01, 0.01]	0.21	.834
Gender [a]	0.00	0.00	[-0.01, 0.01]	-0.05	.962
Prime-target semantic relationship					
Language-based similarity [b]	-0.07	0.01	[-0.09, -0.06]	-8.33	<.001
Visual-strength difference [b]	0.03	0.01	[0.01, 0.04]	3.04	.002
Vision-based similarity [b]	-0.02	0.01	[-0.04, -0.01]	-2.55	.011
Task condition					
Stimulus-onset asynchrony (SOA) [b]	0.06	0.01	[0.04, 0.08]	6.80	<.001
Interactions					
Language-based similarity × Vocabulary size	-0.01	0.01	[-0.02, 0.01]	-0.96	.339
Visual-strength difference × Vocabulary size	-0.01	0.01	[-0.02, 0.01]	-1.02	.309
Vision-based similarity × Vocabulary size	0.00	0.01	[-0.01, 0.01]	-0.01	.991
Language-based similarity × Gender	0.00	0.01	[-0.02, 0.01]	-0.75	.456
Visual-strength difference × Gender	-0.01	0.01	[-0.02, 0.01]	-1.05	.294
Vision-based similarity × Gender	0.00	0.01	[-0.01, 0.01]	0.39	.696
Language-based similarity × SOA [b]	0.00	0.00	[0.00, 0.01]	0.87	.382
Visual-strength difference × SOA [b]	-0.01	0.00	[-0.02, 0.00]	-2.60	.010
Vision-based similarity × SOA [b]	0.01	0.00	[0.00, 0.01]	1.28	.201

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. Covariates shown in next table due to space. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.

[b] By-participant random slopes were included for this effect.

Table 3
Covariates in the semantic priming model that included visual similarity.

	β	*SE*	95% CI	*t*	*p*
Individual-differences covariate					
Attentional control	0.00	0.00	[-0.01, 0.00]	-1.06	.288
Target-word lexical covariates					
Word frequency	-0.15	0.01	[-0.17, -0.14]	-21.97	<.001
Number of syllables	0.02	0.01	[0.01, 0.04]	3.54	<.001
Prime-target semantic covariate					
Word-concreteness difference	0.02	0.01	[0.01, 0.04]	2.73	.006
Covariate interactions					
Word-concreteness difference × Vocabulary size	-0.01	0.00	[-0.01, 0.00]	-1.15	.252
Word-concreteness difference × SOA	0.01	0.00	[0.01, 0.02]	5.64	<.001
Word-concreteness difference × Gender	0.01	0.00	[0.00, 0.01]	1.34	.179
Language-based similarity × Attentional control	0.00	0.00	[-0.01, 0.01]	-0.91	.362
Visual-strength difference × Attentional control	0.00	0.00	[-0.01, 0.01]	0.71	.477
Vision-based similarity × Attentional control	0.00	0.00	[-0.01, 0.01]	0.80	.423

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. Some interactions are split over two lines, with the second line indented.

The results revealed that visual-strength difference had a significantly larger effect. This difference was not due to an excessive collinearity between these measures (r = -0.04). Also importantly, both measures appeared to be valid based on their correlations with language-based similarity and with word concreteness.

Statistical power analysis

Power curves were performed for most effects of interest in the main model (no power analysis was performed for the model that included visual similarity). Figures 9 and 10 show the estimated power for the main effects and the interactions, respectively.

_____ interpret _____

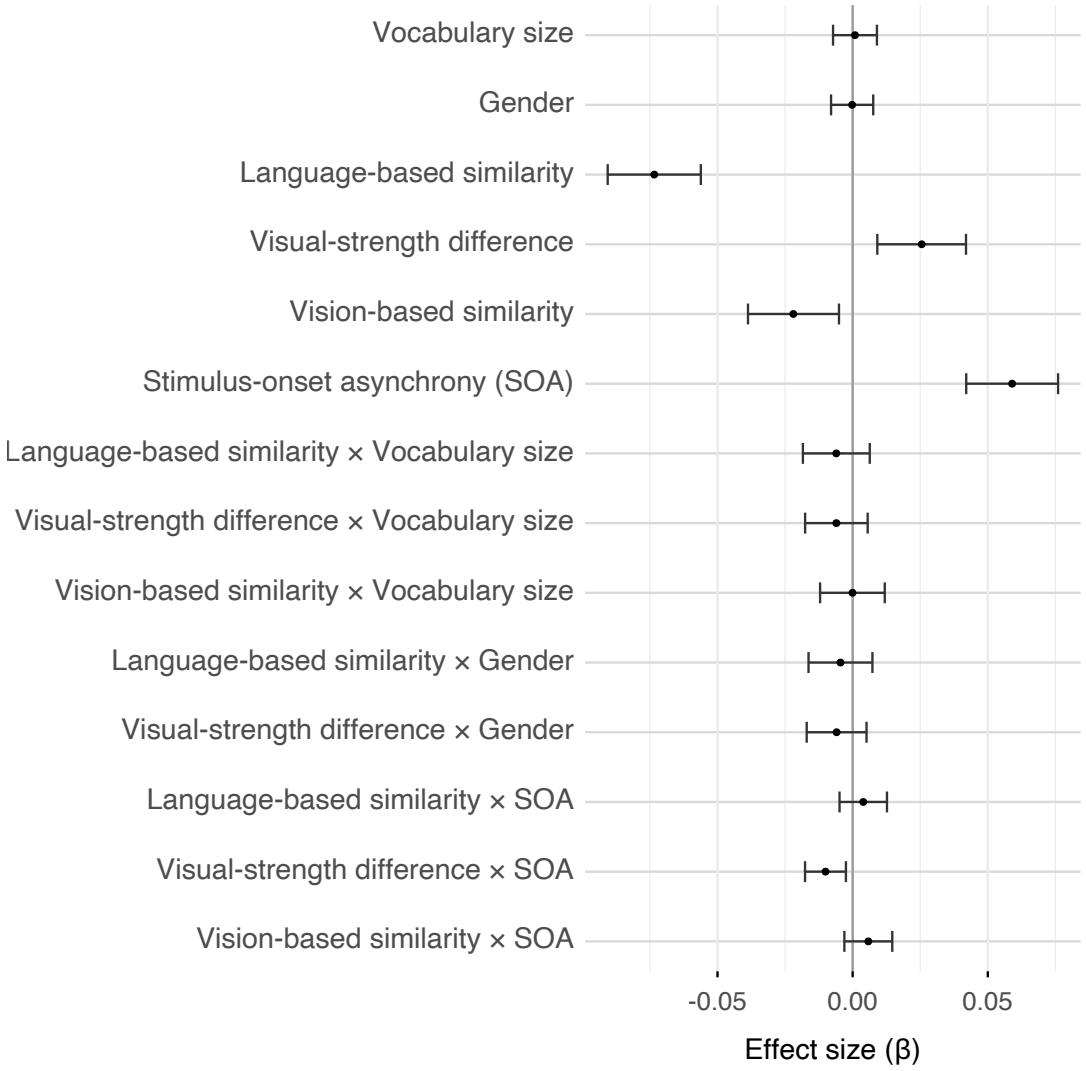

Figure 8

Means and 95% confidence intervals for the effects of interest in the semantic priming model that included visual similarity.

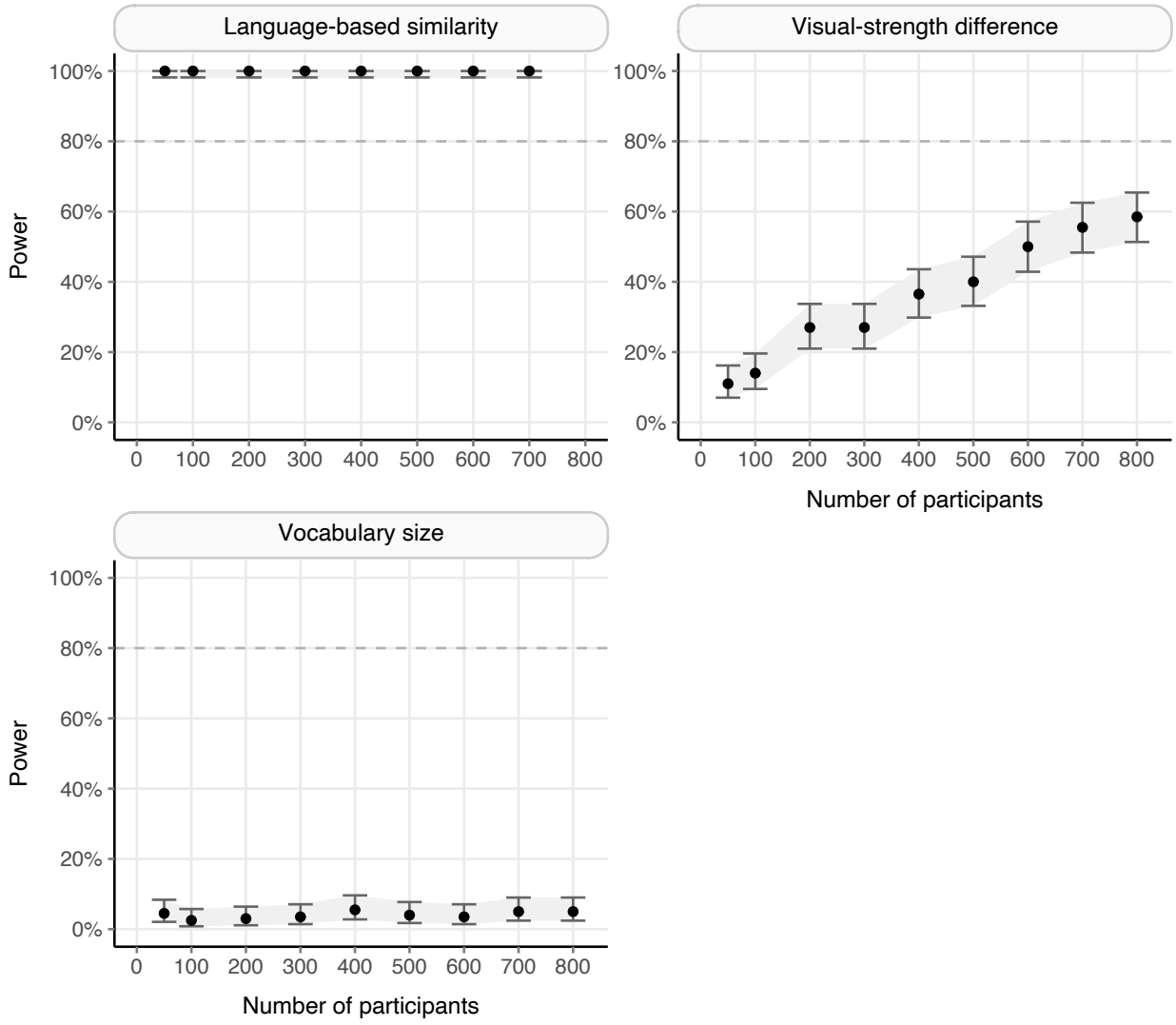

Figure 9

Power curves for some main effects in the semantic priming study.

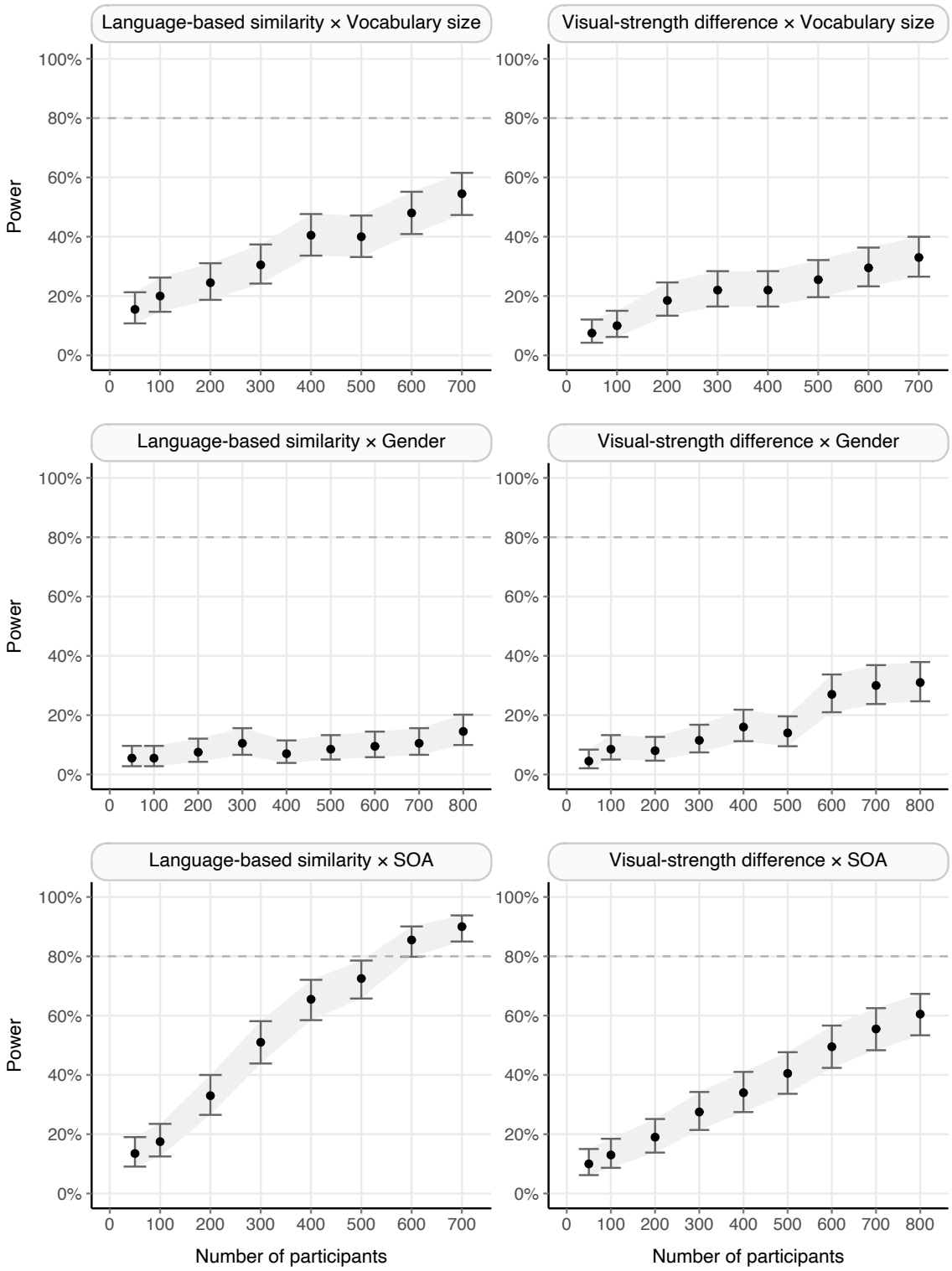

Figure 10

Power curves for some interactions in the semantic priming study.

Discussion

Study 2: Semantic decision

The semantic decision task probes into the role of concreteness in conceptual processing. Specifically, this task requires participants to classify words as abstract or concrete. Researchers then analyse whether the responses can be explained by the sensory experientiality of the referents—that is, the degree to which they can be experienced through our senses—and by other variables, such as word frequency. The core data set in this study was that of the Calgary Semantic Decision Project (Pexman et al., 2017; Pexman & Yap, 2018). The experimental task is semantic decision, in which participants judge whether words are relatively concrete (e.g., 'building') or abstract (e.g., 'thought').

Research has found that the processing of relatively concrete words relies considerably on sensorimotor information (Hultén et al., 2021; Kousta et al., 2011; Vigliocco et al., 2014). In contrast, the processing of relatively abstract words seems to draw more heavily on information from language (Barca et al., 2020; Duñabeitia et al., 2009; Snefjella & Blank, 2020), emotion (Kousta et al., 2011; Ponari et al., 2020; Ponari, Norbury, Rotaru, et al., 2018; Ponari, Norbury, & Vigliocco, 2018; Vigliocco et al., 2014) and sociality (Borghi et al., 2019; Diveica et al., 2022).

Word co-occurrence

Wingfield and Connell (2022) reanalysed the data from Pexman et al. (2017) using language-based variables that are more related to the language system than to the visual system. The task used by Pexman et al. had been semantic decision, in which participants assessed whether words were abstract or concrete. Wingfield and Connell found that the variables that best explained RTs were word co-occurrence measures. Specifically, one of these variables was the correlation distance between each stimulus word and the word 'abstract'. The other variable was the correlation distance between each stimulus word and the word 'concrete'. Wingfield and Connell studied these distance measures in various forms, and found that cosine and correlation distance yielded the best results. We used the correlation distance, following the advice of Kiela and Bottou (2014) (see details below).

Methods

Effects of interest

Vocabulary size. [`z_vocabulary_size`; calculated from `NAART` in Pexman et al. (2017)]. In the test used by Pexman and Yap (2018), participants were presented with 35 rare words, whose pronunciations are not regular (e.g., *gaoled*, *ennui*), and they were asked to read the words aloud. When they pronounced a word correctly, it was inferred that they knew the word. This test was based on NAART35, a short version of the North American Adult Reading Test (Uttl, 2002).

Participant's gender. [`z_recoded_participant_gender`; calculated from `Gender` in Pexman and Yap (2018)]

Word co-occurrence. The zero-order correlation between Wingfield and Connell's (2022) distance to 'abstract' and distance to 'concrete' was $r = 0.98$. To avoid the collinearity between these variables in the model (Dormann et al., 2013; Harrison et al., 2018), and to facilitate the analysis of interactions with other variables, we created a difference score by subtracting the distance to 'abstract' from the distance to 'concrete'. This new variable was named 'word co-occurrence'. As Figure 11 demonstrates, the difference score captured the explanatory power of the two original variables, resulting in an increase of the correlation with word concreteness.

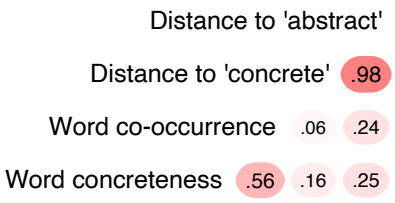

Figure 11
Zero-order correlations among Wingfield and Connell's (2022) distances, the difference score (word co-occurrence) and word concreteness (Brysbaert et al., 2014).

Visual strength. [`z_visual_rating`; calculated from `Visual.mean` in Lynott et al. (2020)]

The final data set contained 306 participants, 8,927 words, and 246,432 RTs. On

average, there were 755 words per participant ($SD = 42.05$), and conversely, 26 participants per word ($SD = 4.80$).

Figure 12 shows the zero-order correlations among the predictors among the predictors and the dependent variable.

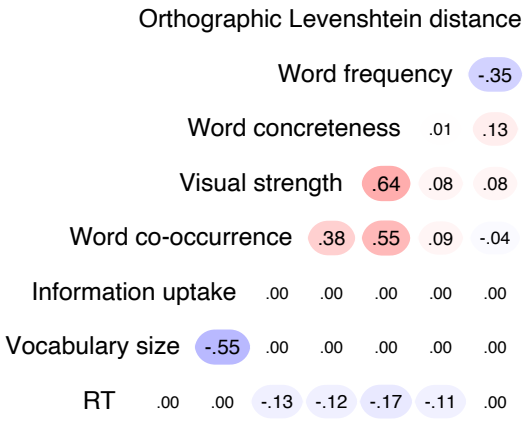

Figure 12
Zero-order correlations in the semantic decision study.

Covariates

The following covariates were included in the model to allow a rigorous analysis of the effects of interest.

- Lexical (see Appendix A): z-scored word frequency and orthographic Levenshtein distance (Balota et al., 2007)
- Semantic: z-scored word concreteness (Brysbaert et al., 2014). This variable is used as a covariate of visual strength, and it is especially fundamental in the semantic decision task, in which participants judge whether words are concrete or abstract (for a more general consideration of concreteness, see Bottini et al., 2021). Indeed, owing to the instructions of the task, word concreteness is likely to be more relevant to the participants' task than our effects of interest.
- Individual differences: z-scored information uptake (Pexman & Yap, 2018). This covariate is related to vocabulary size (Ratcliff et al., 2010; also see James et al.,

2018; Pexman & Yap, 2018).

Diagnostics for the frequentist model

The model presented convergence warnings. To avoid removing important random slopes, which could increase the Type I error (Brauer & Curtin, 2018; Singmann & Kellen, 2019), we examined the model after refitting it using seven optimization algorithms through the 'allFit' function of the 'lme4' package (Bates et al., 2021). The results showed that all optimizers produced virtually identical means for all effects, suggesting that the convergence warnings were not consequential (Bates et al., 2021; see Appendix B).

The residual errors were not normally distributed, and attempts to mitigate this deviation proved unsuccessful (see Appendix B). However, this is not likely to have posed a major problem, as mixed-effects models are fairly robust to deviations from normality (Knief & Forstmeier, 2021; Schielzeth et al., 2020).

The model did not present multicollinearity problems, all VIFs being smaller than 2 (Dormann et al., 2013; Harrison et al., 2018).

Diagnostics for the Bayesian model

Three Bayesian models were run that were respectively characterised by informative, weakly-informative and diffuse priors. In each model, 16 chains were used. In each chain, 2,000 warmup iterations were run, followed by 6,000 post-warmup iterations. Thus, a total of 96,000 post-warmup draws were produced over all the chains.

The maximum $\widehat{R}$ value for the fixed effects across the three models was 1.42, far exceeding the 1.01 threshold (Vehtari et al., 2021; also see Schoot et al., 2021). Similarly, the maximum $\widehat{R}$ value for the random effects was 1.31. Furthermore, the posterior predictive checks revealed major divergences between the observed data and the predicted outcome (see Appendix C). Since the results were not valid, they are not shown in the main text but in Appendix E.

Results

Table 4 presents the results of the frequentist model. The fixed effects explained 4.11% of the variance, and the random effects 17.48% [Nakagawa et al. (2017); for an explanation of this difference, see Results of Study 1). Figure 13 displays these estimates.[8]

Table 4
Frequentist model for the semantic decision study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.05	0.00	[0.04, 0.06]	11.87	<.001
Individual differences					
Information uptake	0.00	0.00	[0.00, 0.00]	0.20	.844
Vocabulary size [a]	0.00	0.00	[-0.01, 0.00]	-1.42	.155
Gender [a]	0.00	0.00	[0.00, 0.00]	-0.47	.636
Lexical covariates					
Word frequency	-0.12	0.00	[-0.13, -0.12]	-28.63	<.001
Orthographic Levenshtein distance	-0.01	0.00	[-0.02, 0.00]	-3.05	.002
Semantic variables					
Word concreteness	-0.13	0.01	[-0.14, -0.11]	-21.39	<.001
Word co-occurrence [b]	-0.03	0.01	[-0.04, -0.02]	-4.48	<.001
Visual strength [b]	-0.02	0.01	[-0.03, -0.01]	-2.91	.004
Interactions					
Word concreteness × Vocabulary size	-0.02	0.00	[-0.03, -0.02]	-7.66	<.001
Word concreteness × Gender	-0.01	0.00	[-0.02, 0.00]	-3.50	<.001
Word co-occurrence × Information uptake	0.01	0.01	[0.00, 0.02]	1.48	.141
Visual strength × Information uptake	0.02	0.01	[0.01, 0.03]	3.05	.003
Word co-occurrence × Vocabulary size	0.01	0.01	[0.00, 0.02]	1.66	.098
Visual strength × Vocabulary size	0.01	0.01	[0.00, 0.02]	2.03	.043
Word co-occurrence × Gender	0.00	0.00	[-0.01, 0.01]	0.86	.393
Visual strength × Gender	0.00	0.00	[-0.01, 0.01]	-0.08	.940

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

[8] Bayesian estimates not shown as they were not valid. They are nonetheless available in Appendix E.

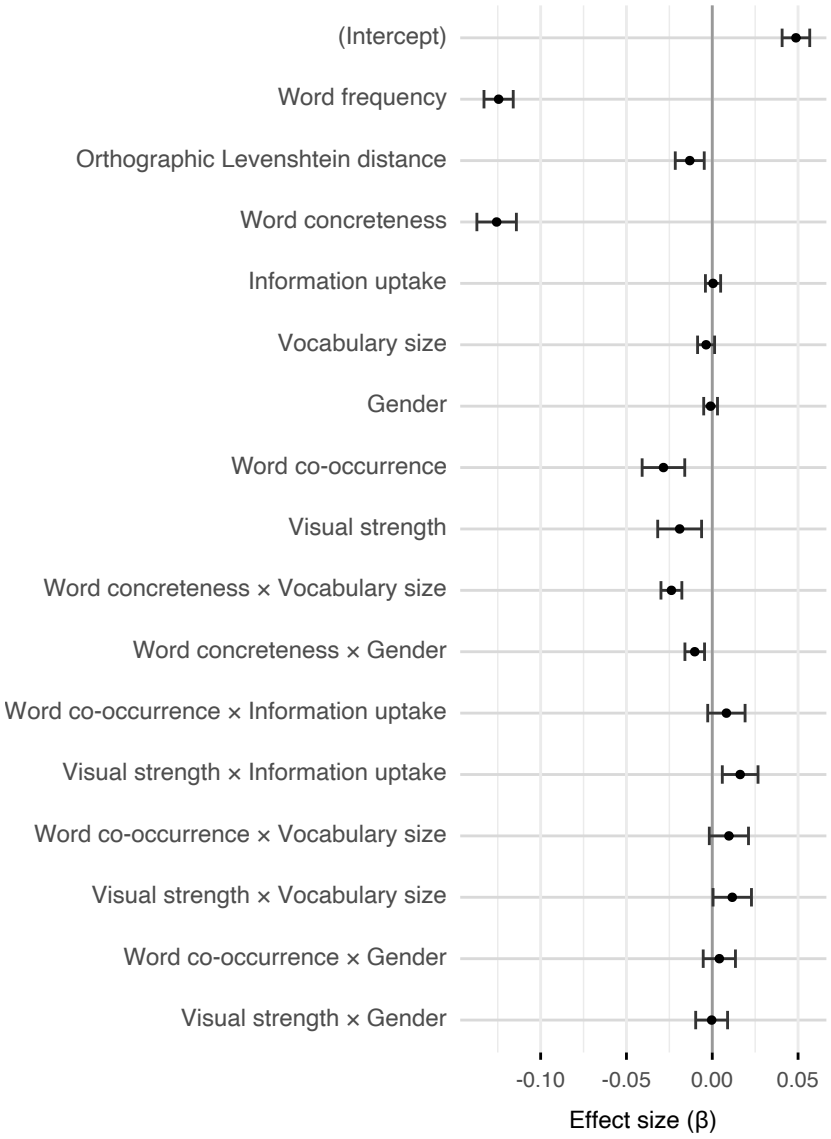

Figure 13

Means and 95% confidence intervals for the effects of interest in the semantic decision study.

Figures 14-a and 14-b demonstrate how lower-vocabulary participants (compared to higher-vocabulary participants) were more strongly influenced by both language- and vision-based similarity. In contrast, Figure 14-c shows that higher-vocabulary participants benefitted more strongly from word concreteness. Since the variable that is most relevant to the semantic decision task is word concreteness, the present interactions suggest that higher-vocabulary participants were better able to focus on the most relevant information, whereas lower-vocabulary participants were sensitive to a greater breadth of information (Lim et al., 2020; Pexman & Yap, 2018; Yap et al., 2009, 2012, 2017).

The present analysis used a continuous measure of word concreteness. In contrast, Pexman and Yap (2018) analysed the same data set after dividing the stimulus words into abstract and concrete subsets, which they analysed separately. Pexman and Yap found that high-vocabulary participants were more sensitive to the relative abstractness of words. Specifically, these participants were faster to classify very abstract words than mid-abstract ones, thus presenting a reverse concreteness effect. Such a reverse effect might stem from the bimodal distributions of concreteness ratings (Brysbaert et al., 2014) and semantic decisions (Pexman & Yap, 2018). The reverse effect has also been found in semantic dementia patients (Connell & Lynott, 2012). Since the reverse effect contrasted with the long-established concreteness effect, the former has sometimes been regarded as an inconsistency.

Notwithstanding the aforementioned bimodal distributions, Troche et al. (2017) suggested that a continuous analysis remains necessary to study word concreteness (also see Cohen, 1983). Consistent with this, our present findings demonstrated the sensitivity of a continuous word concreteness variable to patterns such as the greater role of task-relevant variables in high-vocabulary participants. In conclusion, the literature and our findings suggest that the split-data approach and the continuous approach to word concreteness are both useful. Where feasible, the application of both approaches would provide the most information.

Figure 15 shows the interactions of gender with language-based similarity and

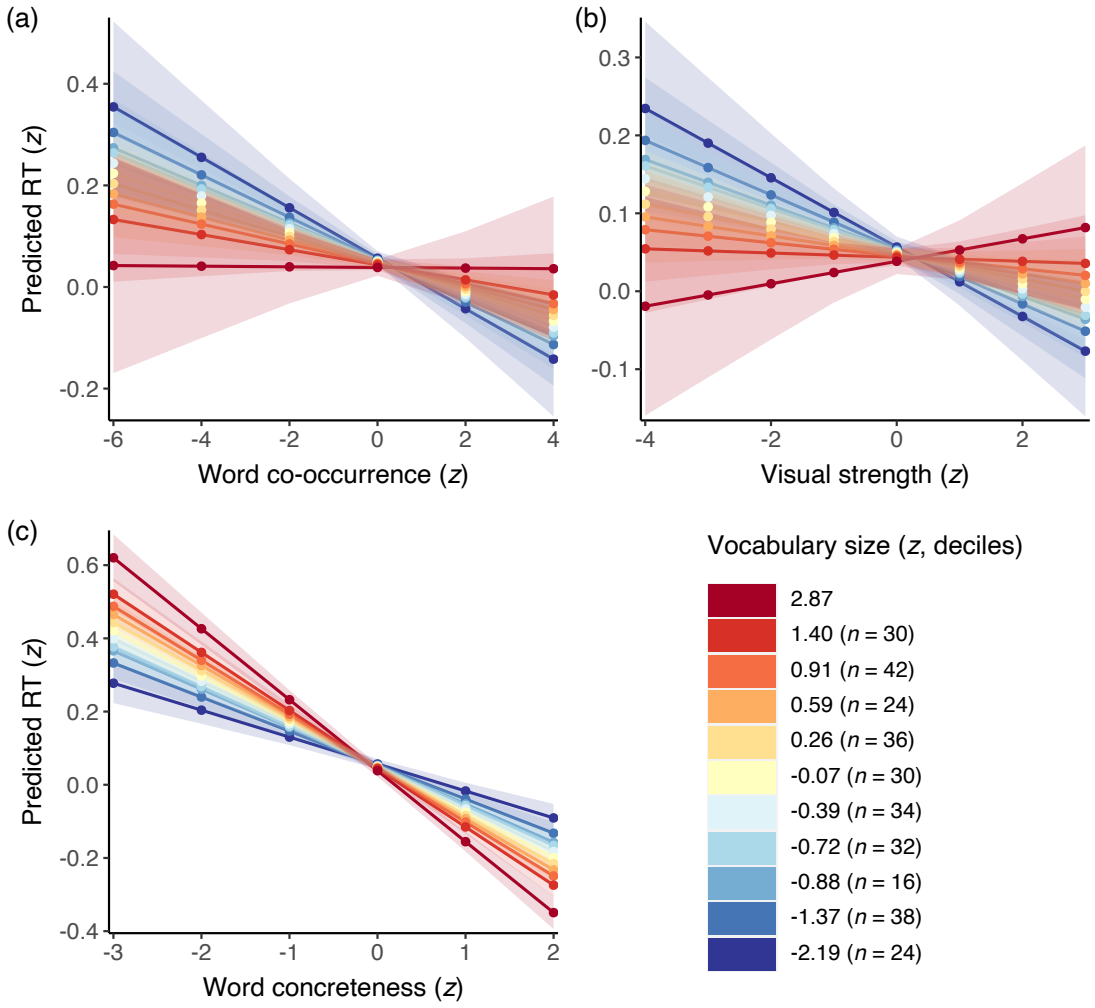

Figure 14

Interactions of vocabulary size with (a) language-based information, (b) visual strength and (c) word concreteness. Vocabulary size is constrained to deciles in this plot, whereas in the statistical analysis it contained more values within the current range. n *= number of participants contained between deciles.*

visual-strength difference, albeit non-significant.[9]

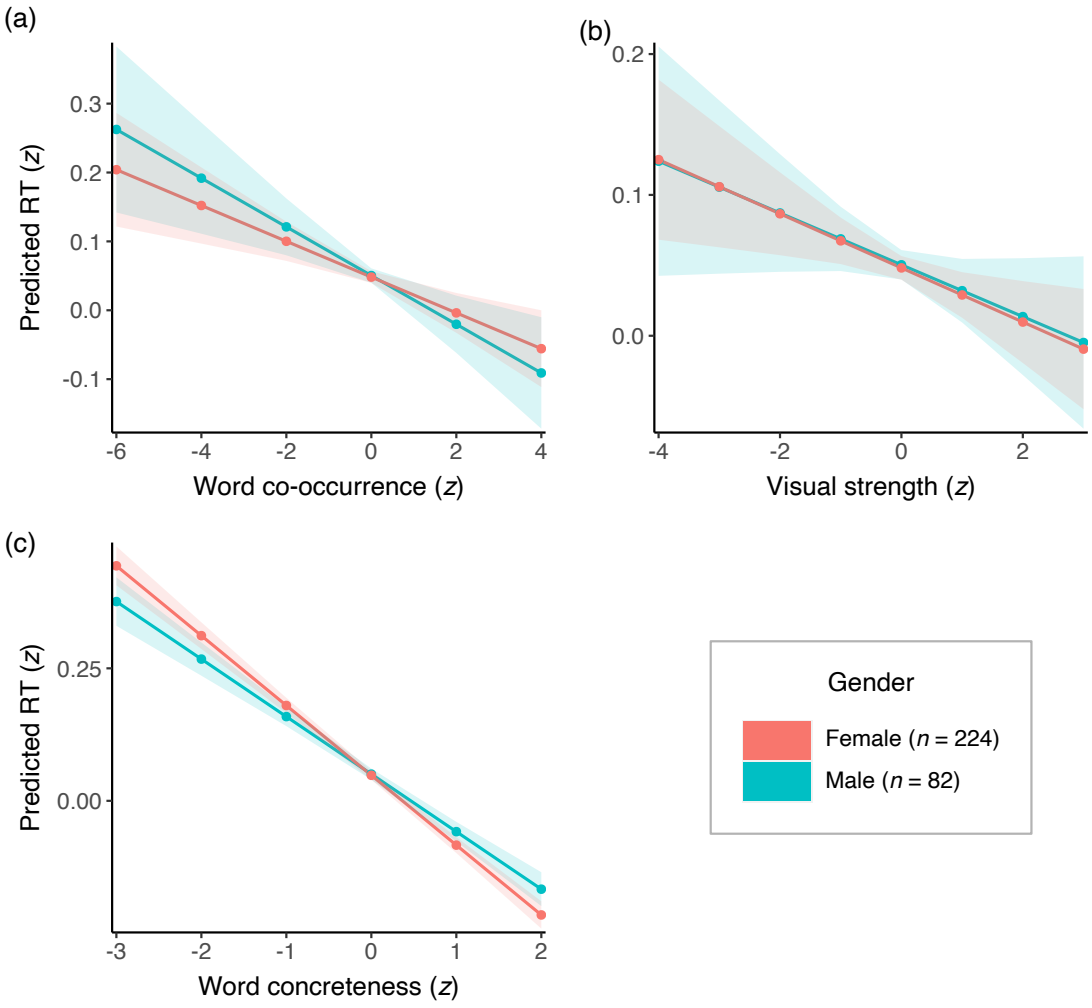

Figure 15

Interactions with gender in the semantic decision study. Gender was analysed using z-scores, but for clarity, the variable is shown in its basic form here.

Statistical power analysis

Figures 16 and 17 show the estimated power available for main effects and interactions of interest, respectively.

_____ interpret _____

[9] Plots of other interactions are available in Appendix D.

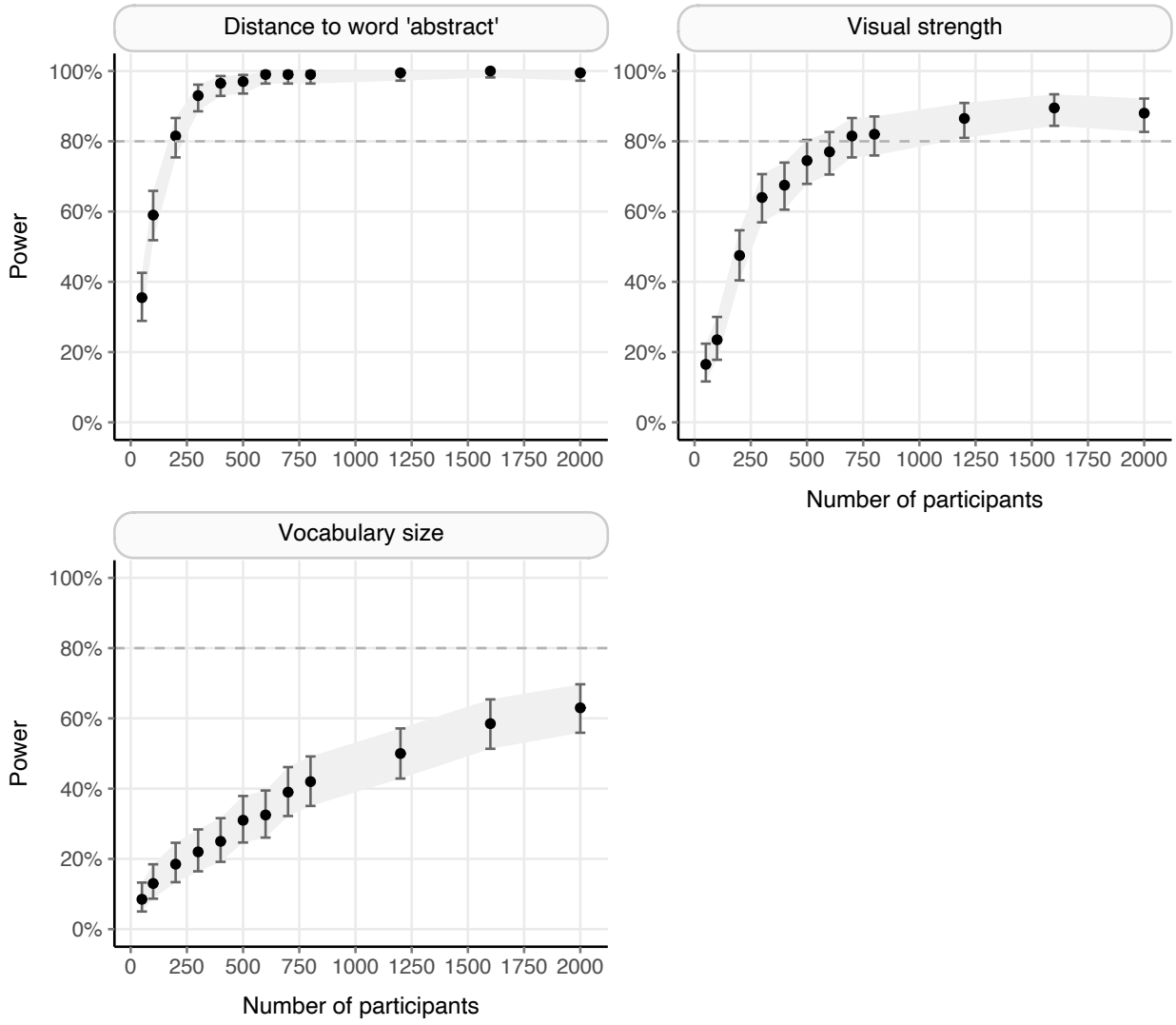

Figure 16

Power curves for some main effects in the semantic decision study.

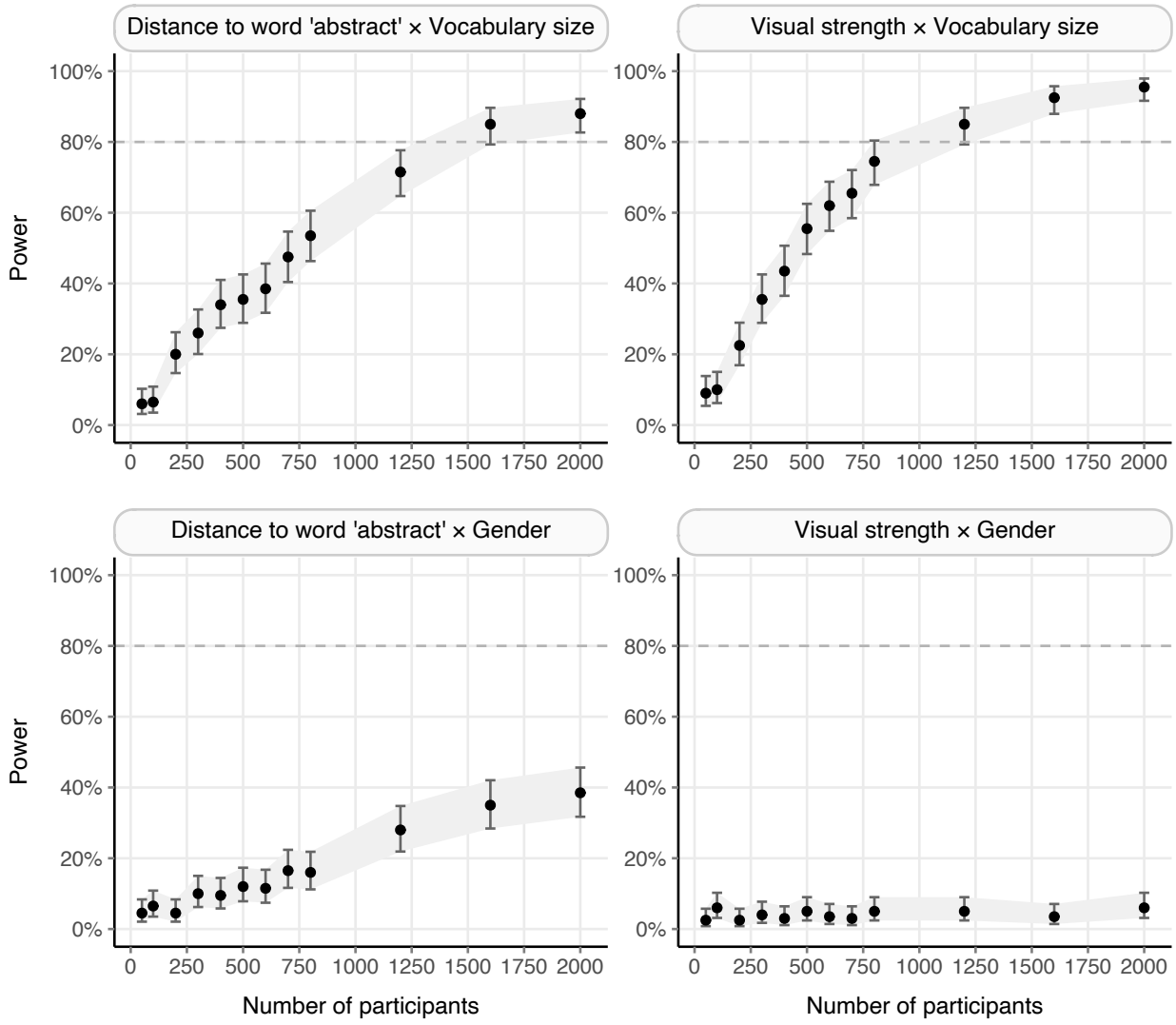

Figure 17

Power curves for some interactions in the semantic decision study.

Discussion

Study 3: Lexical decision

The core data set in this study was the lexical decision subset of the English Lexicon Project—ELP (Balota et al., 2007). The lexical decision task differs from semantic priming and semantic decision in two important aspects.

1. **Less semantic processing.** The lexical decision paradigm is likely to involve less semantic processing than the other paradigms. _____

2. **Single-word measures.** Compared to semantic priming and semantic decision, it is more difficult in the lexical decision paradigm to create word-to-word distance measures to capture language-based and vision-based information. The possibility of calculating the distance between words in consecutive trials is hindered by the need to skip trials, owing to the high prevalence of nonword trials in the lexical decision paradigm. Therefore, the measures must be based on each word alone. Accordingly, vision-based information can be operationalised as the visual strength of each word. Language-based information could be operationalised as any of several lexical variables. In the present study, word frequency was chosen as it had the most explanatory power out of 5 candidates—the other candidates being number of letters, number of syllables, orthographic Levenshtein distance and phonological Levenshtein distance (see Appendix A). It should also be noted that word frequency has been found to be more closely related to semantic variables than to lexical ones, such as word length, orthography, phonology (see Table 4 in Yap et al., 2012).

Word frequency is also of interest due to its varying relationship to vocabulary size across different paradigms. Yap et al. (2012) found that higher-vocabulary participants in the ELP were more strongly influenced by word frequency than lower-vocabulary participants. The same finding also appeared in a lexical decision study in Chinese (Lim et al., 2020). In contrast, deeper semantic tasks, such as semantic priming (Yap et al., 2017) and semantic decision (Pexman & Yap, 2018), have yielded the opposite pattern, with the effect of word frequency decreasing in higher-vocabulary participants.

Effects of interest

- Z-scored vocabulary age [`z_vocabulary_age`; calculated from `vocabAge` in Balota et al. (2007)]
- Z-scored, recoded participant's gender [`z_recoded_participant_gender`; calculated from `Gender` in Balota et al. (2007)]
- Z-scored word frequency [`z_word_frequency`; calculated from `LgSUBTLWF` in Balota et al. (2007)]
- Z-scored vision-based information in words [`z_visual_rating`; calculated from `Visual.mean` in Lynott et al. (2020)]

The final data set contained 795 participants, 12,636 words, and 19,828 RTs. On average, there were 25 words per participant ($SD = 36.04$), and conversely, 2 participants per word ($SD = 0.86$).

Figure 18 shows the zero-order correlations among the predictors and the dependent variable.

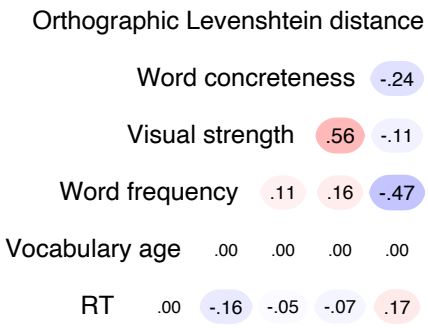

Figure 18
Zero-order correlations in the lexical decision study.

Covariates

The following lexical covariates were included in the model to allow a rigorous analysis of the effects of interest.

- Lexical (see Appendix A): Z-scored orthographic Levenshtein distance (Balota et al., 2007)
- Semantic: z-scored word concreteness (Brysbaert et al., 2014), used as a covariate of visual rating.

Diagnostics for the frequentist model

The model presented convergence warnings. To avoid removing important random slopes, which could increase the Type I error (Brauer & Curtin, 2018; Singmann & Kellen, 2019), we examined the model after refitting it using seven optimization algorithms through the 'allFit' function of the 'lme4' package (Bates et al., 2021). The results showed that all optimizers produced virtually identical means for all effects, suggesting that the convergence warnings were not consequential (Bates et al., 2021; see Appendix B).

The residual errors were not normally distributed, and attempts to mitigate this deviation proved unsuccessful (see Appendix B). However, this is not likely to have posed a major problem, as mixed-effects models are fairly robust to deviations from normality (Knief & Forstmeier, 2021; Schielzeth et al., 2020).

The model did not present multicollinearity problems, all VIFs being smaller than 2 (Dormann et al., 2013; Harrison et al., 2018).

Diagnostics for the Bayesian model

Three Bayesian models were run that were respectively characterised by informative, weakly-informative and diffuse priors. In each model, 5 chains were used. In each chain, 2,000 warmup iterations were run, followed by 18,000 post-warmup iterations. Thus, a total of 90,000 post-warmup draws were produced over all the chains.

The maximum $\widehat{R}$ value for the fixed effects across the three models was 1.00, suggesting that these effects converged (Schoot et al., 2021; Vehtari et al., 2021). For the random effects, the maximum $\widehat{R}$ value was 1.02, barely exceeding the 1.01 threshold (Vehtari et al., 2021).

The posterior predictive checks were sound (see Appendix C). Furthermore, in the prior sensitivity analysis, the results were virtually identical with the three priors that were considered (to recall the priors, see Figure 1 above; to view the results in detail, see Appendix E).

Results

Table 5 presents the results of the frequentist model. The fixed effects explained 5.61% of the variance, and the random effects 10.25% [Nakagawa et al. (2017); for an explanation of this difference, see Results of Study 1). Figure 19 displays the frequentist estimates alongside the Bayesian estimates. The latter are from the weakly-informative prior model. The estimates of the two other models (i.e., with informative and diffuse priors, respectively) were virtually identical to these (see Appendix E).

Table 5
Frequentist model for the lexical decision study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.00	0.01	[-0.01, 0.01]	-0.02	.983
Individual differences					
Vocabulary age [a]	0.00	0.01	[-0.01, 0.01]	-0.06	.950
Gender [a]	0.00	0.01	[-0.01, 0.01]	0.01	.995
Lexical covariate					
Orthographic Levenshtein distance	0.11	0.01	[0.09, 0.12]	13.41	<.001
Semantic variables					
Word concreteness	-0.02	0.01	[-0.04, -0.01]	-2.79	.005
Word frequency [b]	-0.16	0.01	[-0.18, -0.14]	-13.01	<.001
Visual strength [b]	-0.01	0.01	[-0.03, 0.01]	-1.36	.175
Interactions					
Word concreteness × Vocabulary age	0.01	0.01	[-0.01, 0.03]	1.16	.244
Word concreteness × Gender	0.00	0.01	[-0.02, 0.02]	0.16	.876
Word frequency × Vocabulary age	-0.02	0.01	[-0.04, 0.01]	-1.31	.191
Visual strength × Vocabulary age	0.00	0.01	[-0.02, 0.02]	0.05	.962
Word frequency × Gender	-0.02	0.01	[-0.04, 0.00]	-1.75	.080
Visual strength × Gender	-0.01	0.01	[-0.03, 0.01]	-0.86	.390

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. Shaded rows contain covariates.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Figure 20 presents the interactions of vocabulary age with word frequency and with visual strength, albeit non-significant.

Figure 21 shows the interactions of gender with language-based similarity and visual-strength difference, albeit non-significant.[10]

Statistical power analysis

Figures 22 and 23 show the estimated power available for main effects and interactions of interest, respectively.

[10] Plots of other interactions are available in Appendix D.

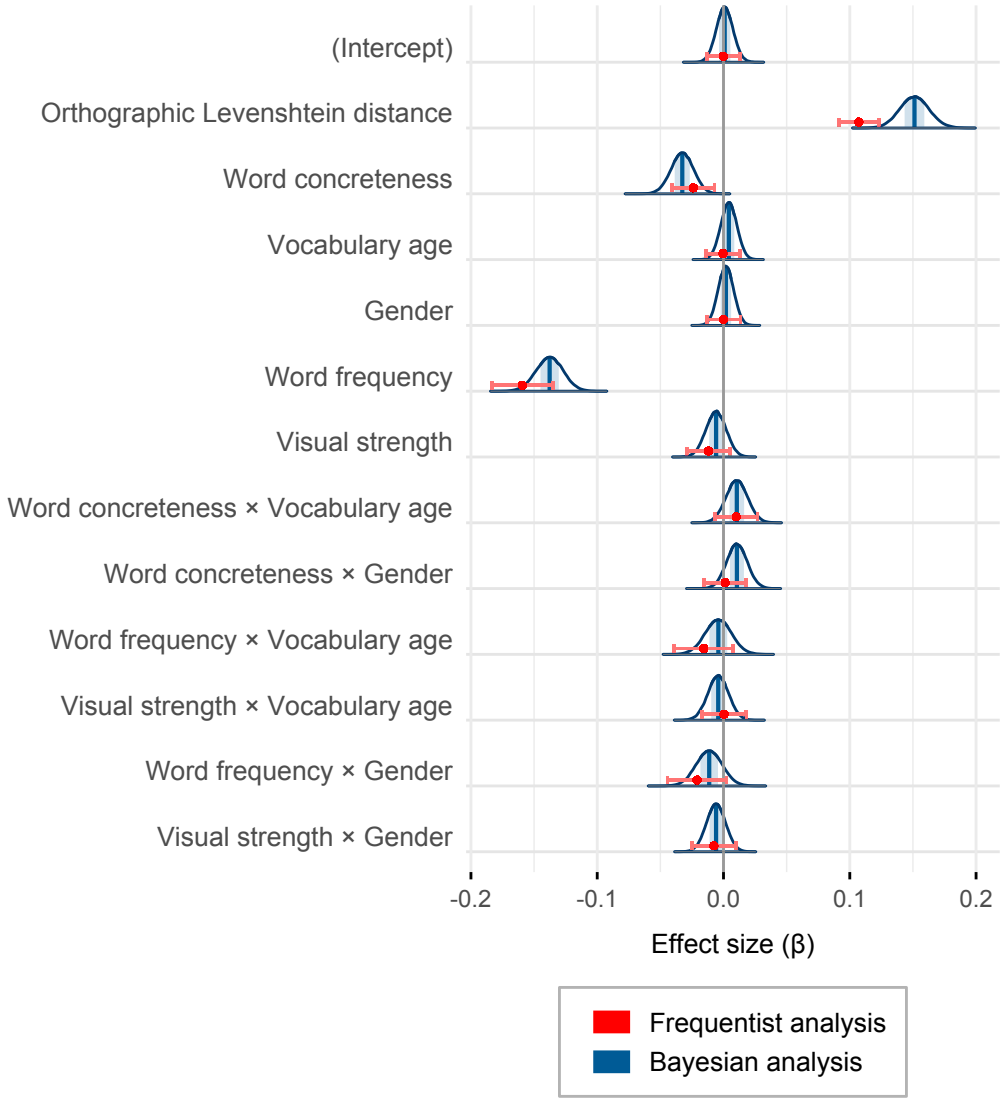

Figure 19

Estimates from the frequentist analysis (in red) and from the Bayesian analysis (in blue) for the lexical decision study. The frequentist means (represented by points) are flanked by 95% confidence intervals. The Bayesian means (represented by vertical lines) are flanked by 95% credible intervals, in light blue (in some cases, the interval is covered up by the bar of the mean).

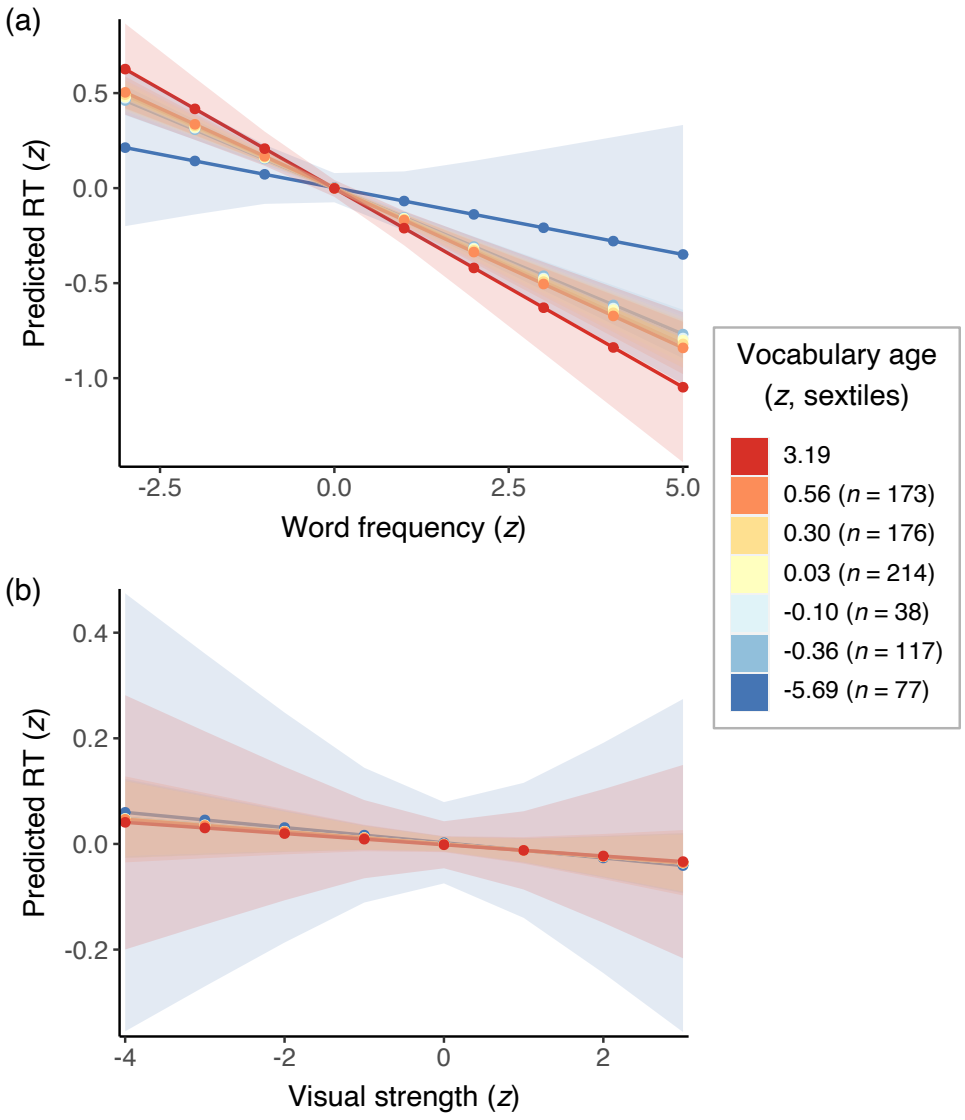

Figure 20

Interactions of vocabulary age with word frequency and visual strength. Vocabulary age is constrained to sextiles (six sections) in this plot, whereas in the statistical analysis it contained more values within the current range. n *= number of participants contained between sextiles.*

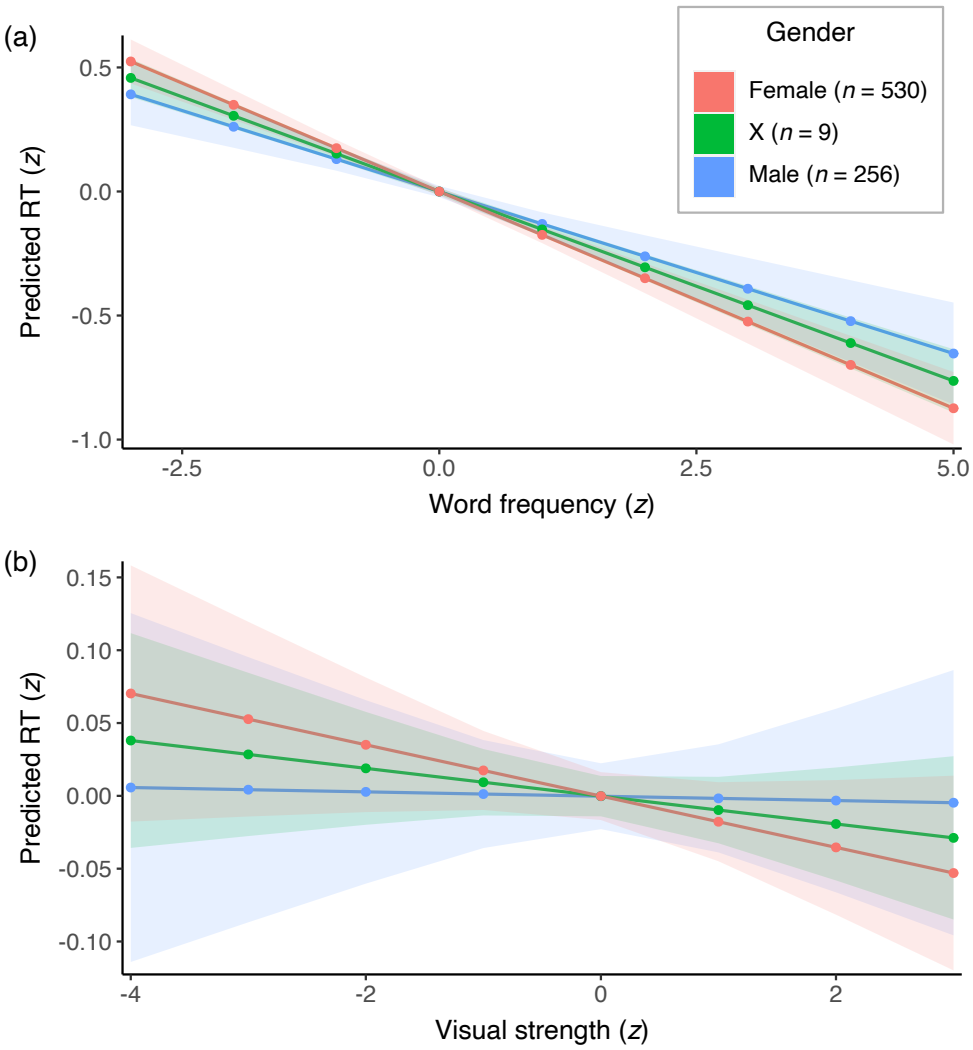

Figure 21

Interactions with gender in the lexical decision study. Gender was analysed using z-scores, but for clarity, the variable is shown in its basic form here.

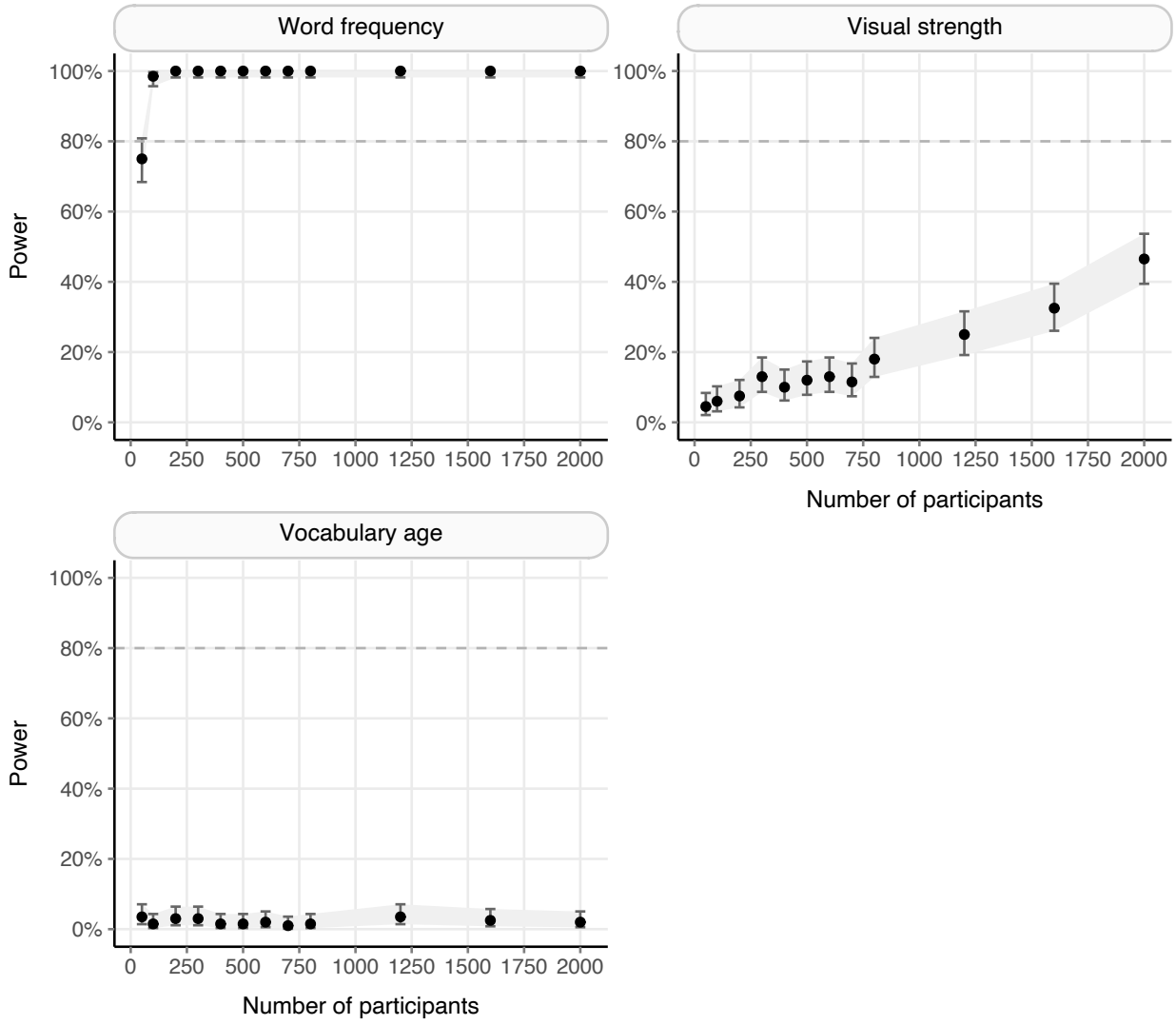

Figure 22

Power curves for some main effects in the lexical decision study.

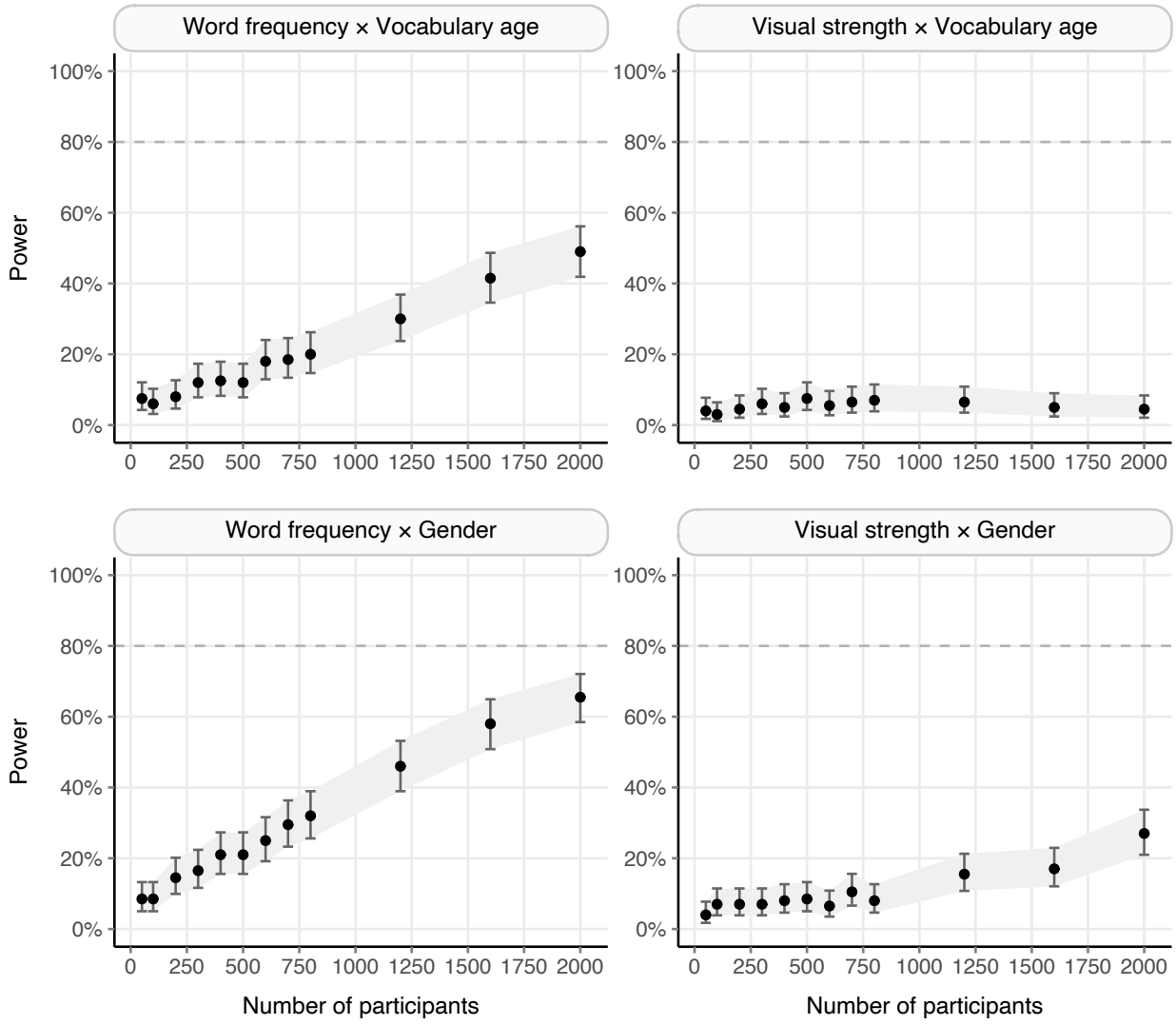

Figure 23

Power curves for some interactions in the lexical decision study.

_____ interpret _____

Discussion

General discussion

In all three studies, the main effects and the interactions of language-based information were larger than those of vision-based information, consistent with previous research [. . .]. Beyond that, the results revealed a dynamic process influenced by three levels of variation: participants, words and tasks. The associations that appeared across these levels revealed the roles of language-based and vision-based mechanisms in conceptual processing.

The RTs of higher-vocabulary participants were influenced by a smaller number of variables than those of lower-vocabulary participants. This converges with previous findings suggesting that higher and lower-vocabulary participants are affected by different variables. Potentially, the variables affecting higher-vocabulary participants are more relevant to the task (Lim et al., 2020; Pexman & Yap, 2018; Yap et al., 2012, 2017).

Differences between measurement instruments and the associated risks

We also compared two measures of visual priming. The first measure was operationalised as the difference in visual strength (Lynott et al., 2020) between the prime and the target word in each trial. The second measure, created by (Petilli et al., 2021), was based on vector representations trained on images. The results revealed that the visual strength measure was significantly superior in explaining RTs. This difference was not due to an excessive collinearity between these measures ($r = .02$). Also importantly, both measures appeared to be valid based on their correlations with language-based similarity and with word concreteness.

If we indeed accept that both the above measures were valid, we must reflect on the possibility that measurement instruments create confounds when different systems are compared. That is, in the case of linguistic and embodied processing, the large difference

between the effect sizes of these systems—found in the three current studies and in the previous literature [. . .]—would not be trustworthy if the instruments that were used to measure the language system were far more precise than the instruments used to measure the embodiment system. In this sense, consider how variables are refined in research: it is done by comparing the performance of different variables. Critically, the literature seems to contain many comparisons of text-based variables, some dating from the 1990s (De Deyne et al., 2016, 2013; Günther et al., 2016a, 2016b; Jones et al., 2006; Lund & Burgess, 1996; Mikolov et al., 2013; Wingfield & Connell, 2022). In contrast, the work on embodiment variables began more than a decade afterwards, and this work has been more concerned with comparisons of different *modalities*—e.g., valence, visual strength, auditory strength, etc. (Lynott et al., 2020; Lynott & Connell, 2009; Newcombe et al., 2012). Thus, this historical accident could account for a portion of the superiority of linguistic information over embodied information (Banks et al., 2021; Kiela & Bottou, 2014; as found in Lam et al., 2015; Louwerse et al., 2015; Pecher et al., 1998; Petilli et al., 2021).

The case of different measurement instruments is one of several factors that can exert a great influence in analyses. In the medium term, it may pay dividends to devote more work to comparing different instruments and, more generally, different analyses (see Barsalou, 2019; Botvinik-Nezer et al., 2020; Wagenmakers et al., 2022).

Statistical power

Power analyses were performed to _____. _____. The sample sizes required for some of the effects are not easily feasible with the usual distribution of funding in psychological research projects.

References

Al-Azary, H., Yu, T., & McRae, K. (2022). Can you touch the N400? The interactive effects of body-object interaction and task demands on N400 amplitudes and decision latencies. *Brain and Language*, *231*, 105147. https://doi.org/10.1016/j.bandl.2022.105147

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia*, *49*(5), 970–983. https://doi.org/10.1016/j.neuropsychologia.2011.01.003

Amsel, B. D., Urbach, T. P., & Kutas, M. (2014). Empirically grounding grounded cognition: The case of color. *NeuroImage*, *99*, 149–157. https://doi.org/10.1016/j.neuroimage.2014.05.025

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., ... Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad9163

Aujla, H. (2021). Language experience predicts semantic priming of lexical decision. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *75*(3), 235. https://doi.org/10.1037/cep0000255

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 336–345. https://doi.org/10.1037/0278-7393.12.3.336

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59*(4), 495–523. https://doi.org/10.1016/j.jml.2007.10.004

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Banks, B., Wingfield, C., & Connell, L. (2021). Linguistic distributional knowledge and sensorimotor grounding both contribute to semantic category production. *Cognitive Science*, *45*(10), e13055. https://doi.org/10.1111/cogs.13055

Barca, L., Mazzuca, C., & Borghi, A. (2020). Overusing the pacifier during infancy sets a footprint on abstract words processing. *Journal of Child Language*, *47*(5), 1084–1099. https://doi.org/10.1017/S0305000920000070

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barsalou, L. W. (2019). Establishing generalizable mechanisms. *Psychological Inquiry*, *30*(4), 220–230. https://doi.org/10.1080/1047840X.2019.1693857

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In *Symbols and Embodiment*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0013

Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*, *31*(5-6), 378–412. https://doi.org/10.1080/02643294.2014.895314

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Brauer, A., & Krivitsky, P. N. (2021). *Package 'lme4'*. CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1059–1082. https://doi.org/10.1037/0278-7393.23.5.1059

Bernabeu, P., Lynott, D., & Connell, L. (2021). *Preregistration: The interplay between linguistic and embodied systems in conceptual processing*. OSF. https://osf.io/ftydw

Bernabeu, P., Willems, R. M., & Louwerse, M. M. (2017). Modality switch effects emerge early and increase throughout conceptual processing: Evidence from ERPs. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1629–1634). Cognitive Science Society. https://mindmodeling.org/cogsci2017/papers/0318

Beyersmann, E., Grainger, J., & Taft, M. (2020). Evidence for embedded word length effects in complex nonwords. *Language, Cognition and Neuroscience*, *35*(2), 235–245. https://doi.org/10.1080/23273798.2019.1659989

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, *29*, 120–153. https://doi.org/10.1016/j.plrev.2018.12.001

Bottini, R., Morucci, P., D'Urso, A., Collignon, O., & Crepaldi, D. (2021). The concreteness advantage in lexical decision does not depend on perceptual simulations. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001090

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810, 7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological*

Methods, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1, 1), 16. https://doi.org/10.5334/joc.72

Brysbaert, M. (2022). Word recognition II. In M. J. Snowling, C. Hulme, & K. Nation, *The science of reading* (pp. 79–101). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119705116.ch4

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526. https://doi.org/10.3758/BF03193020

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://journal.r-project.org/archive/2018/RJ-2018-017/index.html

Bürkner, P.-C., Gabry, J., Weber, S., Johnson, A., Modrak, M., Badr, H. S., Weber, F., Ben-Shachar, M. S., & Rabel, H. (2022). *Package 'brms'*. CRAN. https://cran.r-project.org/web/packages/brms/brms.pdf

Burman, D., Bitan, T., & Both, J. (2008). Sex differences in neural processing of language among children. *Neuropsychologia*, *46, 5*, 1349–1362.

https://doi.org/10.1016/j.neuropsychologia.2007.12.021

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5, 5), 365–376. https://doi.org/10.1038/nrn3475

Cerni, T., Velay, J.-L., Alario, F.-X., Vaugoyeau, M., & Longcamp, M. (2016). Motor expertise for typing impacts lexical decision performance. *Trends in Neuroscience and Education*, *5*(3), 130–138. https://doi.org/10.1016/j.tine.2016.07.007

Charbonnier, J., & Wartena, C. (2020). Predicting the concreteness of German words. *Proceedings of the 5th Swiss Text Analytics Conference (SwissText)*, *2624*. https://doi.org/10.25968/opus-2075

Charbonnier, J., & Wartena, C. (2019). Predicting word concreteness and imagery. *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 176–187. https://doi.org/10.18653/v1/W19-0415

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Collins, J., Pecher, D., Zeelenberg, R., & Coulson, S. (2011). Modality switching in a property verification task: An erp study of what happens when candles flicker after high heels click. *Frontiers in Psychology*, *2*. https://www.frontiersin.org/article/10.3389/fpsyg.2011.00010

Connell, L. (2019). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, *34*(10), 1308–1318. https://doi.org/10.1080/23273798.2018.1471512

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452–465. https://doi.org/10.1016/j.cognition.2012.07.010

Connell, L., & Lynott, D. (2013). Flexible and fast: Linguistic shortcut affects both shallow and deep conceptual processing. *Psychonomic Bulletin & Review*, *20, 3*, 542–550. https://doi.org/10.3758/s13423-012-0368-x

Connell, L., & Lynott, D. (2014). I see/hear what you mean: Semantic activation in visual

word recognition depends on perceptual attention. *Journal of Experimental Psychology: General*, *143*(2), 527–533. https://doi.org/10.1037/a0034626

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170143. https://doi.org/10.1098/rstb.2017.0143

Corker, K. S., Lynott, D., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). High quality direct replications matter: Response to Williams (2014). *Social Psychology*, *45*(4), 324–326.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Davies, R. A., Arnell, R., Birchenough, J. M., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1298. https://doi.org/10.1037/xlm0000366

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922. https://doi.org/10.1111/cogs.12922

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

De Deyne, S., Perfors, A., & Navarro, D. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861–1870.

de Wit, B., & Kinoshita, S. (2015). The masked semantic priming effect is task dependent:

Reconsidering the automatic spreading activation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1062–1075. https://doi.org/10.1037/xlm0000074

Di Lollo, V., Mühlenen, A. von, Enns, J. T., & Bridgeman, B. (2004). Decoupling stimulus duration from brightness in metacontrast masking: Data and models. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 733–745. https://doi.org/10.1037/0096-1523.30.4.733

Diaz, M. T., Karimi, H., Troutman, S. B. W., Gertel, V. H., Cosgrove, A. L., & Zhang, H. (2021). Neural sensitivity to phonological characteristics is stable across the lifespan. *NeuroImage*, *225*, 117511. https://doi.org/10.1016/j.neuroimage.2020.117511

Dijkstra, T., Wahl, A., Buytenhuijs, F., Halem, N. V., Al-Jibouri, Z., Korte, M. D., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences*, *107*(37), 16396–16400. https://doi.org/10.1073/pnas.1009438107

Diveica, V., Pexman, P. M., & Binney, R. J. (2022). Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01810-x

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. *Cognition*, *110*(2), 284–292. https://doi.org/10.1016/j.cognition.2008.11.012

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799. https://doi.org/10.1037/0033-2909.125.6.777

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6). https://doi.org/10.1073/pnas.2108091119

Fetterman, A. K., Wilkowski, B. M., & Robinson, M. D. (2018). On feeling warm and being warm: Daily perceptions of physical warmth fluctuate with interpersonal warmth. *Social Psychological and Personality Science*, *9*(5), 560–567. https://doi.org/10.1177/1948550617712032

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. https://doi.org/10.1016/j.cognition.2020.104335

Flores d'Arcais, G. B., Schreuder, R., & Glazenborg, G. (1985). Semantic activation during recognition of referential words. *Psychological Research*, *47*(1), 39–49. https://doi.org/10.1007/BF00309217

Fox, J. (2016). Generalized linear models. In *Applied regression analysis and generalized linear models* (Third Edition, pp. 418–472). SAGE.

Gagné, C. L., Spalding, T. L., & Nisbet, K. A. (2016). Processing english compounds: Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics*, *13*(2), 2–22. https://link.gale.com/apps/doc/A469757337/LitRC?u=anon~b6a332f4&xid=9960afc7

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: The role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3-4), 455–479. https://doi.org/10.1080/02643290442000310

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–807.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Günther, F., Dudschig, C., & Kaup, B. (2016a). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Günther, F., Dudschig, C., & Kaup, B. (2016b). Predicting lexical priming effects from distributional semantic similarities: A replication with extension. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01646

Hald, L. A., Bastiaansen, M. C. M., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and Language*, *96*(1), 90–105. https://doi.org/10.1016/j.bandl.2005.06.007

Hald, L. A., Hocking, I., Vernon, D., Marshall, J. A., & Garnham, A. (2013). Exploring modality switching effects in negated sentences: Further evidence for grounded representations. *Frontiers in Psychology*, *4*, 93. https://doi.org/10.3389/fpsyg.2013.00093

Hald, L. A., Marshall, J. A., Janssen, D. P., & Garnham, A. (2011). Switching modalities in a sentence verification task: ERP evidence for embodied language processing. *Frontiers in Psychology*, *2*, 45. https://doi.org/10.3389/fpsyg.2011.00045

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, 4794. https://doi.org/10.7717/peerj.4794

Hauk, O. (2016). Only time will tell – why temporal information is essential for our

neuroscientific understanding of semantics. *Psychonomic Bulletin & Review*, *23*(4), 1072–1079. https://doi.org/10.3758/s13423-015-0873-9

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Hoedemaker, R. S., & Gordon, P. C. (2014). It takes time to prime: Semantic priming in the ocular lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2179–2197. https://doi.org/10.1037/a0037677

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Holt, L. E., & Beilock, S. L. (2006). Expertise and its embodiment: Examining the impact of sensorimotor skill expertise on the representation of action-related text. *Psychonomic Bulletin & Review*, *13*(4), 694–701. https://doi.org/10.3758/BF03193983

Hultén, A., Vliet, M. van, Kivisaari, S., Lammi, L., Lindh-Knuutila, T., Faisal, A., & Salmelin, R. (2021). The neural representation of abstract words may arise through grounding word meaning in language itself. *Human Brain Mapping*, *42*(15), 4973–4984. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25593

Hutchinson, S., & Louwerse, M. M. (2013). Language statistics and individual differences in processing primary metaphors. *Cognitive Linguistics*, *24*(4), 667–687. https://doi.org/10.1515/cog-2013-0023

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. https://doi.org/10.3758/BF03196544

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology*, *61*(7), 1036–1066. https://doi.org/10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse,

C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*, 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

Joordens, S., & Becker, S. (1997). The long and short of semantic priming effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1083–1105. https://doi.org/10.1037/0278-7393.23.5.1083

Jung, M., Mody, M., Fujioka, T., Kimura, Y., Okazawa, H., & Kosaka, H. (2019). Sex differences in white matter pathways related to language ability. *Frontiers in Human Neuroscience*, *13*, 898. https://doi.org/10.3389/fnins.2019.00898

Kiefer, M., Pielke, L., & Trumpp, N. M. (2022). Differential temporo-spatial pattern of electrical brain activity during the processing of abstract concepts related to mental states and verbal associations. *NeuroImage*, *252*, 119036. https://doi.org/10.1016/j.neuroimage.2022.119036

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*, 36–45. https://doi.org/10.3115/v1/D14-1005

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*(5), 1155–1180. https://doi.org/10.1007/s11145-018-9833-x

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01587-5

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects

models. *Journal of Statistical Software*, *75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Kos, M., Van den Brink, D., & Hagoort, P. (2012). Individual Variation in the Late Positive Complex to Semantic Anomalies. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/article/10.3389/fpsyg.2012.00318

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14–34. https://doi.org/10.1037/a0021446

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2261–2276. https://doi.org/10.1037/xlm0000793

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01546-0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lam, K. J., Dijkstra, T., & Rueschemeyer, S. A. (2015). Feature activation during word recognition: Action, visual, and associative-semantic priming effects. *Frontiers in Psychology*, *6*, 659. https://doi.org/10.3389/fpsyg.2015.00659

Lamiell, J. T. (2019). Statistical thinking in psychology: Some needed critical perspective on what "everyone knows." In J. T. Lamiell (Ed.), *Psychology's Misuse of Statistics and Persistent Dismissal of its Critics* (pp. 99–121). Springer International Publishing. https://doi.org/10.1007/978-3-030-12131-0_5

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic

analysis. *Discourse Processes*, *25*(2-3), 259–284. https://doi.org/10.1080/01638539809545028

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lim, R. Y., Yap, M. J., & Tse, C.-S. (2020). Individual differences in cantonese chinese word recognition: Insights from the chinese lexicon project. *Quarterly Journal of Experimental Psychology*, *73*(4), 504–518. https://doi.org/10.1177/1747021820906566

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*(2), 381–398. https://doi.org/10.1111/j.1551-6709.2010.01157.x

Louwerse, M. M., Hutchinson, S., Tillman, R., & Recchia, G. (2015). Effect size matters: The role of language statistics and perceptual simulation in conceptual processing. *Language, Cognition and Neuroscience*, *30*(4), 430–447. https://doi.org/10.1080/23273798.2014.981552

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the Cognitive Science Society*, 660–665.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558–564. https://doi.org/10.3758/BRM.41.2.558

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*, 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, *45*(3), 216–222. https://doi.org/10.1027/1864-9335/a000187

Mak, M., & Willems, R. M. (2019). Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience*, *34*(4), 511–535. https://doi.org/10.1080/23273798.2018.1552007

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., ... Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 1–7. https://doi.org/10.1038/s41586-022-04492-9

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin &*

Review, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

McDonald, S., & Brew, C. (2002). A distributional model of semantic context effects in lexical processing. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 17–24. http://dblp.uni-trier.de/db/conf/acl/acl2004.html#McDonaldB04

Mendes, P. S., & Undorf, M. (2021). On the pervasive effect of word frequency in metamemory. *Quarterly Journal of Experimental Psychology*, 17470218211053329. https://doi.org/10.1177/17470218211053329

Miceli, A., Wauthia, E., Lefebvre, L., Vallet, G. T., Ris, L., & Loureiro, I. S. (2022). Differences related to aging in sensorimotor knowledge: Investigation of perceptual strength and body object interaction. *Archives of Gerontology and Geriatrics*, *102*, 104715. https://doi.org/10.1016/j.archger.2022.104715

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (Version 3). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Milek, A., Butler, E. A., Tackman, A. M., Kaplan, D. M., Raison, C. L., Sbarra, D. A., Vazire, S., & Mehl, M. R. (2018). "Eavesdropping on happiness" revisited: A pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality. *Psychological Science*, *29*(9), 1451–1462. https://doi.org/10.1177/0956797618774252

Milton, F., Fulford, J., Dance, C., Gaddum, J., Heuerman-Williamson, B., Jones, K., Knight, K. F., MacKisack, M., Winlove, C., & Zeman, A. (2021). Behavioral and neural signatures of visual imagery vividness extremes: Aphantasia versus hyperphantasia. *Cerebral Cortex Communications*, *2*(2), 035. https://doi.org/10.1093/texcom/tgab035

Montero-Melis, G. (2021). Consistency in motion event encoding across languages. *Frontiers in Psychology*, *12*(625153). https://doi.org/10.3389/fpsyg.2021.625153

Montero-Melis, G., Eisenbeiss, S., Narasimhan, B., Ibarretxe-Antuñano, I., Kita, S., Kopecka, A., Lüpke, F., Nikitina, T., Tragel, I., Jaeger, T. F., & Bohnemeyer, J. (2017). Satellite- vs. Verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations. *Cognitive Semantics*, *3*(1),

36–61. https://doi.org/10.1163/23526416-00301002

Montero-Melis, G., van Paridon, J., Ostarek, M., & Bylund, E. (2022). No evidence for embodiment: The motor system is not needed to keep action verbs in working memory. *Cortex*, *150*, 108–125. https://doi.org/10.1016/j.cortex.2022.02.006

Moran, G. E., Cunningham, J. P., & Blei, D. M. (2021). *Posterior predictive null checks* (No. arXiv:2112.03333). arXiv. https://doi.org/10.48550/arXiv.2112.03333

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Newcombe, P., Campbell, C., Siakaluk, P., & Pexman, P. (2012). Effects of emotional and sensorimotor knowledge in semantic processing of concrete and abstract nouns. *Frontiers in Human Neuroscience*, *6*. https://www.frontiersin.org/article/10.3389/fnhum.2012.00275

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Ostarek, M., & Bottini, R. (2021). Towards strong inference in research on embodiment – Possibilities and limitations of causal paradigms. *Journal of Cognition*, *4*(1), 5. https://doi.org/10.5334/joc.139

Ostarek, M., & Huettig, F. (2017). A task-dependent causal role for low-level visual processes in spoken word comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1215–1224. https://doi.org/10.1037/xlm0000375

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science*, *28*(6), 593–599. https://doi.org/10.1177/0963721419866441

Pacini, A. M., & Barnard, P. J. (2021). Exocentric coding of the mapping between valence and regions of space: Implications for embodied cognition. *Acta Psychologica*, *214*, 103264. https://doi.org/10.1016/j.actpsy.2021.103264

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models.

Computational Linguistics, *33*(2), 161–199. https://doi.org/10.1162/coli.2007.33.2.161

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, *112*(33), 10089–10092. https://doi.org/10.1073/pnas.1504933112

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science*, *14*(2), 119–124. https://doi.org/10.1111/1467-9280.t01-1-01429

Pecher, D., Zeelenberg, R., & Raaijmakers, J. G. W. (1998). Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. *Journal of Memory and Language*, *38*(4), 401–418. https://doi.org/10.1006/jmla.1997.2557

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Studies in Written Language and Literacy* (Vol. 11, pp. 189–213). John Benjamins Publishing Company. https://doi.org/10.1075/swll.11.14per

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language*, *117*, 104194. https://doi.org/10.1016/j.jml.2020.104194

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*(2), 407–417. https://doi.org/10.3758/s13428-016-0720-6

Pexman, P. M., & Yap, M. J. (2018). Individual differences in semantic processing: Insights from the Calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1091–1112. https://doi.org/10.1037/xlm0000499

Ponari, M., Norbury, C. F., Rotaru, A., Lenci, A., & Vigliocco, G. (2018). Learning abstract words and concepts: Insights from developmental language disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170140. https://doi.org/10.1098/rstb.2017.0140

Ponari, M., Norbury, C. F., & Vigliocco, G. (2020). The role of emotional valence in learning novel abstract concepts. *Developmental Psychology*, *56*(10), 1855–1865. https://doi.org/10.1037/dev0001091

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), 12549. https://doi.org/10.1111/desc.12549

Pregla, D., Lissón, P., Vasishth, S., Burchert, F., & Stadie, N. (2021). Variability in sentence comprehension in aphasia in German. *Brain and Language*, *222*, 105008. https://doi.org/10.1016/j.bandl.2021.105008

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1–24. https://doi.org/10.1037/h0034650

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajananda, S., Lau, H., & Odegaard, B. (2018). A random-dot kinematogram for web-based vision research. *Journal of Open Research Software*, *6*(1, 1), 6. https://doi.org/10.5334/jors.194

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127–157. https://doi.org/10.1016/j.cogpsych.2009.09.001

Reilly, J., Flurie, M., & Peelle, J. E. (2020). The English lexicon mirrors functional brain activation for a sensory hierarchy dominated by vision and audition: Point-counterpoint. *Journal of Neurolinguistics*, *55*, 100895. https://doi.org/10.1016/j.jneuroling.2020.100895

Rodríguez-Ferreiro, J., Aguilera, M., & Davies, R. (2020). Semantic priming and schizotypal personality: Reassessing the link between thought disorder and enhanced spreading of semantic activation. *PeerJ*, *8*, e9511. https://doi.org/10.7717/peerj.9511

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for

psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. https://doi.org/10.3758/s13423-017-1420-7

Santos, A., Chaigneau, S. E., Simmons, W. K., & Barsalou, L. W. (2011). Property generation reflects word association and situated simulation. *Language and Cognition*, *3*(1), 83–119. https://doi.org/10.1515/langcog.2011.004

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language*, *162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2021). What is a Bayes factor? *Psychological Methods*. https://doi.org/10.1037/met0000421

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 421–439. https://doi.org/10.1037/xlm0000442

Schoot, R. van de, Depaoli, S., Gelman, A., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*, 3. https://doi.org/10.1038/s43586-020-00003-0

Schreuder, R., Flores d'Arcais, G. B., & Glazenborg, G. (1984). Effects of perceptual and conceptual similarity in semantic priming. *Psychological Research*, *45*(4), 339–354. https://doi.org/10.1007/BF00309710

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Psychology Press.

Sleegers, W. W. A., Proulx, T., & van Beest, I. (2021). Pupillometry and hindsight bias: Physiological arousal predicts compensatory behavior. *Social Psychological and Personality Science*, *12*(7), 1146–1154. https://doi.org/10.1177/1948550620966153

Snefjella, B., & Blank, I. (2020). *Semantic norm extrapolation is a missing data problem*. PsyArXiv. https://doi.org/10.31234/osf.io/y2gav

Solovyev, V. (2021). Concreteness/abstractness concept: State of the art. In B. M. Velichkovsky, P. M. Balaban, & V. L. Ushakov (Eds.), *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics* (pp. 275–283). Springer International Publishing. https://doi.org/10.1007/978-3-030-71637-0_33

Speed, L. J., van Dam, W. O., Hirath, P., Vigliocco, G., & Desai, R. H. (2017). Impaired comprehension of speed verbs in parkinson's disease. *Journal of the International Neuropsychological Society*, *23*(5), 412–420. https://doi.org/10.1017/S1355617717000248

Stasenko, A., Garcea, F. E., Dombovy, M., & Mahon, B. Z. (2014). When concepts lose their color: A case of object-color knowledge impairment. *Cortex*, *58*, 217–238. https://doi.org/10.1016/j.cortex.2014.05.013

Stone, K., Malsburg, T. von der, & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ*, *8*, e10438. https://doi.org/10.7717/peerj.10438

Stone, K., Veríssimo, J., Schad, D. J., Oltrogge, E., Vasishth, S., & Lago, S. (2021). The interaction of grammatically distinct agreement dependencies in predictive processing. *Language, Cognition and Neuroscience*, *36*(9), 1159–1179. https://doi.org/10.1080/23273798.2021.1921816

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*(3), 605–611. https://doi.org/10.3758/s13423-011-0078-9

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Tendeiro, J. N., & Kiers, H. A. L. (in press). On the white, the black, and the many shades of gray in between: Our reply to van Ravenzwaaij and Wagenmakers (2021).

Psychological Methods.

Tillman, R., Hutchinson, S., & Louwerse, M. M. (2015). How sharp is Occam's razor? Language statistics in cognitive processing. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2404–2409). Cognitive Science Society. https://cogsci.mindmodeling.org/2015/papers/0413/paper0413.pdf

Troche, J., Crutch, S. J., & Reilly, J. (2017). Defining a conceptual topography of word concreteness: Clustering properties of emotion, sensation, and magnitude among 750 english words. *Frontiers in Psychology*, *8*, 1787. https://doi.org/10.3389/fpsyg.2017.01787

Trumpp, N. M., Traub, F., & Kiefer, M. (2013). Masked priming of conceptual features reveals differential brain activation during unconscious access to conceptual action and sound information. *PLOS ONE*, *8*(5), e65910. https://doi.org/10.1371/journal.pone.0065910

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Ullman, M. T., Miranda, R. A., & Travers, M. L. (2008). Sex differences in the neurocognition of language. In J. B. Becker, K. J. Berkley, N. Geary, E. Hampson, J. Herman, & E. Young (Eds.), *Sex on the brain: From genes to behavior* (pp. 291–309). Oxford University Press.

Uttl, B. (2002). North American Adult Reading Test: Age norms, reliability, and validity. *Journal of Clinical and Experimental Neuropsychology*, *24*(8), 1123–1137. https://doi.org/10.1076/jcen.24.8.1123.8375

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2021). Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods*. https://doi.org/10.1037/met0000415

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance

filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Burkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Versace, R., Bailloud, N., Magnan, A., & Ecalle, J. (2021). The impact of embodied simulation in vocabulary learning. *The Mental Lexicon*, *16*(1), 2–22. https://doi.org/10.1075/ml.20011.ver

Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *7*(24), 1767–1777. https://doi.org/10.1093/cercor/bht025

Villalonga, M. B., Sussman, R. F., & Sekuler, R. (2021). Perceptual timing precision with vibrotactile, auditory, and multisensory stimuli. *Attention, Perception, & Psychophysics*, *83*(5), 2267–2280. https://doi.org/10.3758/s13414-021-02254-9

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Vukovic, N., & Williams, J. N. (2015). Individual differences in spatial cognition influence mental simulation of language. *Cognition*, *142*, 110–122. https://doi.org/10.1016/j.cognition.2015.05.017

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wallentin, M. (2020). Chapter 6 - Gender differences in language are small but matter for disorders. In R. Lanzenberger, G. S. Kranz, & I. Savic (Eds.), *Handbook of Clinical Neurology* (Vol. 175, pp. 81–102). Elsevier.

https://doi.org/10.1016/B978-0-444-64123-6.00007-2

Williams, L. E. (2014). Improving psychological science requires theory, data, and caution: Reflections on Lynott et al. (2014). *Social Psychology*, *45*(4), 321–323.

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, 1–51. https://doi.org/10.1080/23273798.2022.2069278

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213–220. https://doi.org/10.1016/j.cognition.2018.05.008

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III tests of cognitive abilities*. Riverside Publishing.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38, 1*, 53–79. https://doi.org/10.1037/a0024177

Yap, M. J., Balota, D. A., & Tan, S. E. (2013). Additive and interactive effects in semantic priming: Isolating lexical and decision processes in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 140–158. https://doi.org/10.1037/a0028520

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2017). Individual differences in semantic priming performance: Insights from the semantic priming project. In M. N. Jones (Ed.), *Frontiers of cognitive psychology. Big data in cognitive science* (pp. 203–226). Routledge/Taylor & Francis Group.

Yap, M. J., Tse, C.-S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language*, *61*(3), 303–325. https://doi.org/10.1016/j.jml.2009.07.001

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yee, E., Ahmed, S. Z., & Thompson-Schill, S. L. (2012). Colorless green ideas (can) prime furiously. *Psychological Science*, *23*(4), 364–369. https://doi.org/10.1177/0956797611430691

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, *140*(3), 348–363. https://doi.org/10.1037/a0022840

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heuerman-Williamson, B., Jones, K., & MacKisack, M. (2020). Phantasia—the psychological significance of lifelong visual imagery vividness extremes. *Cortex*, *130*, 426–440. https://doi.org/10.1016/j.cortex.2020.04.003

Appendix A: Selection of lexical covariates

Lexical covariates are usually used in conceptual processing studies due to the widespread connections among lexical and semantic variables. Including these covariates—or nuisance variables—in the model allows a more rigorous analysis of the predictors of interest (Sassenhagen & Alday, 2016).

In each study, the covariates were selected out of a group of 5 variables that had been used as covariates in Wingfield and Connell (2022), and are widely used (e.g., Petilli et al., 2021). Some of these covariates were highly intercorrelated ($r > .70$), as shown below. To avoid the problem of multicollinearity, the maximum zero-order correlation allowed between any two covariates was of $r = \pm.70$ (Dormann et al., 2013; Harrison et al., 2018). In cases of higher correlations, the covariate with the largest effect in the model, based on the estimate (β), was selected.

In Studies 1 (semantic priming) and 2 (semantic decision), the lexical covariates were selected out of five variables, which mirrored those used by Wingfield and Connell (2022): namely, number of letters (i.e., orthographic length, which we computed in R), word frequency, number of syllables (both the latter from Balota et al., 2007), orthographic Levenshtein distance (Yarkoni et al., 2008) and phonological Levenshtein distance (Suárez et al., 2011). In Study 3 (lexical decision), the procedure was identical except that word frequency could not be selected as a covariate because it constituted a predictor of interest, specifically corresponding to linguistic information (for details, see Study 3 in the main article).

All the models included by-participant and by-word random intercepts, as well as by-participant random slopes for every predictor. Below, the correlations and the selection model are shown for each study.

Study 1: Semantic priming

All lexical covariates considered in the semantic priming study were based on the target words. Figure A1 shows the zero-order correlations among the lexical covariates considered in the selection.

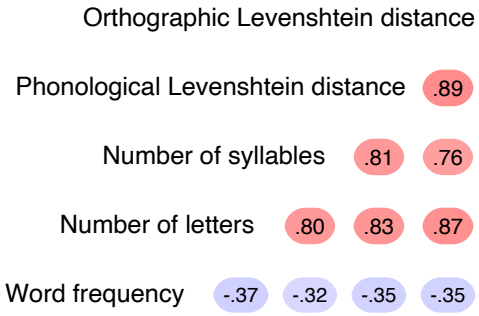

Figure A1
Zero-order correlations among lexical covariates pretested in the semantic priming study.

Table A1 shows the results of the selection model.

Table A1
Mixed-effects model for the selection of lexical covariates in the semantic priming study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.01	0.00	[0.00, 0.02]	1.19	.236
Word frequency	-0.14	0.01	[-0.15, -0.13]	-24.19	<.001
Number of letters	0.00	0.01	[-0.02, 0.02]	0.12	.903
Number of syllables	0.04	0.01	[0.02, 0.06]	4.02	<.001
Orthographic Levenshtein distance	0.03	0.01	[0.00, 0.05]	2.19	.029
Phonological Levenshtein distance	0.02	0.01	[-0.01, 0.04]	1.28	.199

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. By-participant random slopes were included for every effect.

Considering the maximum correlation allowed ($r = \pm.70$) and the results of the model, the variables that will be included as covariates in the main analysis are word frequency and number of syllables.

Study 2: Semantic decision

Figure A2 shows the zero-order correlations among the lexical covariates considered in the selection.

Table A2 shows the results of the selection model.

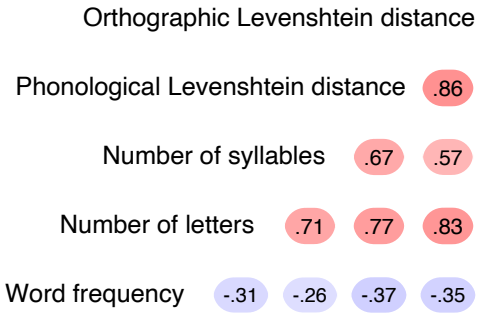

Figure A2
Zero-order correlations for the lexical covariates pretested in the semantic decision study.

Table A2
Mixed-effects model for the selection of lexical covariates in the semantic decision study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.05	0.00	[0.05, 0.06]	12.35	<.001
Word frequency	-0.13	0.01	[-0.14, -0.11]	-20.01	<.001
Number of letters	0.05	0.01	[0.03, 0.07]	5.20	<.001
Number of syllables	0.08	0.01	[0.07, 0.10]	10.80	<.001
Orthographic Levenshtein distance	-0.13	0.01	[-0.15, -0.10]	-10.23	<.001
Phonological Levenshtein distance	0.01	0.01	[-0.01, 0.03]	0.91	.361

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. By-participant random slopes were included for every effect.

Considering the maximum correlation allowed ($r = \pm.70$) and the results of the model, the variables that will be included as covariates in the main analysis are word frequency and orthographic Levenshtein distance.

Study 3: Lexical decision

Figure A3 shows the zero-order correlations among the lexical covariates considered in the selection.

Table A3 shows the results of the selection model.

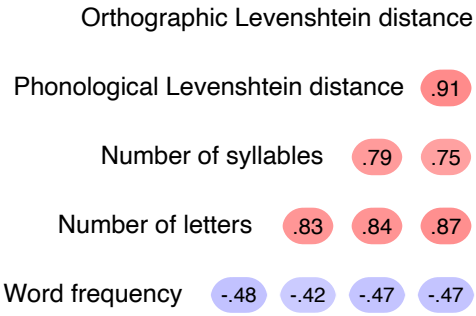

Figure A3
Zero-order correlations for the lexical covariates pretested in the lexical decision study.

Table A3
Mixed-effects model for the selection of lexical covariates in the lexical decision study.

	β	*SE*	95% CI	*t*	*p*
(Intercept)	0.00	0.01	[-0.01, 0.01]	-0.02	.981
Word frequency	-0.12	0.01	[-0.15, -0.10]	-11.60	<.001
Number of letters	0.05	0.02	[0.01, 0.09]	2.73	.006
Number of syllables	0.06	0.01	[0.03, 0.09]	4.43	<.001
Orthographic Levenshtein distance	0.10	0.02	[0.05, 0.14]	4.52	<.001
Phonological Levenshtein distance	-0.02	0.02	[-0.06, 0.02]	-1.18	.238

Note. β = Estimate based on *z*-scored variables; *SE* = standard error; CI = confidence interval. By-participant random slopes were included for every effect.

Considering the maximum correlation allowed ($r = \pm.70$), the results of the model, and the use of word frequency as a predictor of interest in the model, the variable that will be included as a covariate in the main analysis is orthographic Levenshtein distance.

Conclusion

Word frequency presented the largest effect in the three models. Orthographic Levenshtein distance was the second largest effect in the semantic decision and the lexical decision studies, whereas its phonological counterpart was not significant in any of the studies. The latter difference makes sense, as participants read the stimulus words in the three studies (Brysbaert, 2022).

Appendix B: Diagnostics for the frequentist analyses

Below, the convergence warnings and the non-normal residuals are first addressed generally, and then in more detail in the context of each study.

Convergence

The challenge of convergence is well known in the area of mixed-effects models. These models often struggle to reach reliable-enough estimates due to an insufficiency of data relative to the complexity of the model (Baayen et al., 2008; Bates et al., 2015; Brauer & Curtin, 2018). The solutions proposed range from the removal of random slopes under certain conditions (Matuschek et al., 2017) to the maintenance of random slopes in spite of convergence warnings, which seeks to avoid an inflation of the Type I error due to dependencies in the data (Brauer & Curtin, 2018; Singmann & Kellen, 2019).

The multiple-optimizers sanity check from `lme4::allFit()`

Framed within the drive to maintain random slopes wherever possible, the developers of the 'lme4' package propose a sanity check that uses a part of the 'lme4' *engine* called 'optimizer'. Every model has a default optimizer, unless a specific one is chosen through `control = lmerControl(optimizer = '...')` (in `lmer` models) or `control = glmerControl(optimizer = '...')` (in `glmer` models). The 7 widely-available optimizers are:

- bobyqa
- Nelder_Mead
- nlminbwrap
- nmkbw
- optimx.L-BFGS-B
- nloptwrap.NLOPT_LN_NELDERMEAD
- nloptwrap.NLOPT_LN_BOBYQA

To assess whether convergence warnings render the results invalid, or on the contrary, the results can be deemed valid in spite of the warnings, Bates et al. (2021)

suggest refitting models affected by convergence warnings with a variety of optimizers. The authors argue that if the different optimizers produce practically-equivalent results, the results are valid. For this purpose, the 'allFit' function from the 'lme4' package allows the refitting of models using a number of optimizers. To use the 7 optimizers listed above, two extra packages were installed: 'dfoptim' and 'optimx' (see lme4 manual). The output from 'allFit' contains several statistics on the fixed and the random effects fitted by each optimizer (see example).

The severity of convergence problems in each study will be examined below using the 'allFit' function from the 'lme4' package.

Residual errors not normally distributed

The residuals of the linear mixed-effects models in all three studies violated the assumption of normality. Even though linear mixed-effects models tend to be quite robust to deviations from normality (Knief & Forstmeier, 2021; Schielzeth et al., 2020), we sought to verify our results. To this end, two robust models were performed in each study, as described below.

Method A:* robustlmm *model

The first method uses the R package 'robustlmm' v2.4-4 (Koller, 2016). To calculate the p values, we followed the procedure of Sleegers et al. (2021), but used the Kenward-Roger method instead of Satterthwaite (see Luke, 2017).

Method B: Inverse Gaussian model with identity link function

We followed a method proposed by Lo and Andrews (2015), based on generalized linear mixed-effects models (GLMM) implementing an identity link function. According to Lo and Andrews (2015), the link function helps avoid directly transforming the dependent variable, which can hinder the interpretability of the results (also see Knief & Forstmeier, 2021).

GLMMs require the use of families of distributions. Lo and Andrews (2015) tested the Gaussian, Gamma and Inverse Gaussian families, with either an identity or an inverse

link function. The authors found that the Inverse Gaussian family with an identity link yielded the most normal residuals. The Inverse Gaussian and the Gamma families only accept positive values in the outcome variable (see Table 15.2 in Fox, 2016). Due to this restriction, the dependent variable in the present model is raw RT, unlike the standardised RT that was used in the main analysis.

P values were to be calculated through parametric bootstrapping, which is the most robust method for GLMMs, as the Kenward-Roger and Satterthwaite methods are not available for these models (Luke, 2017; Singmann et al., 2021).

Neither Method A nor Method B could finally be used, as the code produced errors. These errors are shown in the corresponding scripts inside the 'model_diagnostics' folder in each study. Nonetheless, the residuals are shown below.

Study 1: Semantic priming

Convergence

In the initial model, the optimizer used (the default one in 'lmerTest') was 'nloptwrap', and the convergence warning read: 'boundary (singular) fit: see ?isSingular'.

Based on the reanalysis using 7 optimizers, Figure B1 shows the fixed, main effects, and Figure B2 shows the fixed interactions.

Residual errors not normally distributed

Figure B3 shows the deviation from normality of the residuals of the linear mixed-effects model.

Semantic priming model including visual similarity

Convergence. In the initial model, the optimizer used (the default one in 'lmerTest') was 'nloptwrap', and the convergence warning read: 'boundary (singular) fit: see ?isSingular'.

Based on the reanalysis using 7 optimizers, Figure B4 shows the fixed, main effects, and Figure B5 shows the fixed interactions.

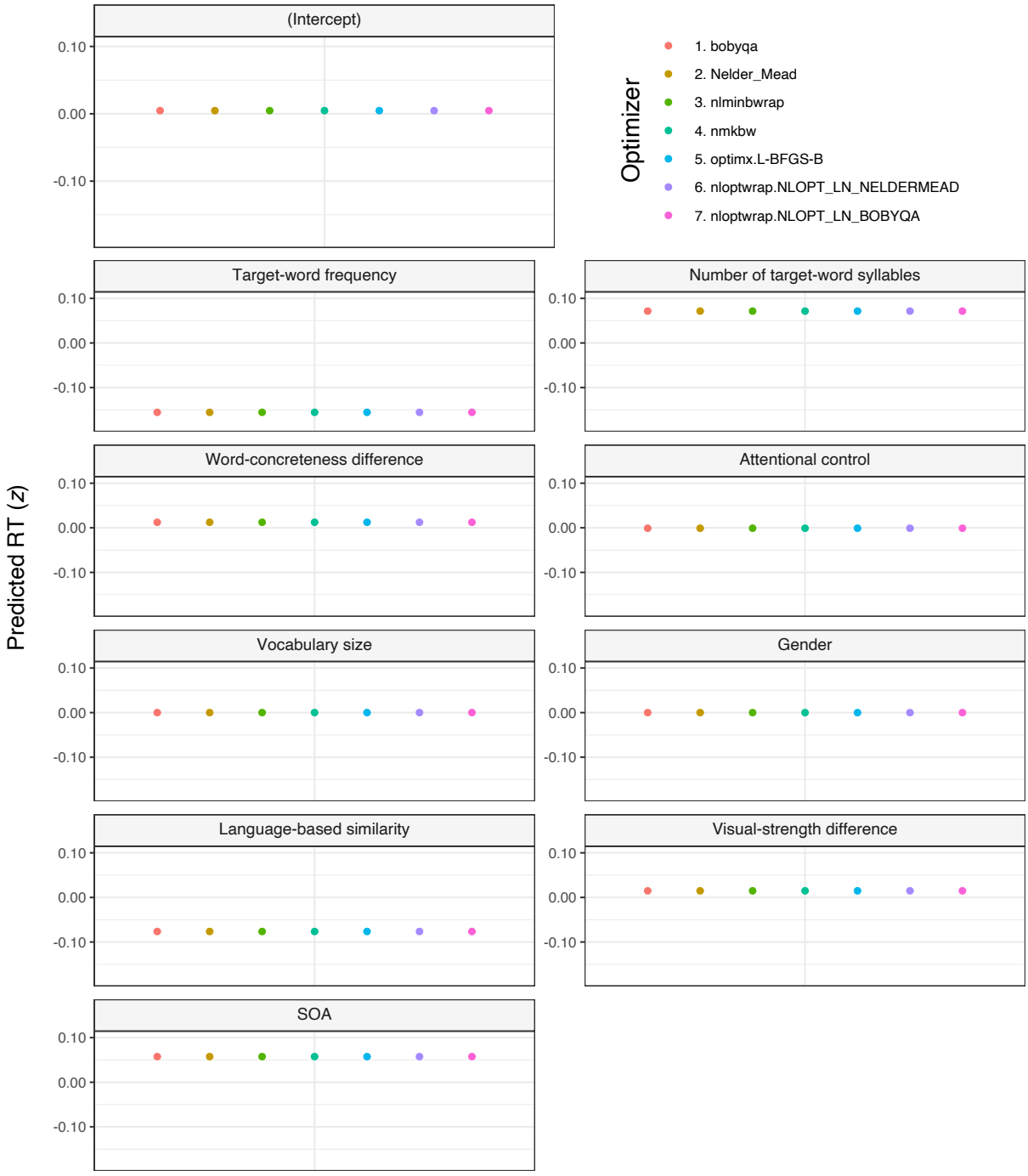

Figure B1

Fixed, main effects from the semantic priming study fitted by 7 optimizers.

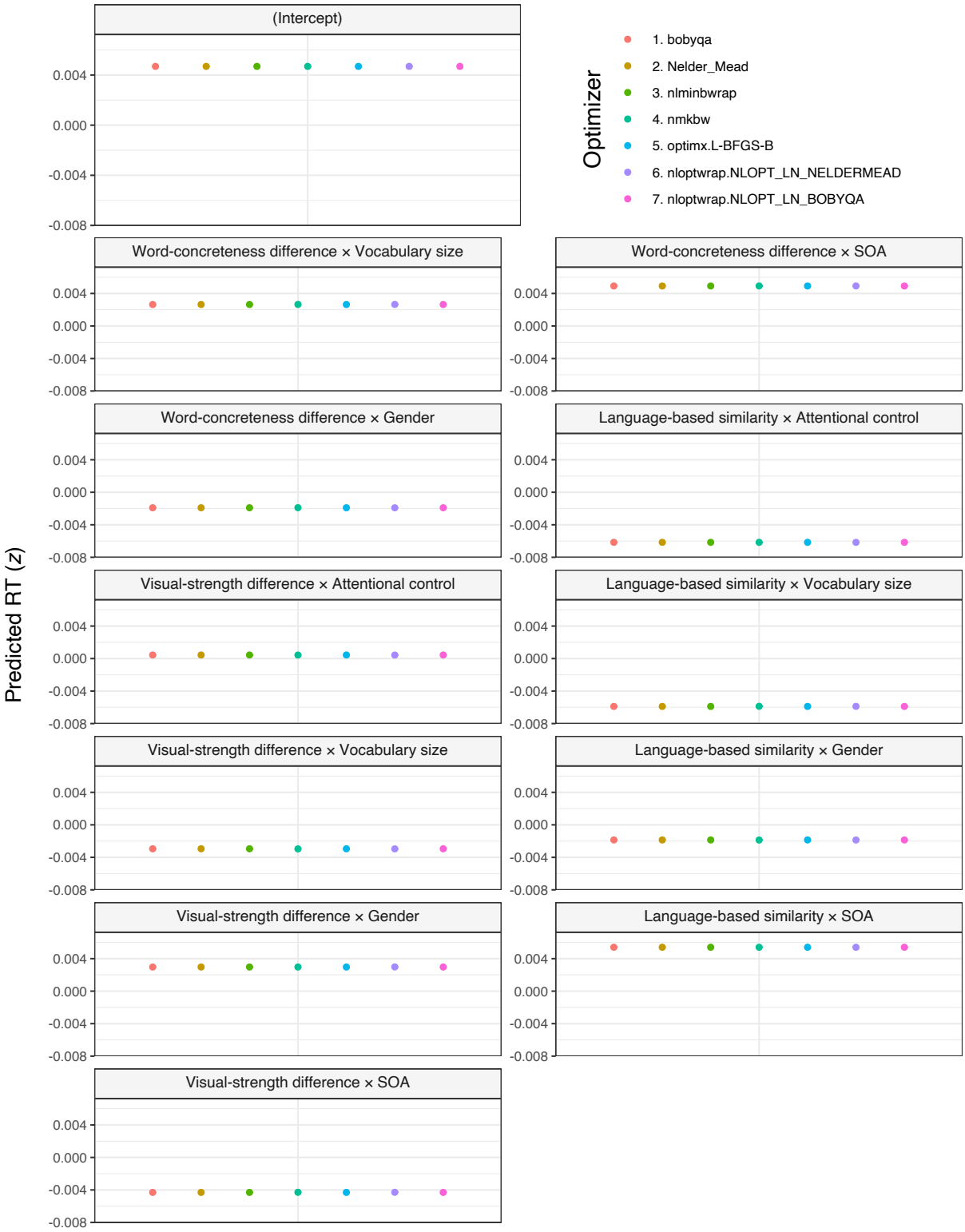

Figure B2

Fixed interaction effects from the semantic priming study fitted by 7 optimizers.

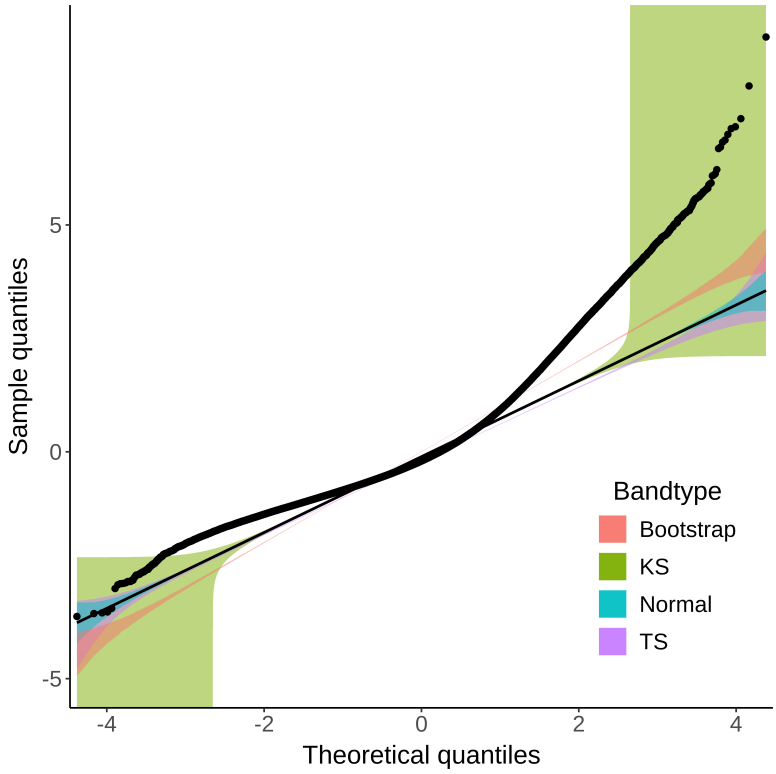

Figure B3
Residuals of the linear mixed-effects model from the semantic priming study. KS = Kolmogorov-Smirnov test; TS = tail-sensitive confidence bands.

Residual errors not normally distributed. Figure B6 shows the deviation from normality of the residuals of the linear mixed-effects model.

Study 2: Semantic decision

Convergence

In the initial model, the optimizer used (the default one in 'lmerTest') was 'nloptwrap', and the convergence warning read: 'boundary (singular) fit: see ?isSingular'.

Based on the reanalysis using 7 optimizers, Figure B7 shows the fixed, main effects, and Figure B8 shows the fixed interactions.

Residual errors not normally distributed

Figure B9 shows the deviation from normality of the residuals.

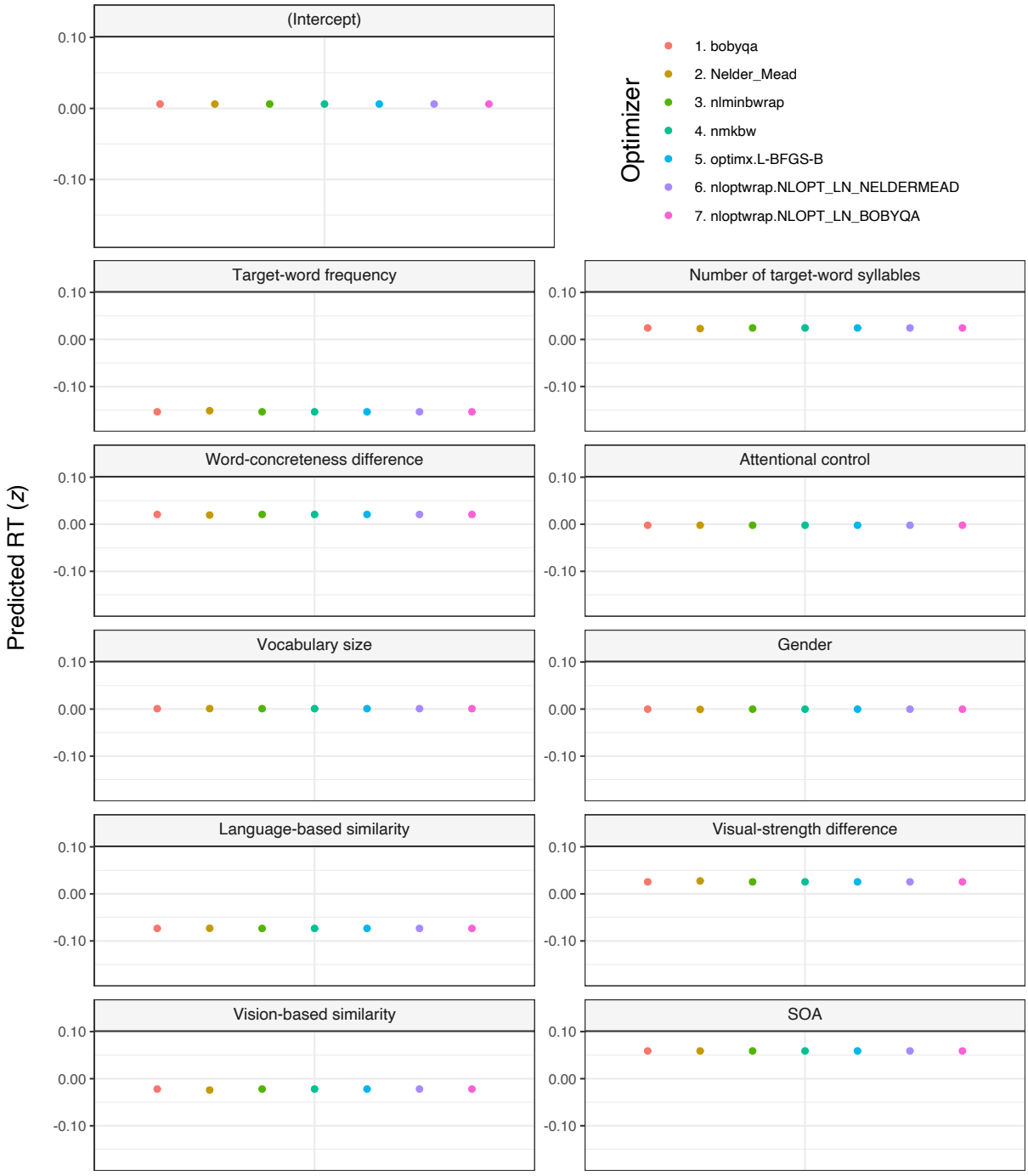

Figure B4

Fixed, main effects from the semantic priming study fitted by 7 optimizers.

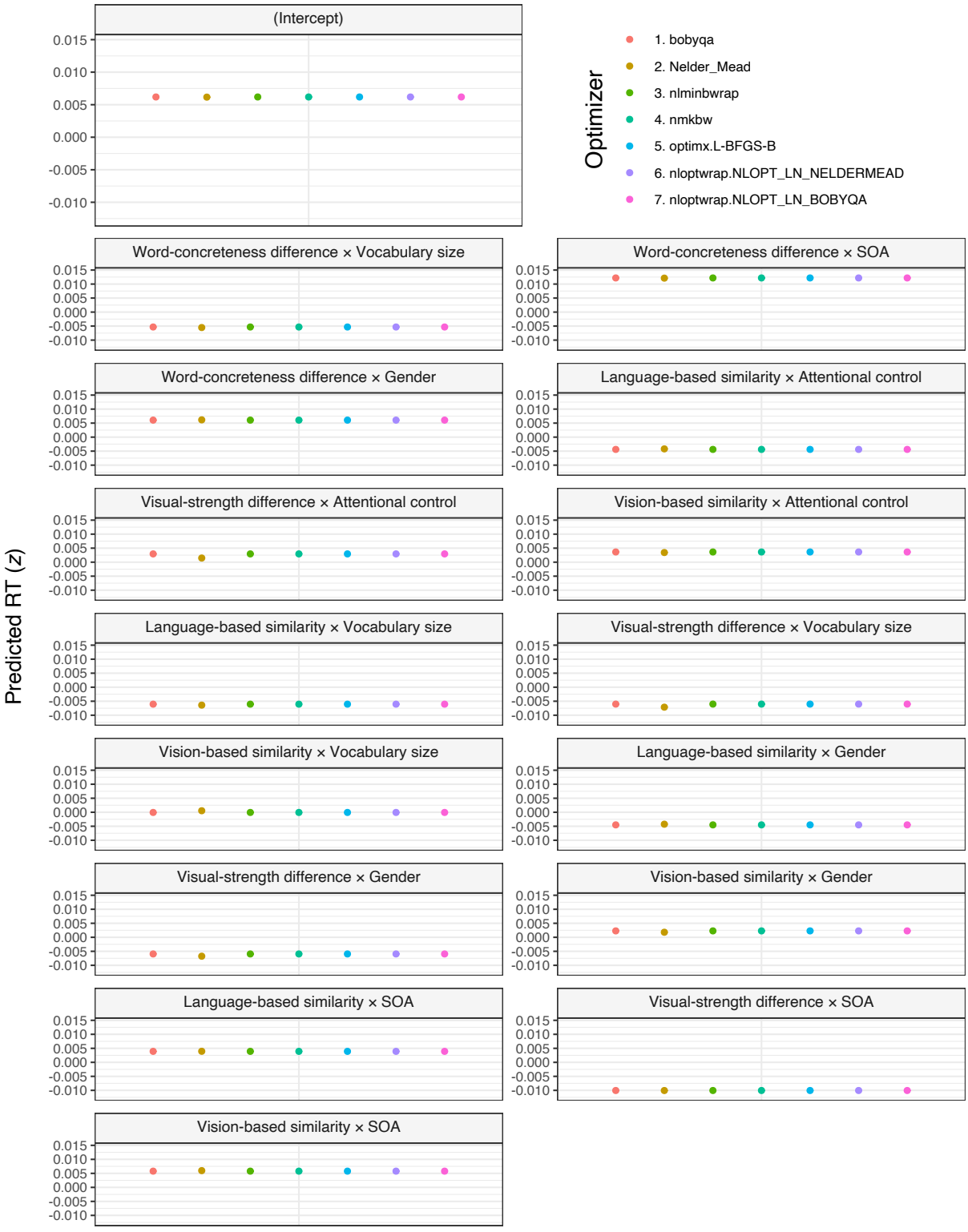

Figure B5

Fixed interaction effects from the semantic priming study fitted by 7 optimizers.

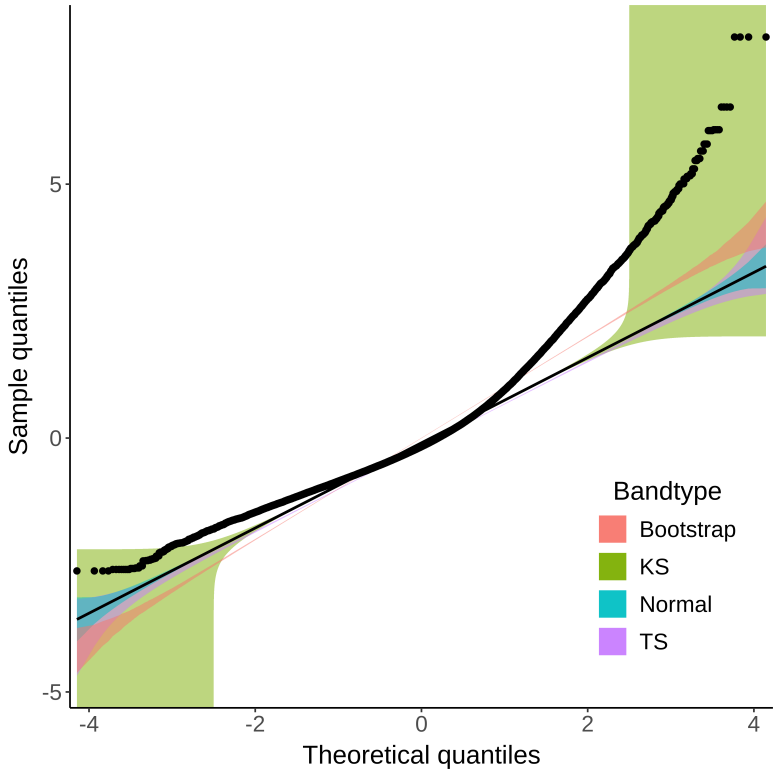

Figure B6

Residuals of the linear mixed-effects model from the semantic priming study. KS = Kolmogorov-Smirnov test; TS = tail-sensitive confidence bands.

Study 3: Lexical decision

Convergence

In the initial model, the optimizer used (the default one in 'lmerTest') was 'nloptwrap', and the convergence warning read: 'boundary (singular) fit: see ?isSingular'.

Based on the reanalysis using 7 optimizers, Figure B10 shows the fixed, main effects, and Figure B11 shows the fixed interactions.

Residual errors not normally distributed

Figure B12 shows the deviation from normality of the residuals.

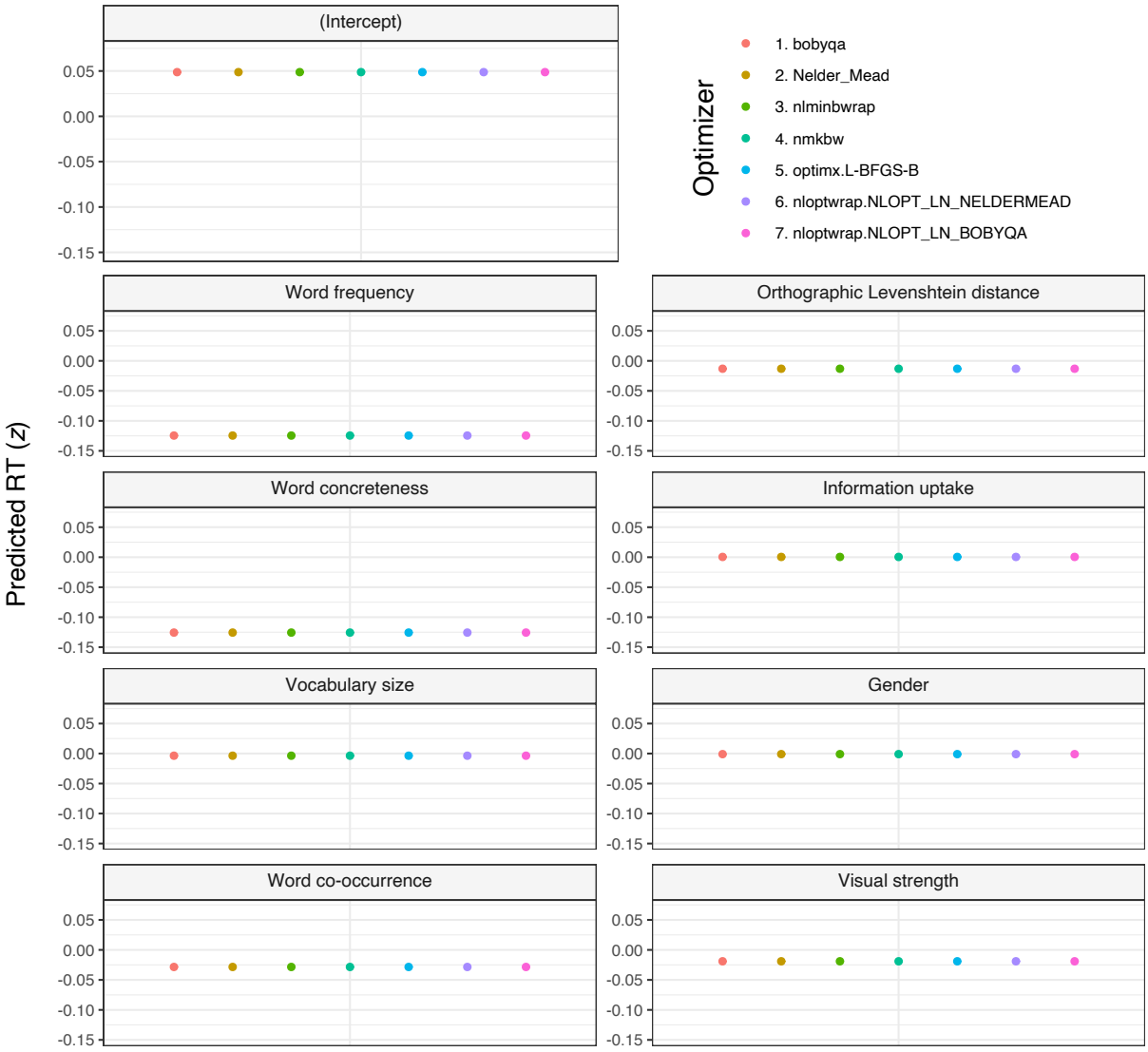

Figure B7

Fixed, main effects from the semantic decision study fitted by 7 optimizers.

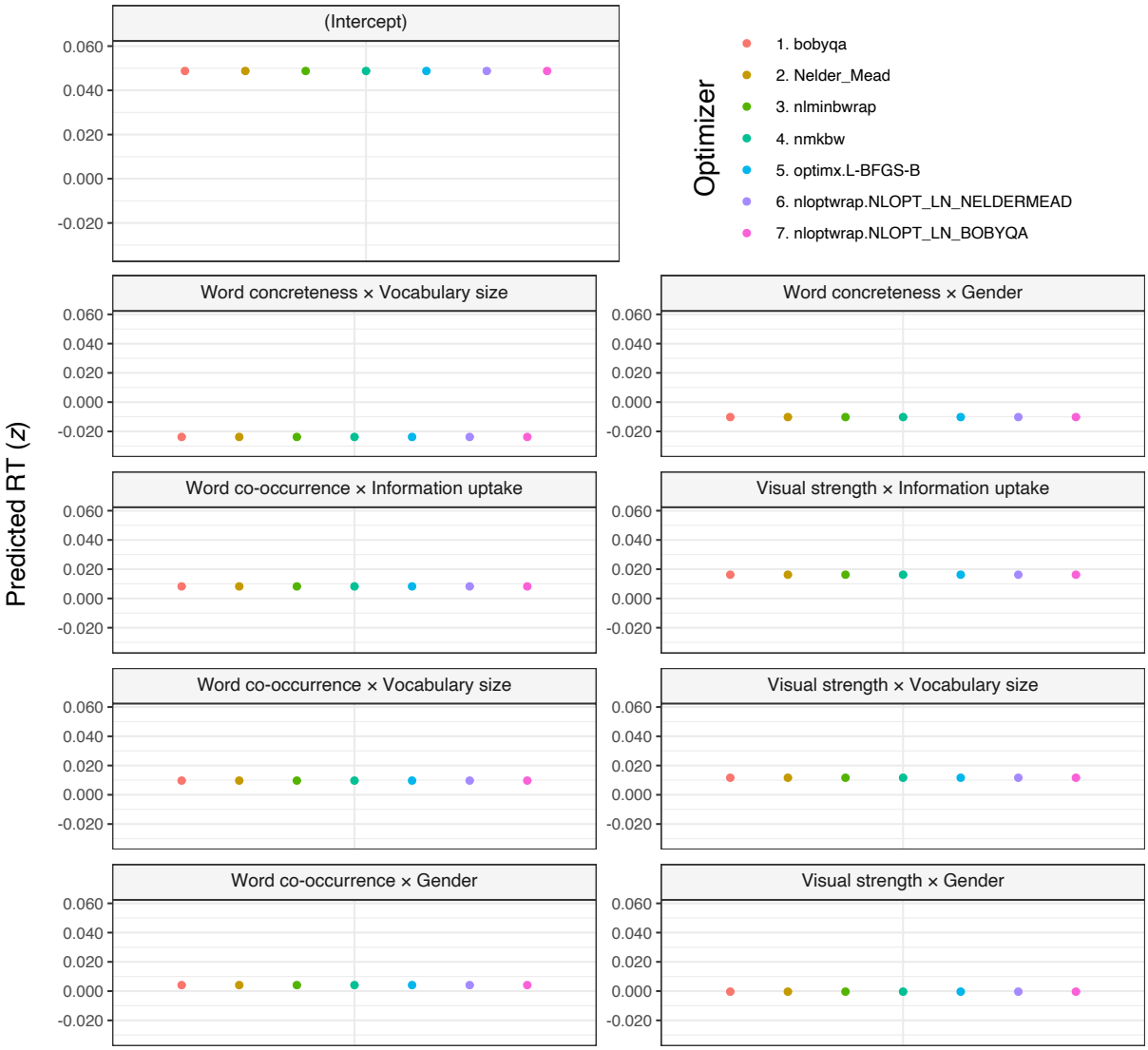

Figure B8

Fixed interaction effects from the semantic decision study fitted by 7 optimizers.

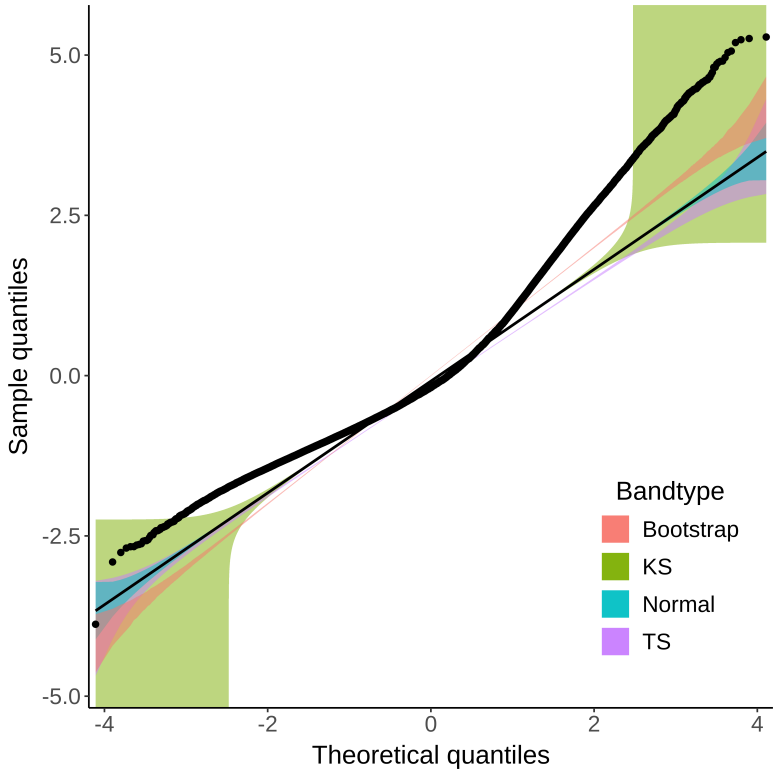

Figure B9

Residuals of the linear mixed-effects model from the semantic decision study. KS = Kolmogorov-Smirnov test; TS = tail-sensitive confidence bands.

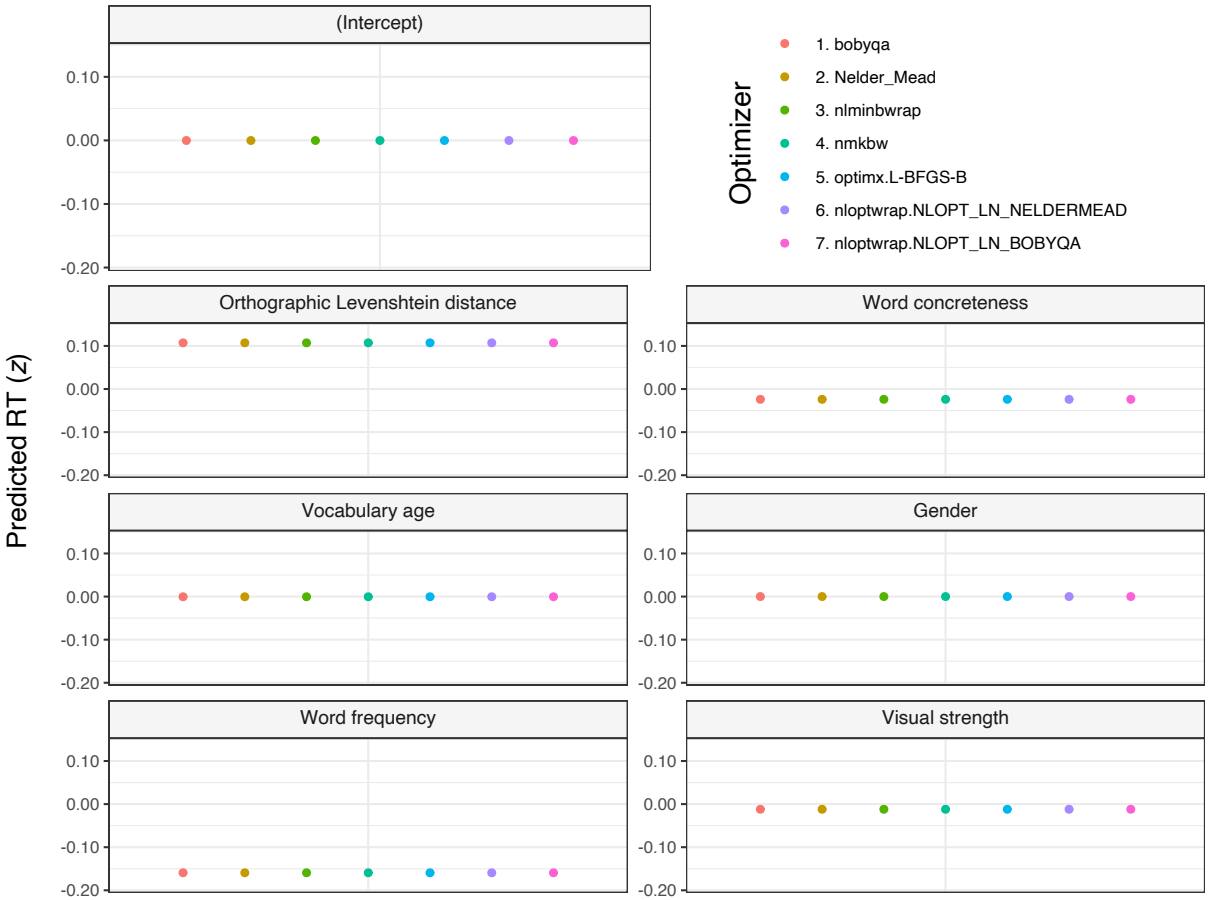

Figure B10

Fixed, main effects from the lexical decision study fitted by 7 optimizers.

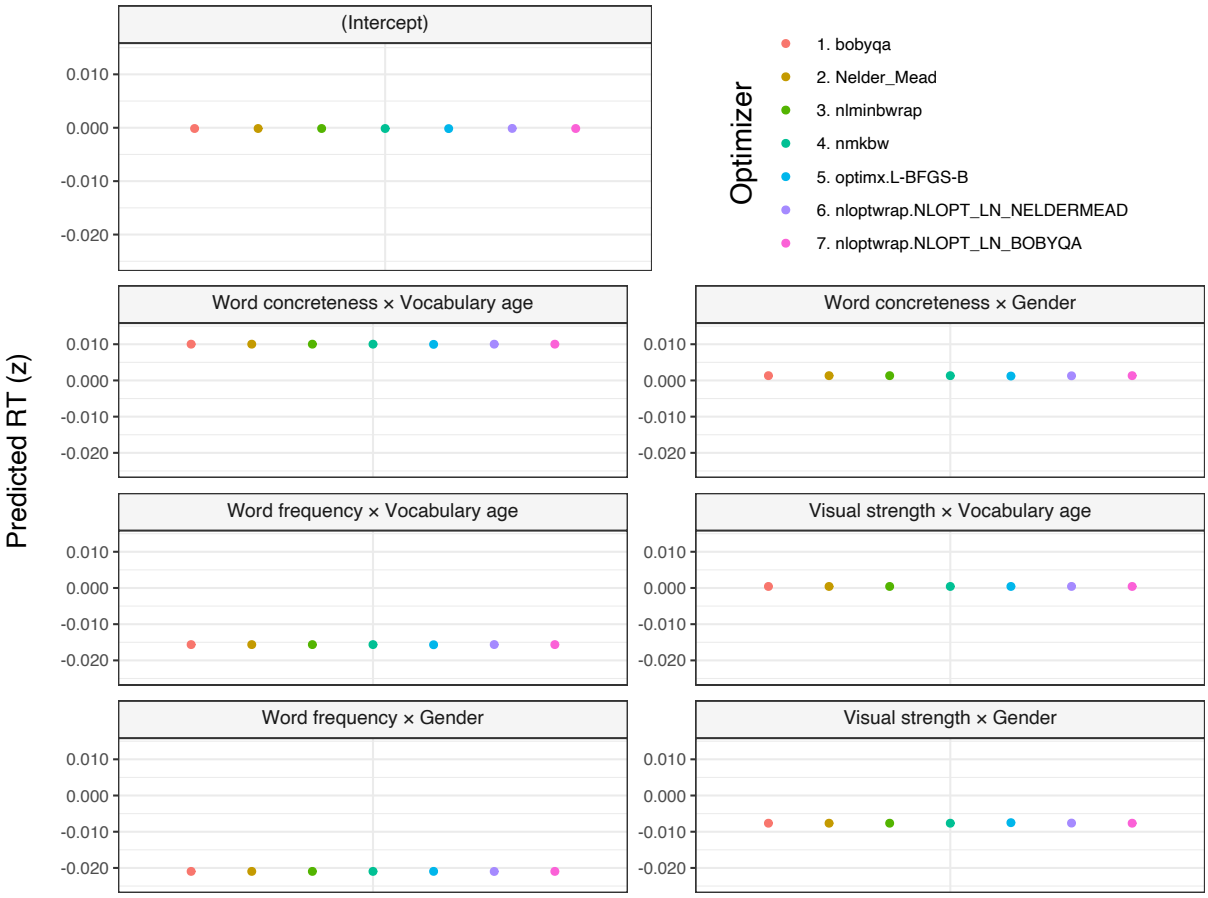

Figure B11

Fixed interaction effects from the lexical decision study fitted by 7 optimizers.

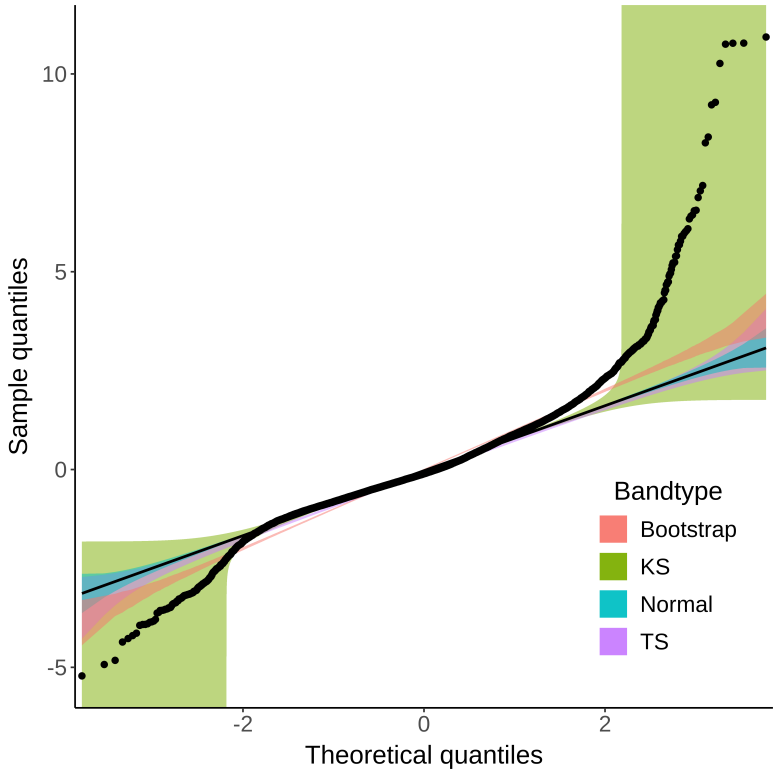

Figure B12

Residuals of the linear mixed-effects model from the lexical decision study. The outliers in the residuals deviate from the coloured areas indicating an acceptable normality. KS = Kolmogorov-Smirnov test; TS = tail-sensitive confidence bands.

Appendix C: Diagnostics for the Bayesian analyses

This appendix presents diagnostics for the Bayesian analyses. In each study, prior predictive checks are presented before posterior predictive checks. Furthermore, in each of these checks, the models presented first have the default Gaussian distribution, whereas the next series have an exponentially modified Gaussian (dubbed 'ex-Gaussian') distribution with an identity link function (for details, see the section titled 'Distributions and prior predictive checks' in the main article). Eyeball estimation is used to assess the outcomes of these checks (for background on predictive checks and for alternative estimation procedures, see Gelman et al., 1996; Moran et al., 2021; Schoot et al., 2021). One diagnostic not shown in this appendix is the $\widehat{R}$, which is shown in Appendix E instead.

Study 1: Semantic priming

Prior predictive checks

Figures C1, C2 and C3 show the prior predictive checks for the Gaussian models. These plots show the maximum, mean and minimum values of the observed data (y) and those of the predicted distribution (y_{rep}, which stands for *rep*lications of the outcome). The way of interpreting these plots is by comparing the observed data to the predicted distribution. The specifics of this vary across the three plots in the series. Firstly, in the upper plot, which shows the maximum values, the ideal scenario would have the observed (y) value overlapping with the maximum value of the predicted distribution (y_{rep}). Secondly, in the middle plot, showing the mean values, the ideal scenario would show the observed (y) value overlapping with the mean value of the predicted distribution (y_{rep}). Lastly, in the lower plot, which shows the minimum values, the ideal scenario would have the observed (y) value overlapping with the minimum value of the predicted distribution (y_{rep}). The overlap need not be absolute but the closer the observed and the predicted values on the X axis, the better. As such, the three predictive checks below—corresponding to models that used the default Gaussian distribution—show that the priors fitted the data acceptably but not very well.

In contrast to the above results, Figures C4, C5 and C6 demonstrate that, when an

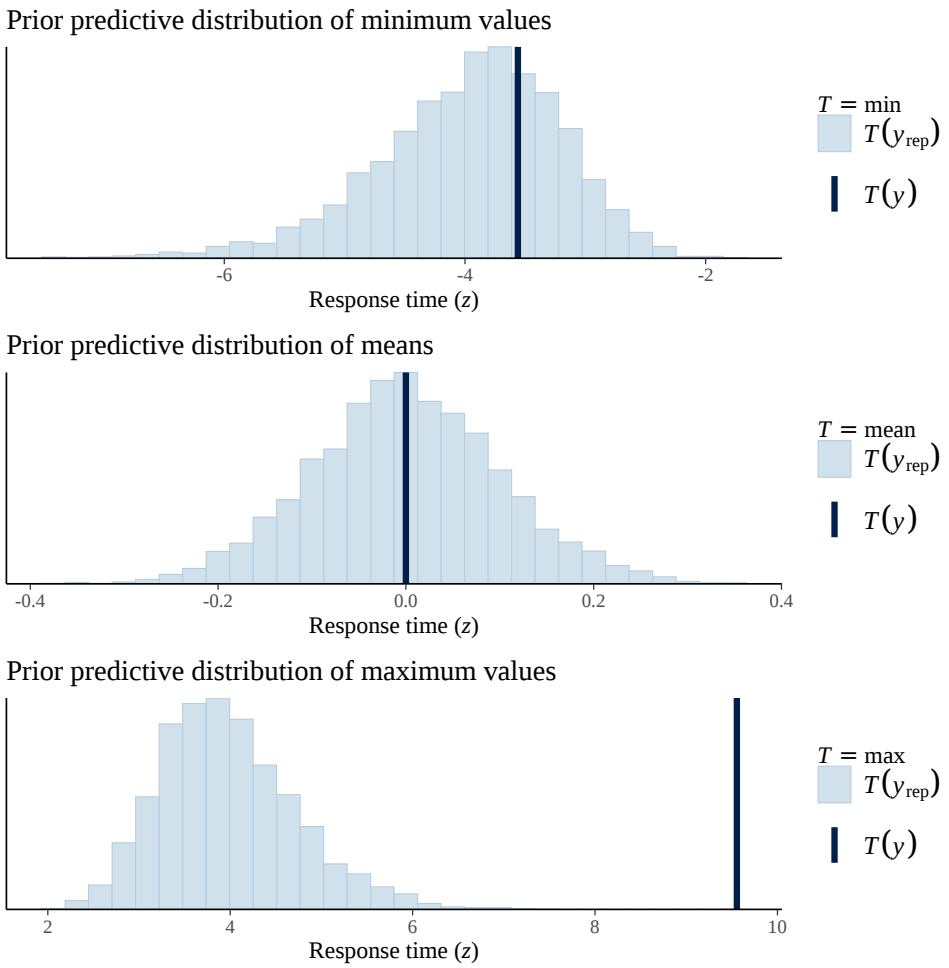

Figure C1

Prior predictive checks for the Gaussian, informative prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

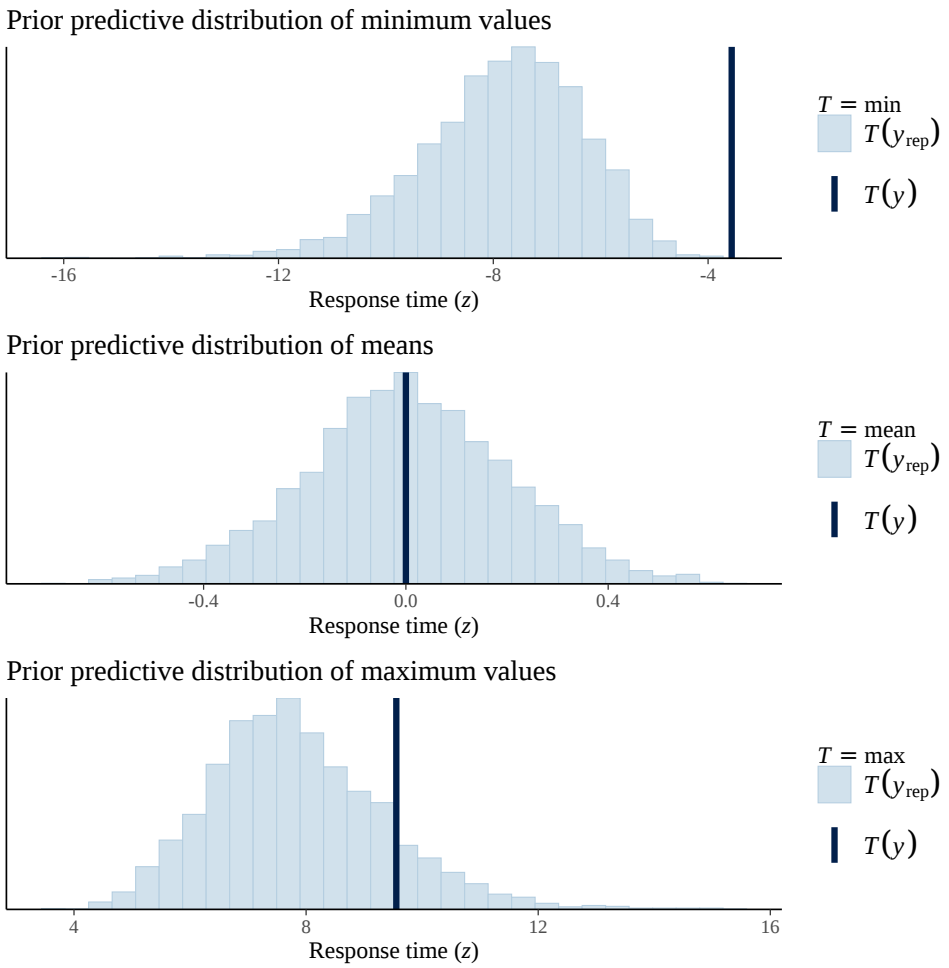

Figure C2

Prior predictive checks for the Gaussian, weakly-informative prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

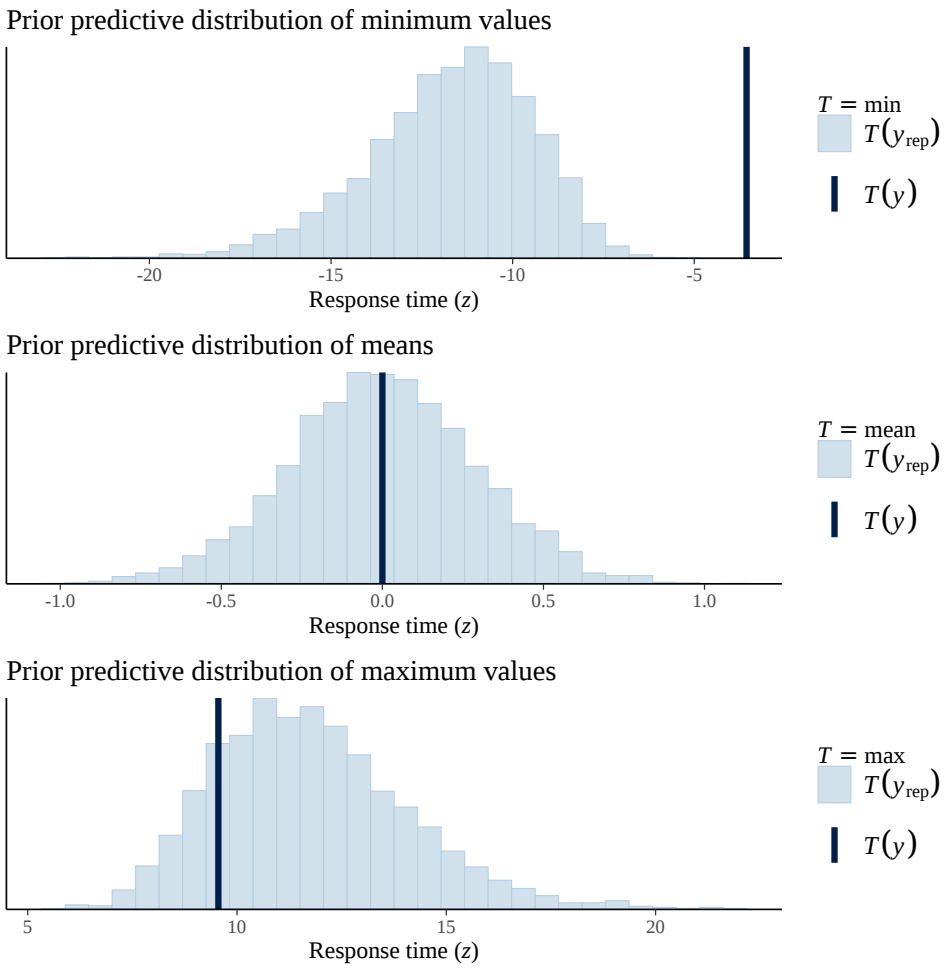

Figure C3

Prior predictive checks for the Gaussian, diffuse prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

ex-Gaussian distribution was used, the priors fitted the data far better, which converged with the results of a similar comparison performed by Rodríguez-Ferreiro et al. (2020) (see supplementary materials of the latter study).

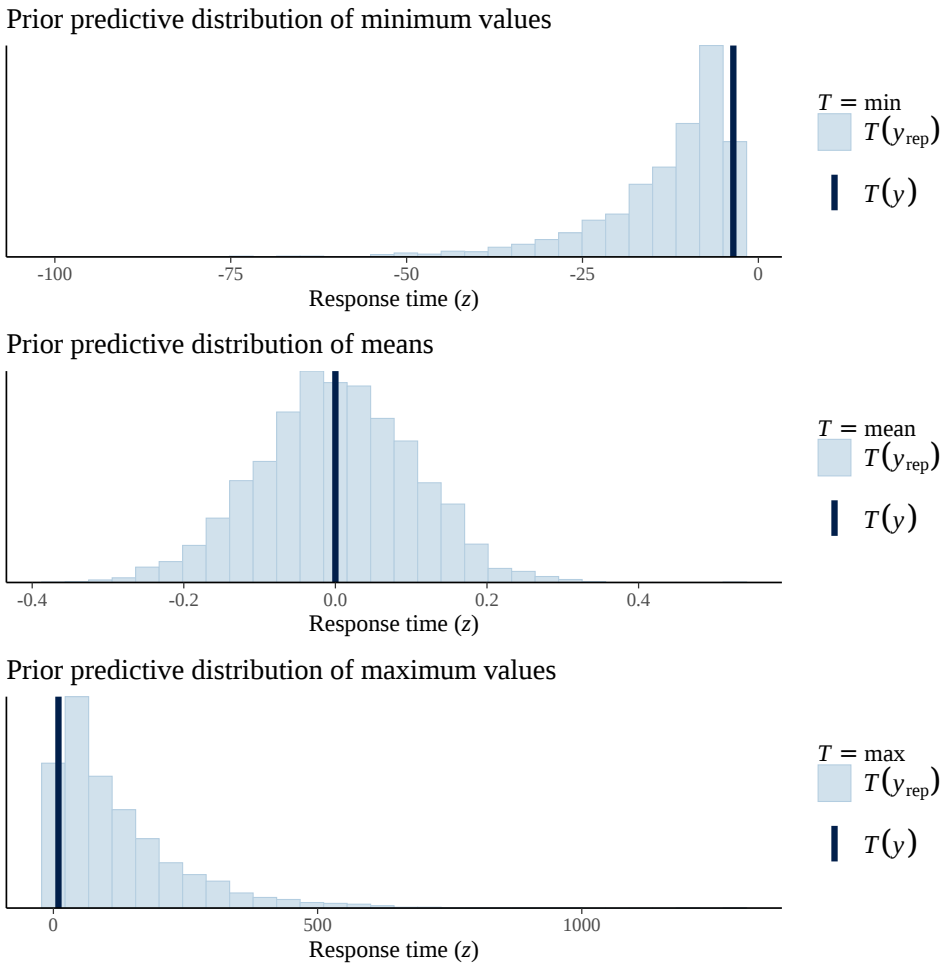

Figure C4

Prior predictive checks for the ex-Gaussian, informative prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

Posterior predictive checks

Based on the above results, the ex-Gaussian distribution was used in the final models. Figure C7 presents the posterior predictive checks the latter models. The interpretation of these plots is simple: the distributions of the observed (y) and the predicted data (y_{rep}) should be as similar as possible. As such, the plots below suggest that the results are trustworthy.

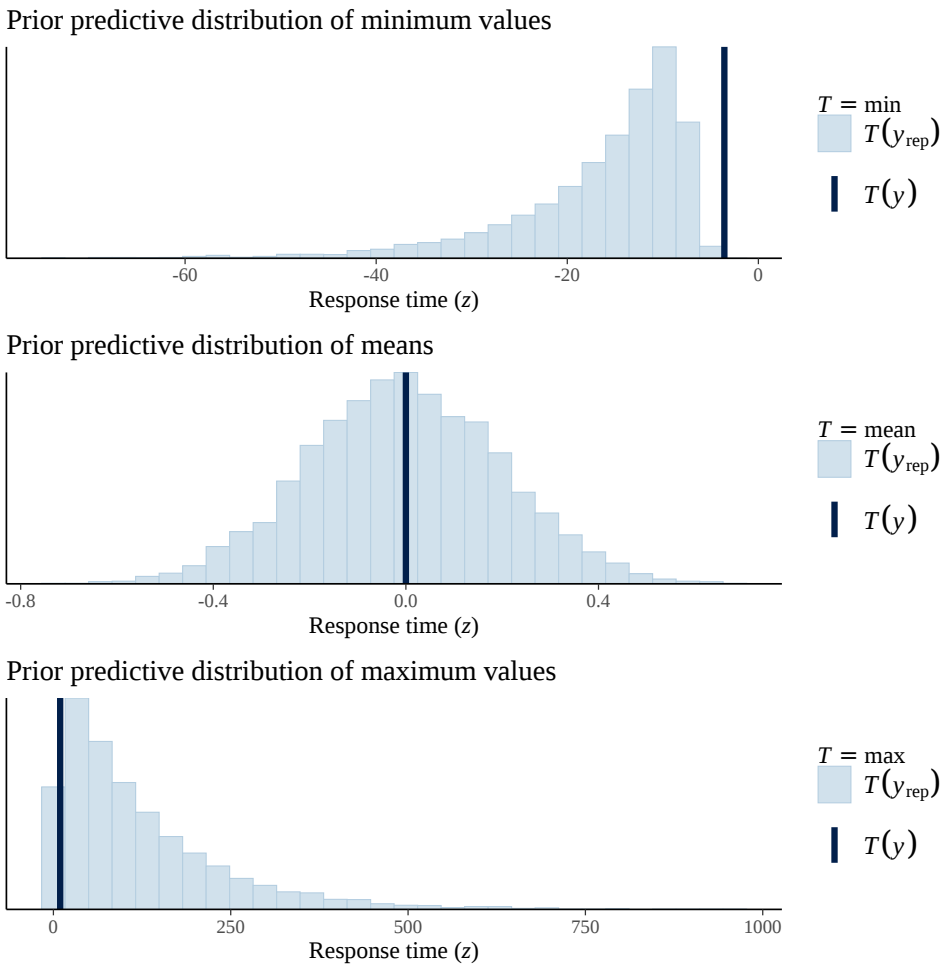

Figure C5

Prior predictive checks for the ex-Gaussian, weakly-informative prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

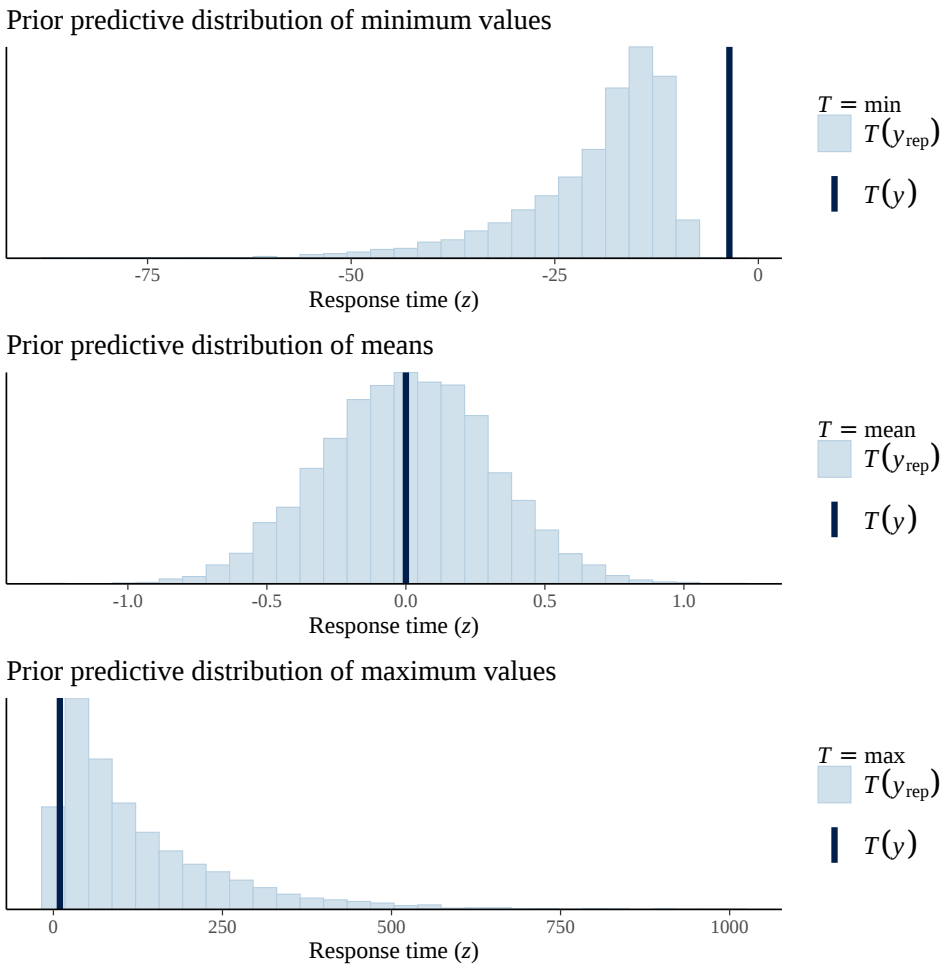

Figure C6

Prior predictive checks for the ex-Gaussian, diffuse prior model from the semantic priming study. y = observed data; y_{rep} = predicted data.

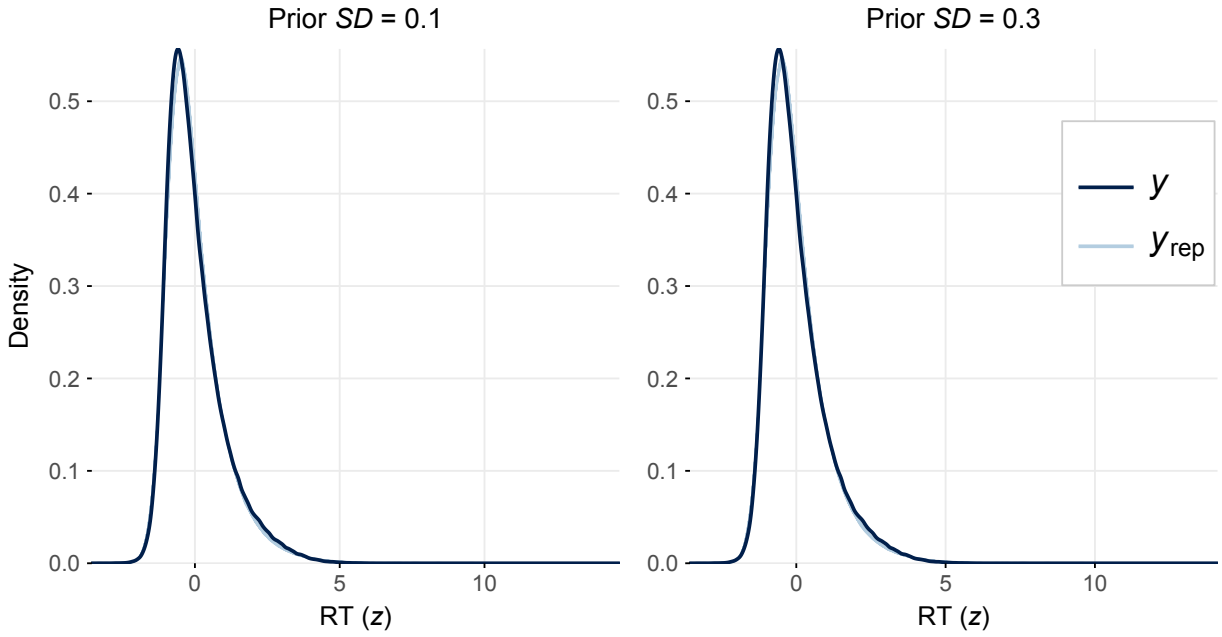

Figure C7
Posterior predictive checks for the (ex-Gaussian) models from the semantic priming study. y = observed data; y_{rep} = predicted data.

Study 2: Semantic decision

Prior predictive checks

Figures C8, C9 and C10 show the prior predictive checks for the Gaussian models (for background on these checks, see Study 1 above). The three plots—corresponding to models that used the default Gaussian distribution—show that the priors fitted the data acceptably but not very well.

In contrast to the results from the Gaussian models, Figures C11, C12 and C13 demonstrate that, when an ex-Gaussian distribution was used, the priors fitted the data far better, which converged with the results found in Study 1.

Posterior predictive checks

Based on the above results, the ex-Gaussian distribution was used in the final models. Figure C14 presents the posterior predictive checks for the latter models. The interpretation of these plots is simple: the distributions of the observed (y) and the

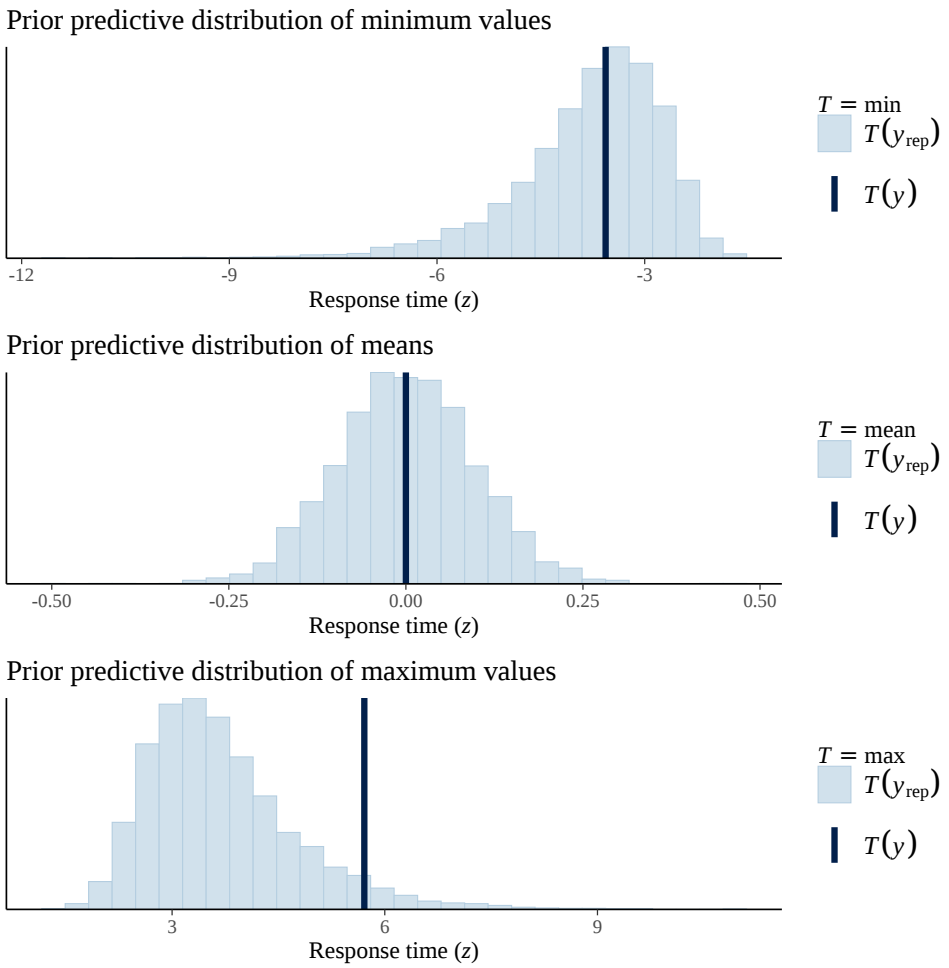

Figure C8

Prior predictive checks for the Gaussian, informative prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

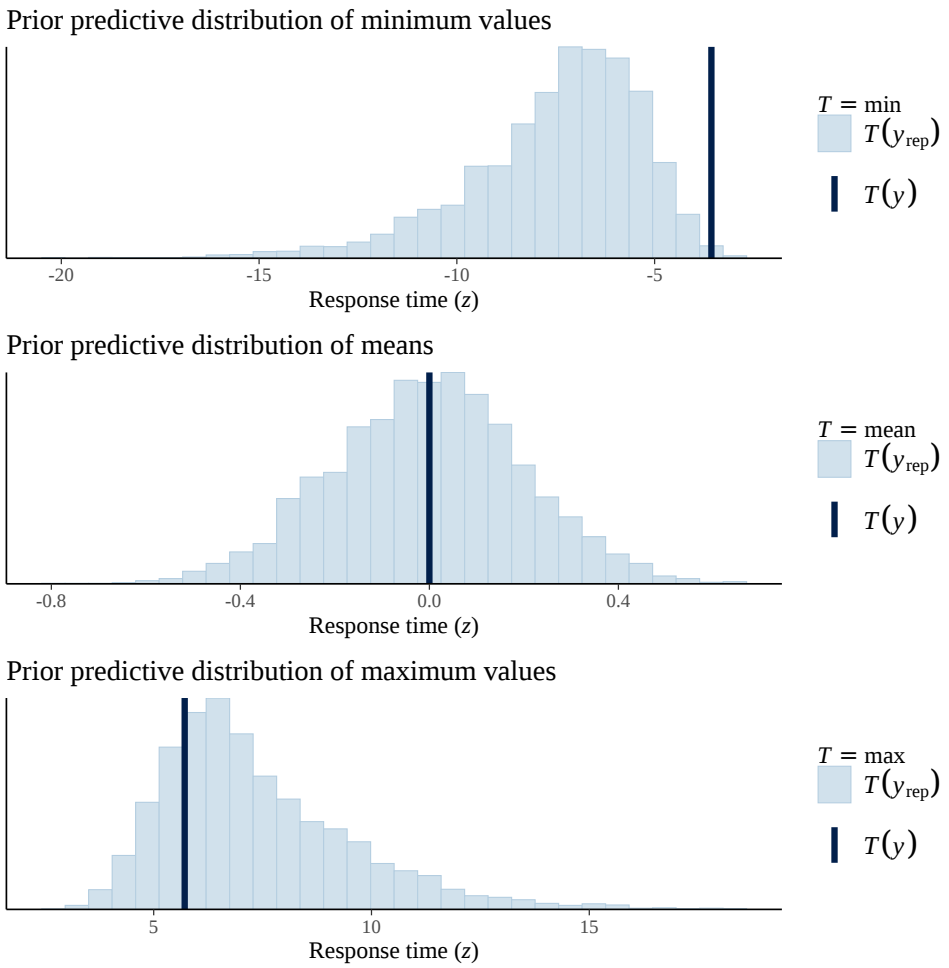

Figure C9

Prior predictive checks for the Gaussian, weakly-informative prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

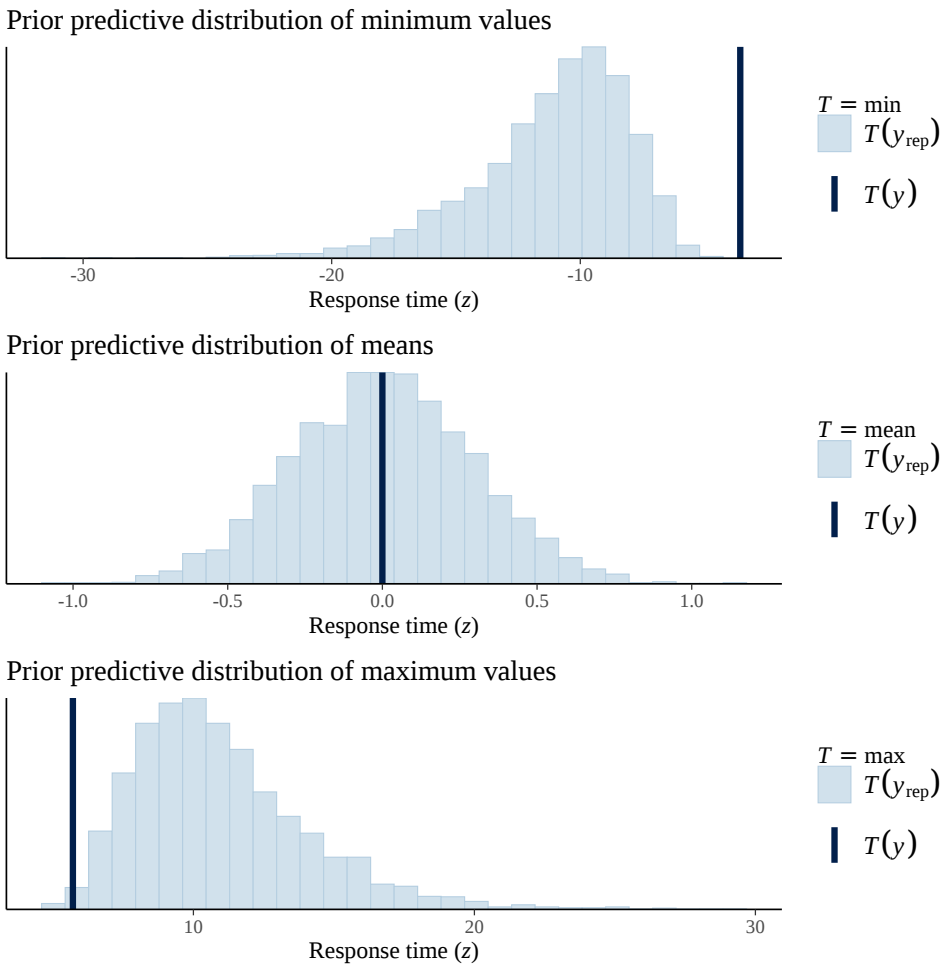

Figure C10

Prior predictive checks for the Gaussian, diffuse prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

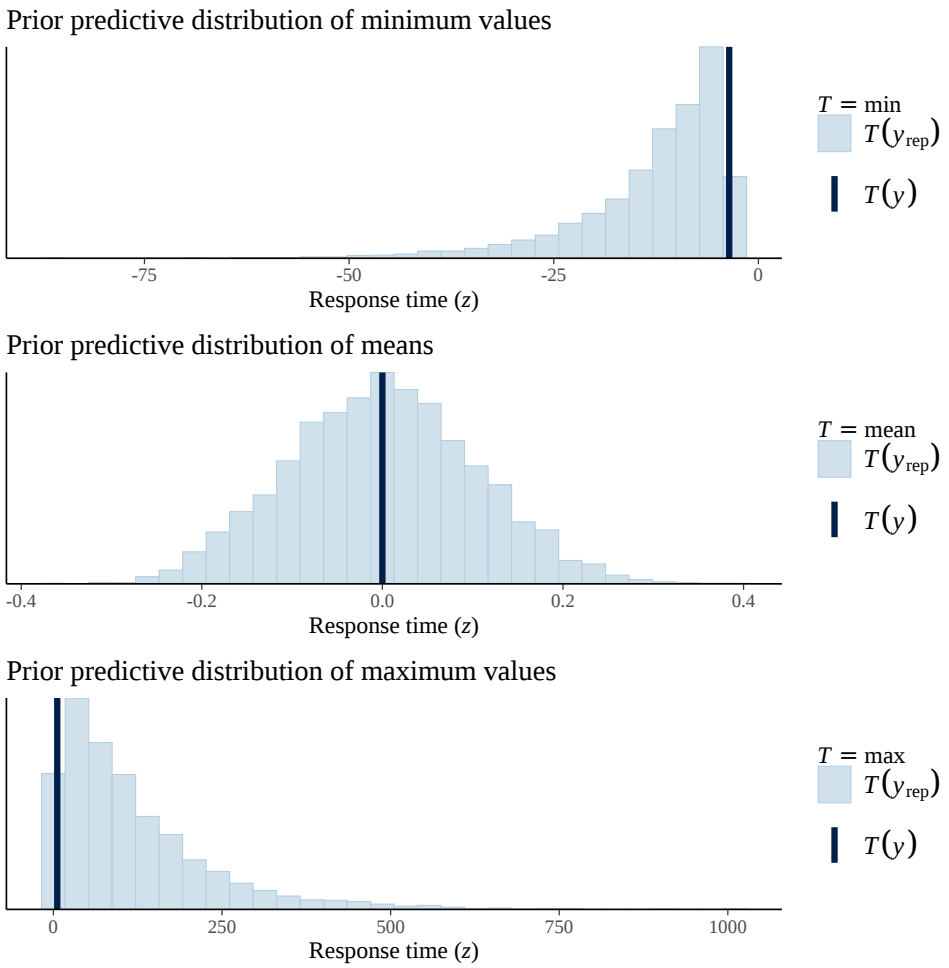

Figure C11

Prior predictive checks for the ex-Gaussian, informative prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

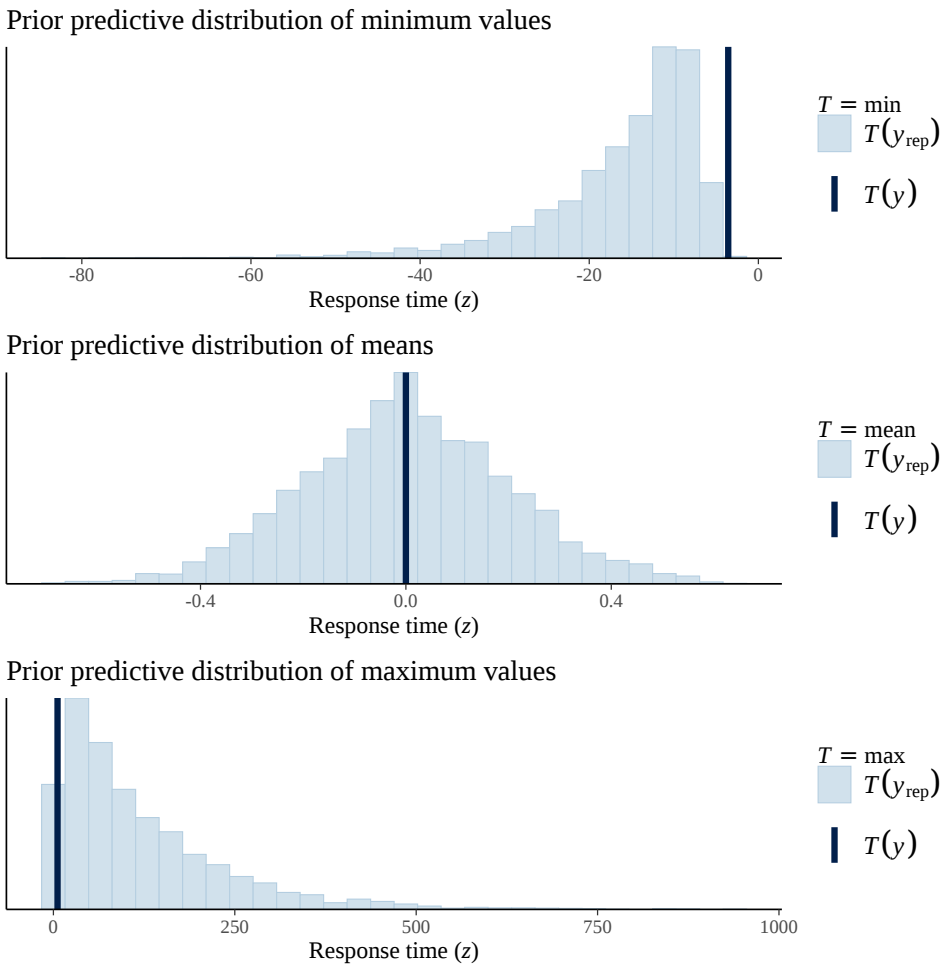

Figure C12

Prior predictive checks for the ex-Gaussian, weakly-informative prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

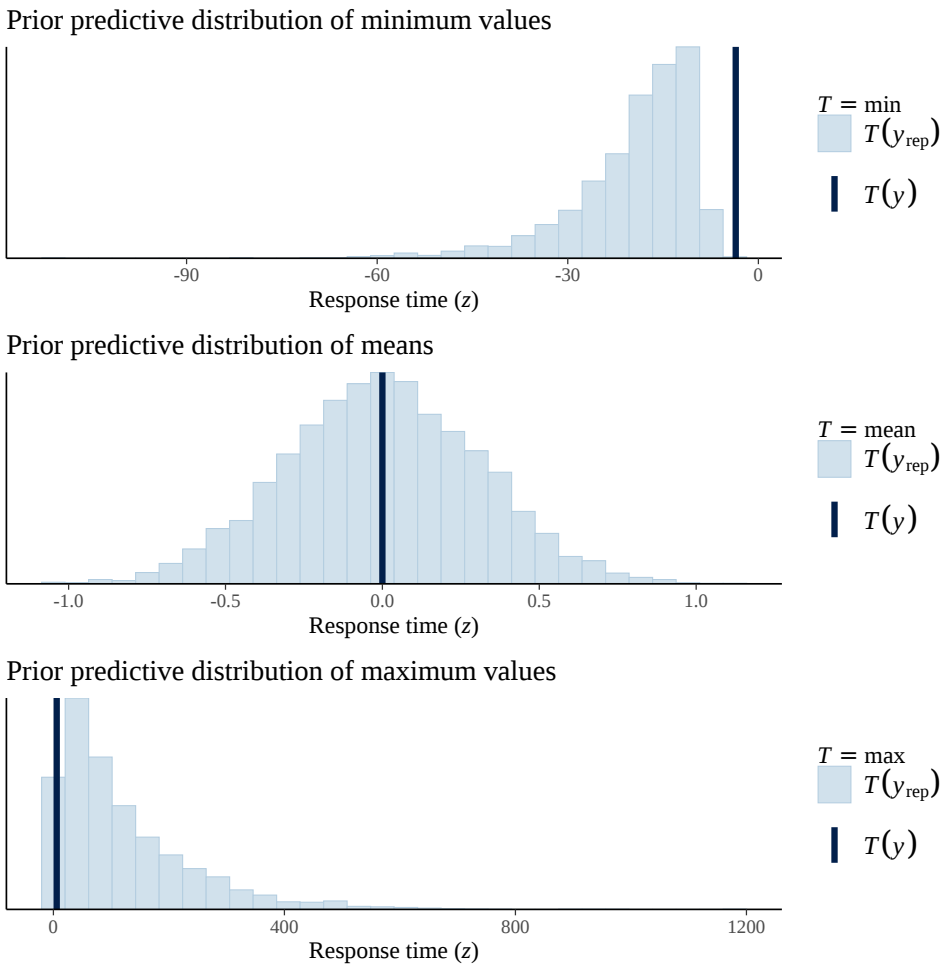

Figure C13

Prior predictive checks for the ex-Gaussian, diffuse prior model from the semantic decision study. y = observed data; y_{rep} = predicted data.

predicted data (y_{rep}) should be as similar as possible. As such, the plots below suggest that the results are not entirely trustworthy. Indeed, the results themselves (Appendix E) are clearly not valid.

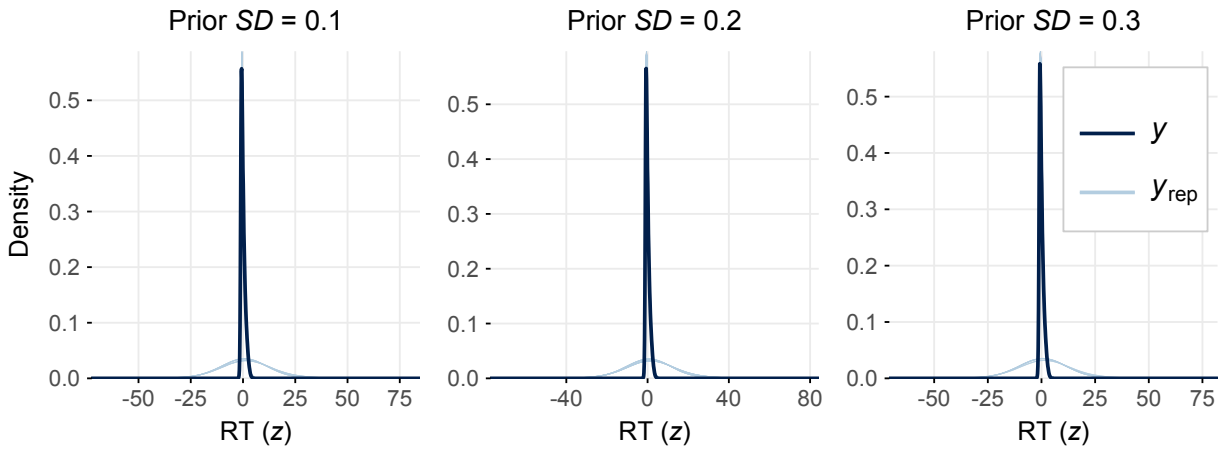

Figure C14
Posterior predictive checks for the (ex-Gaussian) models from the semantic decision study. y = observed data; y_{rep} = predicted data.

Study 3: Lexical decision

Prior predictive checks

Figures C15, C16 and C17 show the prior predictive checks for the Gaussian models (for background on these checks, see Study 1 above). The three plots—corresponding to models that used the default Gaussian distribution—show that the priors fitted the data acceptably but not very well.

In contrast to the results from the Gaussian models, Figures C18, C19 and C20 demonstrate that, when an ex-Gaussian distribution was used, the priors fitted the data far better, which converged with the results found in Studies 1 and 2.

Posterior predictive checks

Based on the above results, the ex-Gaussian distribution was used in the final models. Figure C21 presents the posterior predictive checks for the latter models. The interpretation of these plots is simple: the distributions of the observed (y) and the

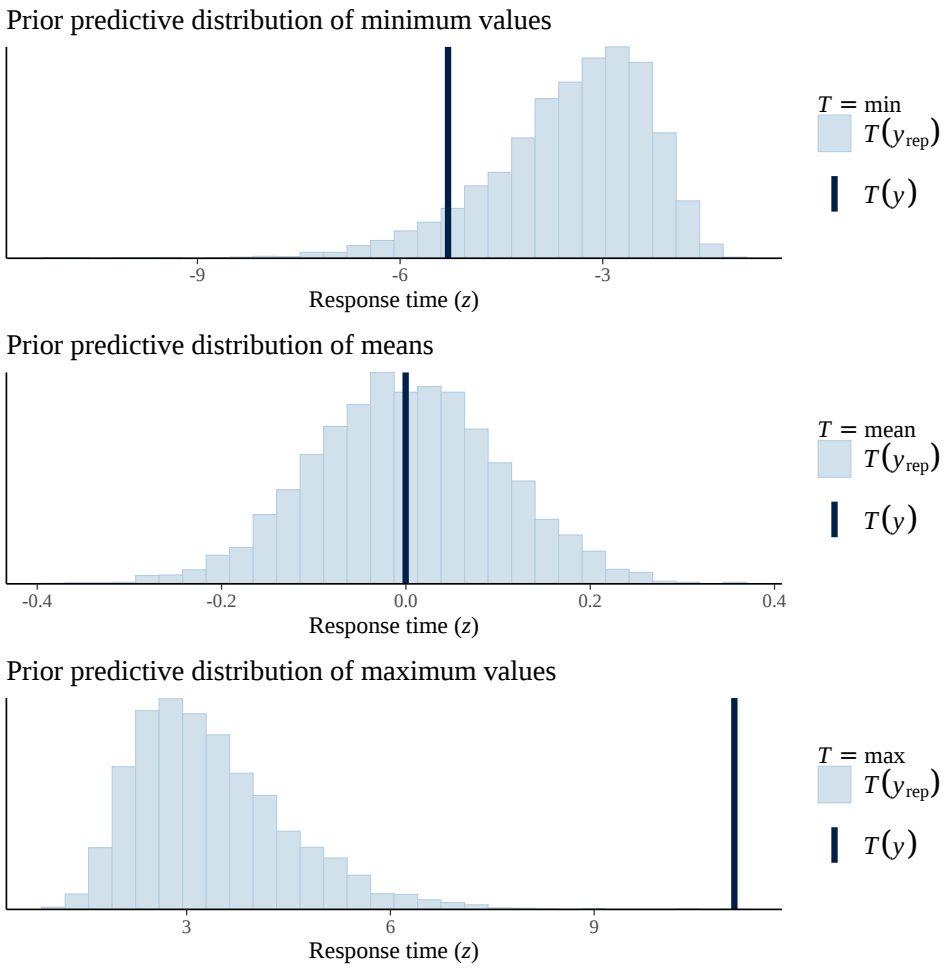

Figure C15

Prior predictive checks for the Gaussian, informative prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

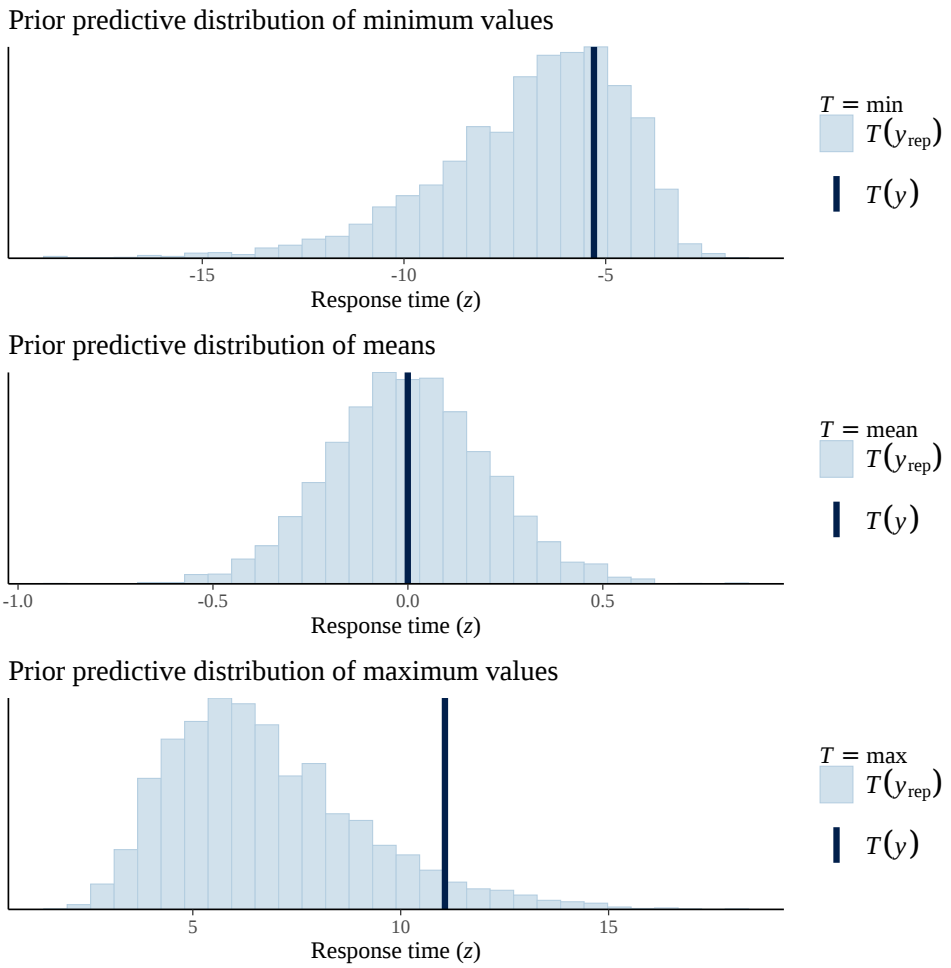

Figure C16

Prior predictive checks for the Gaussian, weakly-informative prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

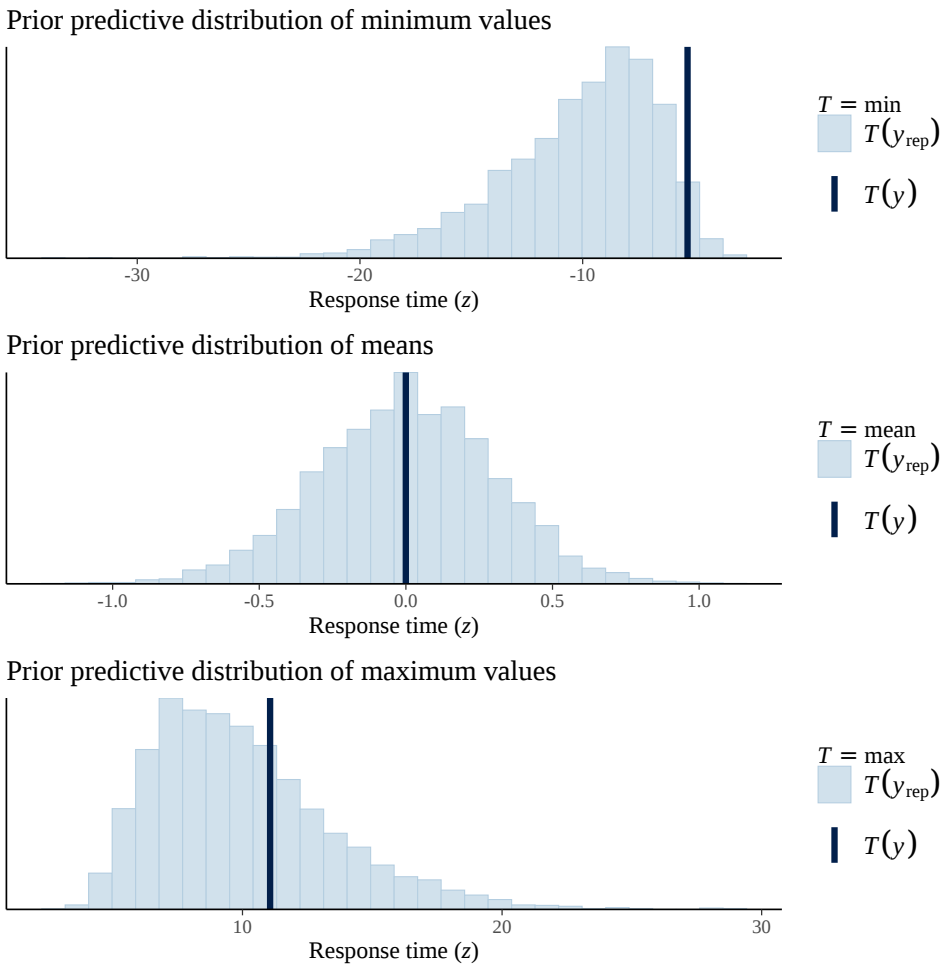

Figure C17

Prior predictive checks for the Gaussian, diffuse prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

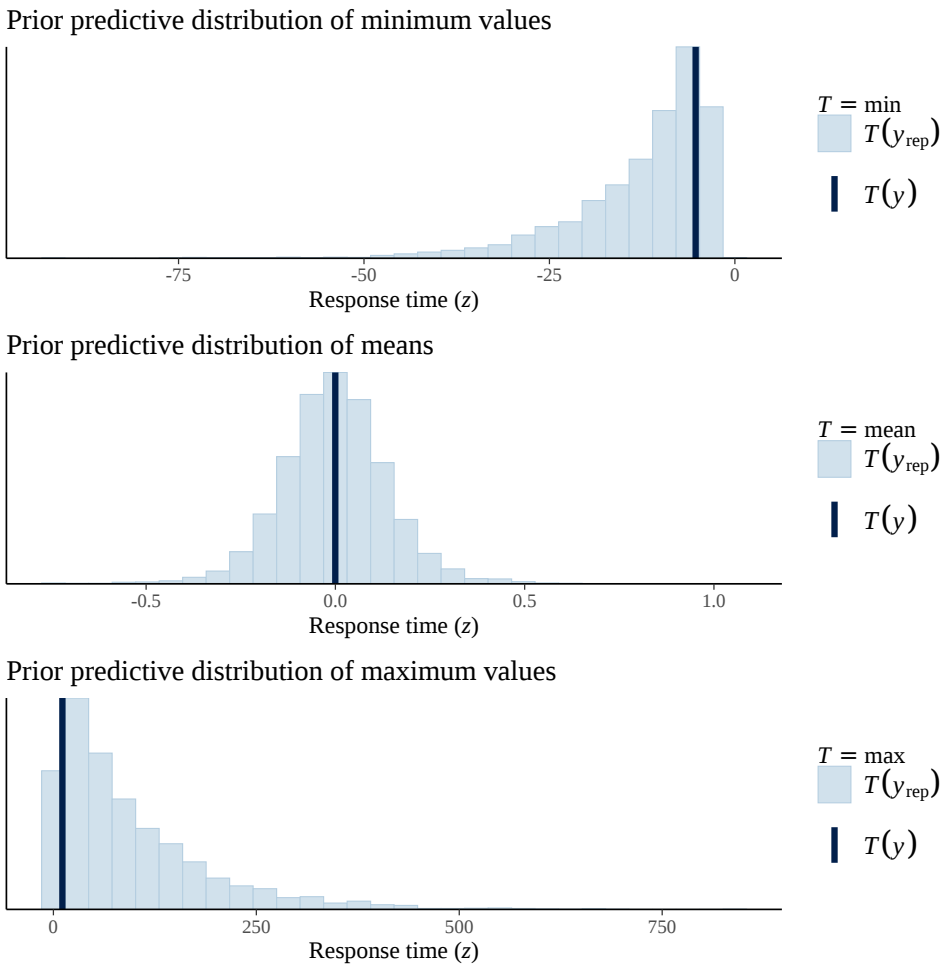

Figure C18

Prior predictive checks for the ex-Gaussian, informative prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

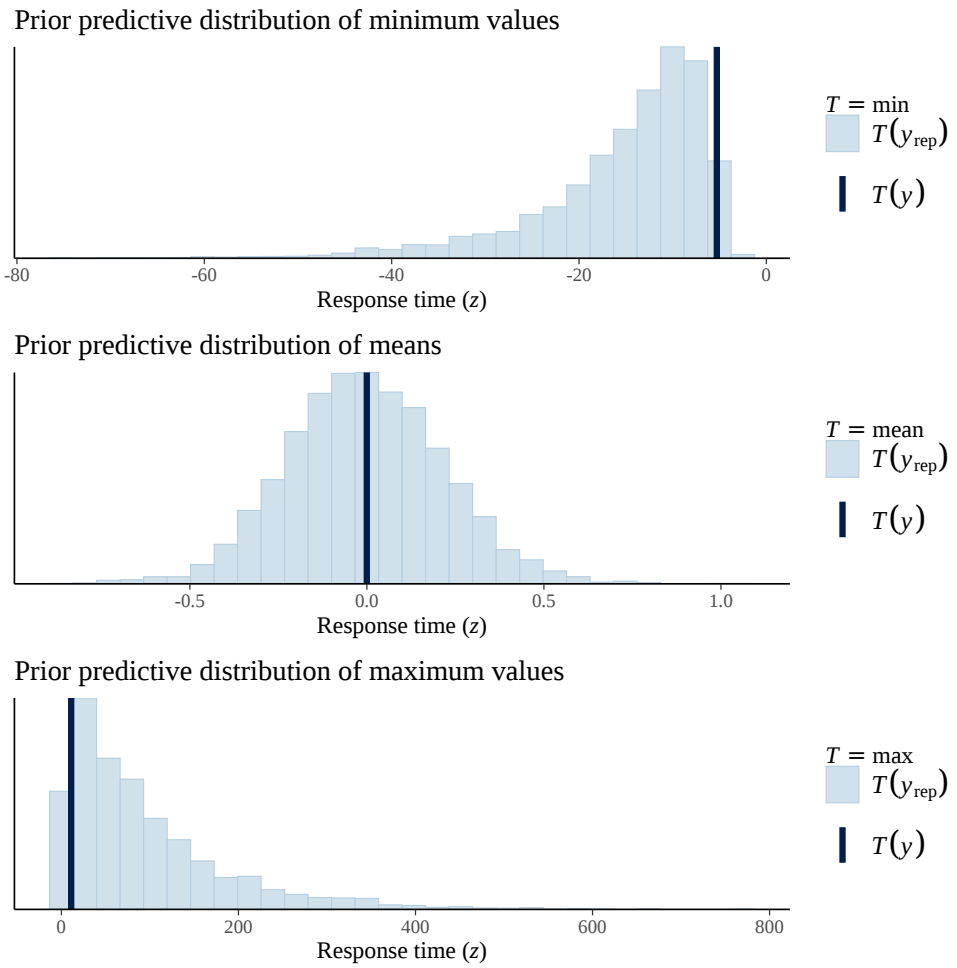

Figure C19

Prior predictive checks for the ex-Gaussian, weakly-informative prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

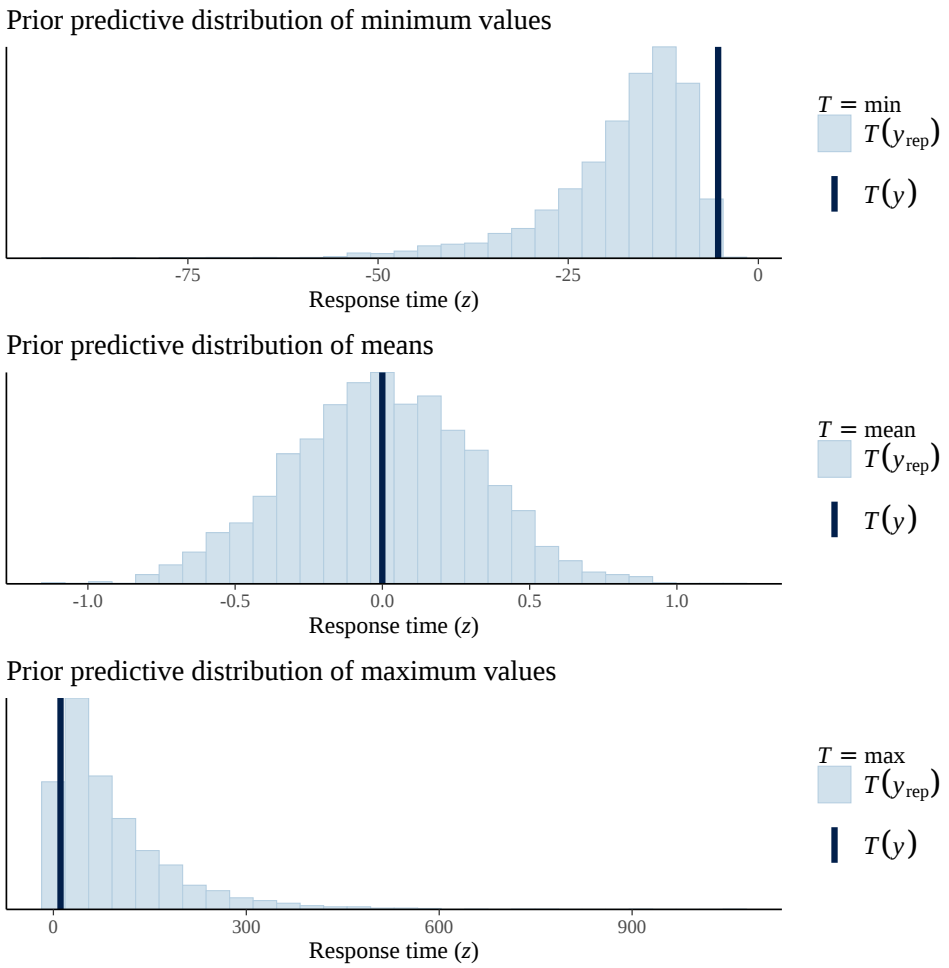

Figure C20

Prior predictive checks for the ex-Gaussian, diffuse prior model from the lexical decision study. y = observed data; y_{rep} = predicted data.

predicted data (y_{rep}) should be as similar as possible. As such, the plots below suggest that the results are trustworthy.

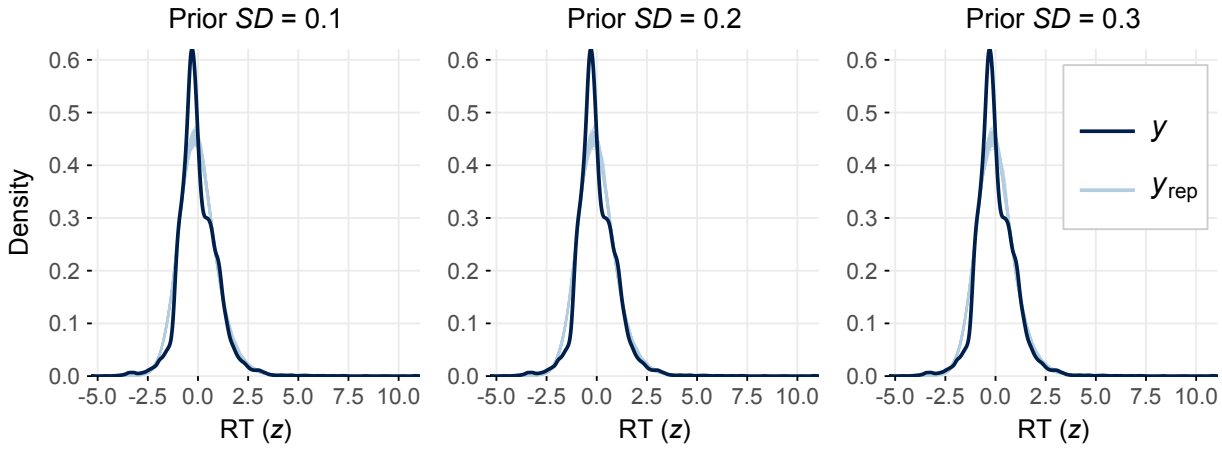

Figure C21

Posterior predictive checks for the (ex-Gaussian) models from the lexical decision study. y = observed data; y_{rep} = predicted data.

Appendix D: Further interaction plots

This appendix presents plots of interactions that were not shown in the main text.

Study 1: Semantic priming

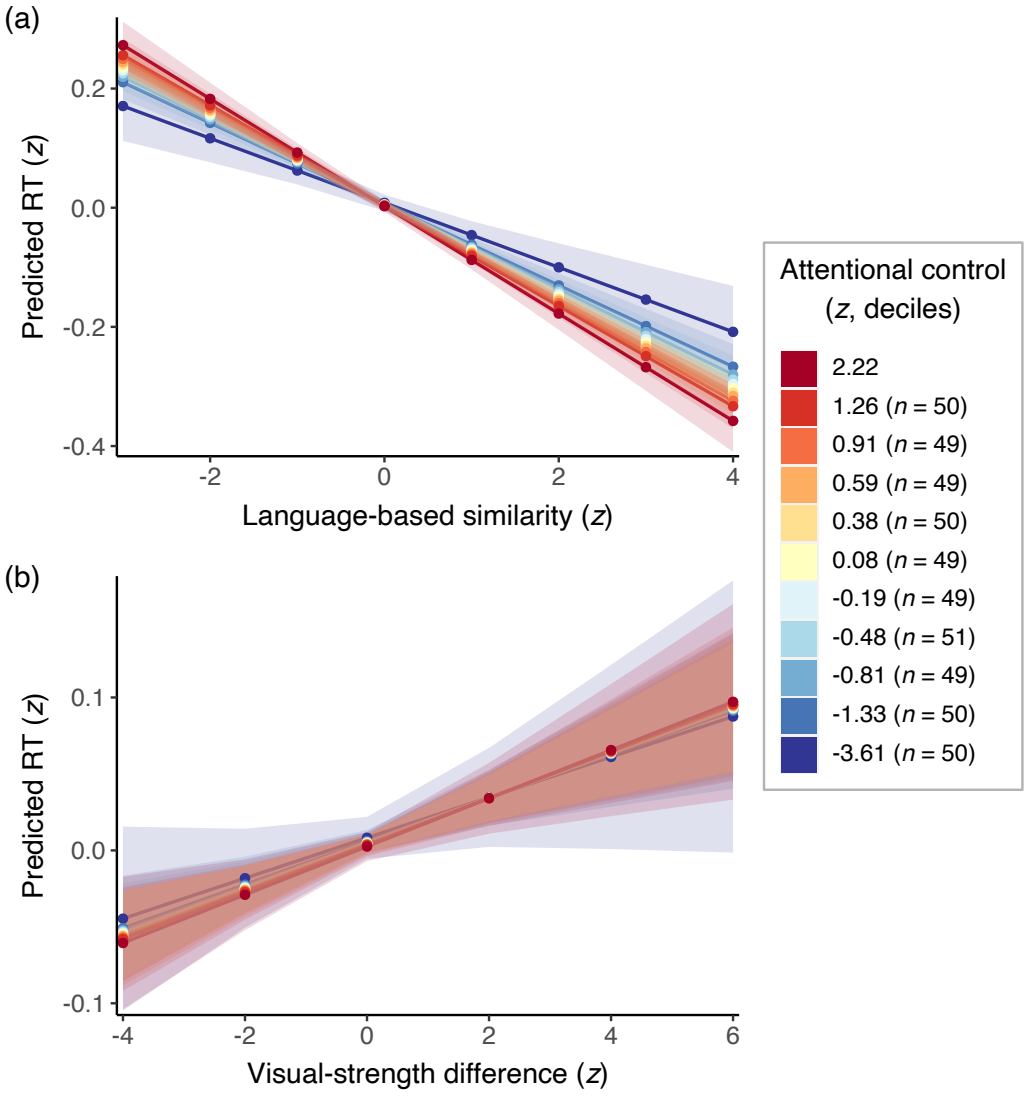

Figure D1

Interactions of attentional control with language-based similarity and visual-strength difference. Attentional control is constrained to deciles (ten sections) in this plot, whereas in the statistical analysis it contained more values within the current range. n = *number of participants contained between decile values.*

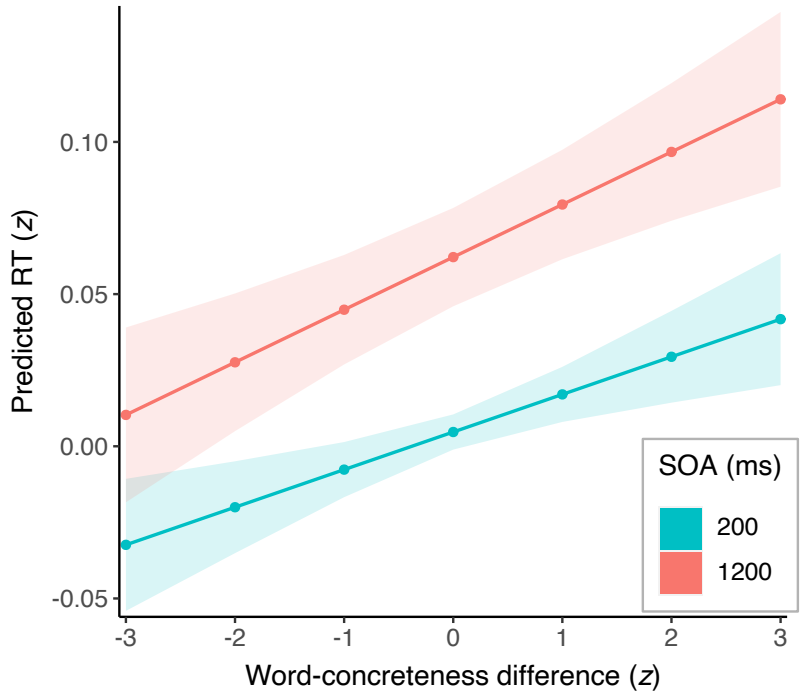

Figure D2

Interaction of stimulus-onset asynchrony (SOA) with word-concreteness difference. SOA was analysed using z-scores, but for clarity, the variable is shown in its basic form here.

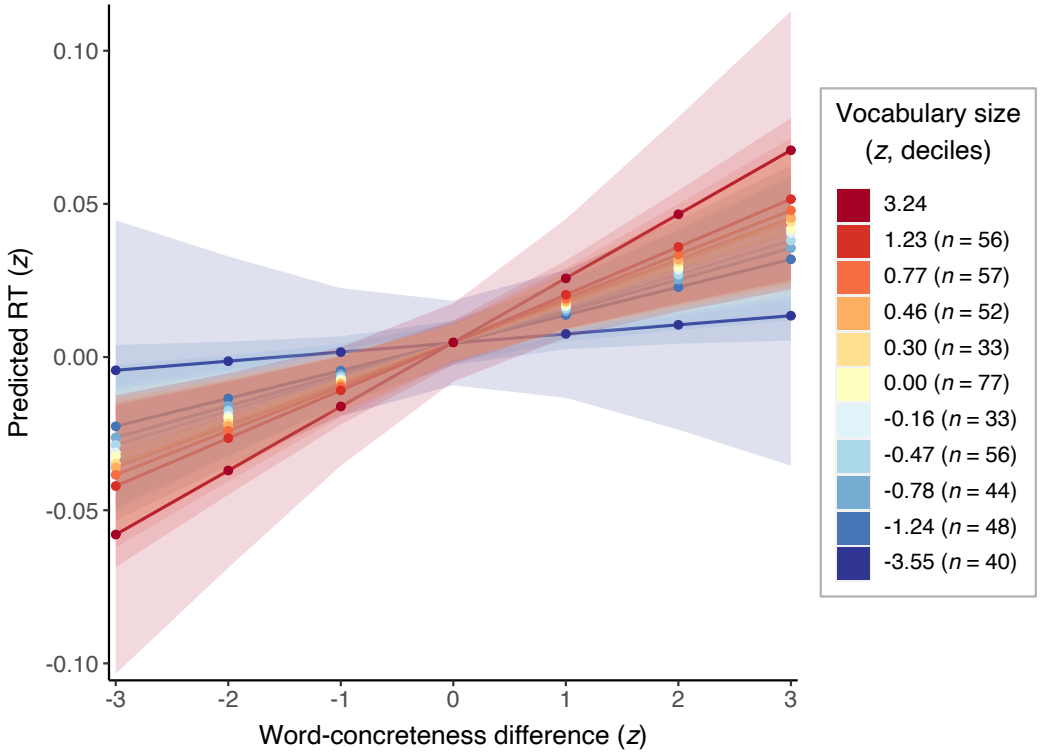

Figure D3

Interaction of word-concreteness difference with vocabulary size. Vocabulary size is constrained to deciles in this plot, whereas in the statistical analysis it contained more values within the current range. n = *number of participants contained between decile values.*

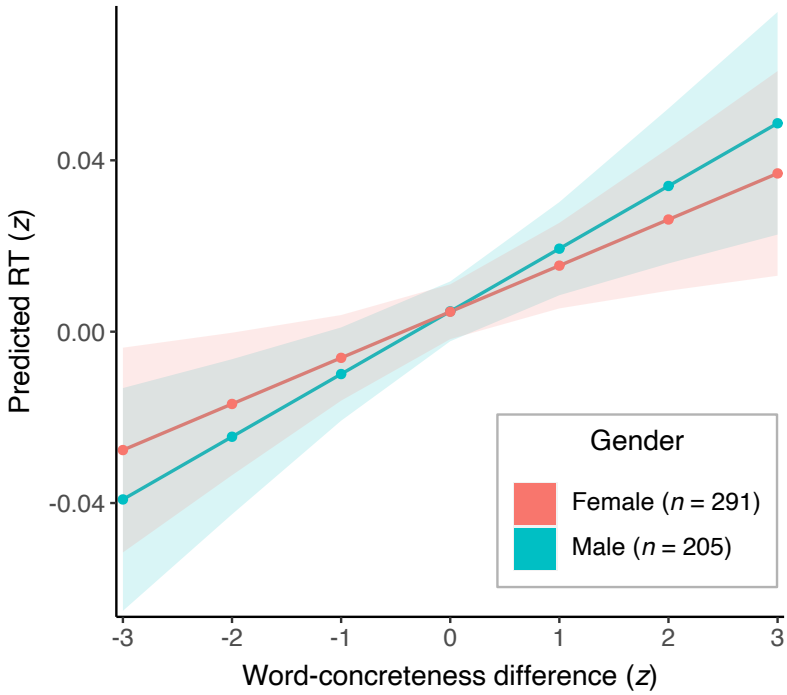

Figure D4

Interaction of word-concreteness difference with gender. Gender was analysed using z-scores, but for clarity, the variable is shown in its basic form here. n *= number of participants contained between decile values.*

Study 2: Semantic decision

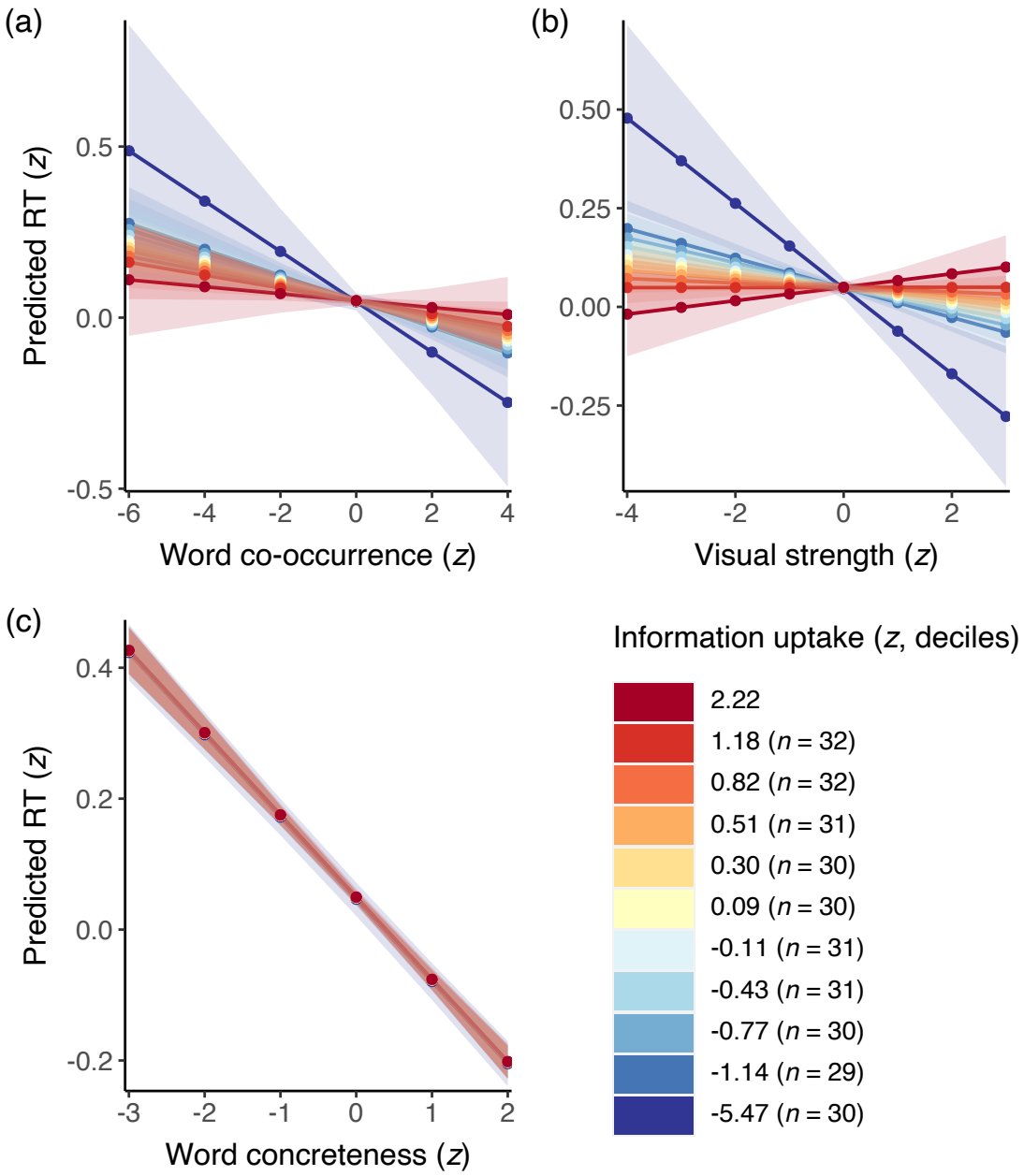

Figure D5

Interactions of information uptake with language-based similarity and visual-strength difference. Information uptake is constrained to deciles in this plot, whereas in the statistical analysis it contained more values within the current range. n *= number of participants contained between decile values.*

Study 3: Lexical decision

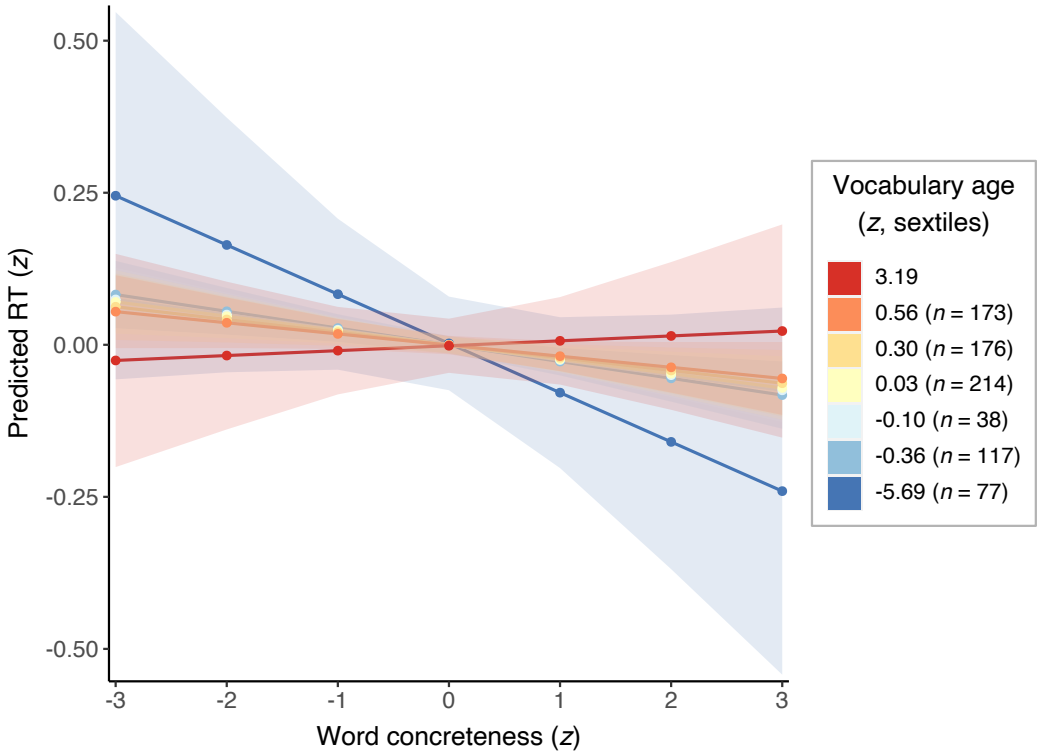

Figure D6

Interaction of word concreteness with vocabulary age. Vocabulary age is constrained to sextiles (six sections) in this plot, whereas in the statistical analysis it contained more values within the current range. n = *number of participants contained between sextile values.*

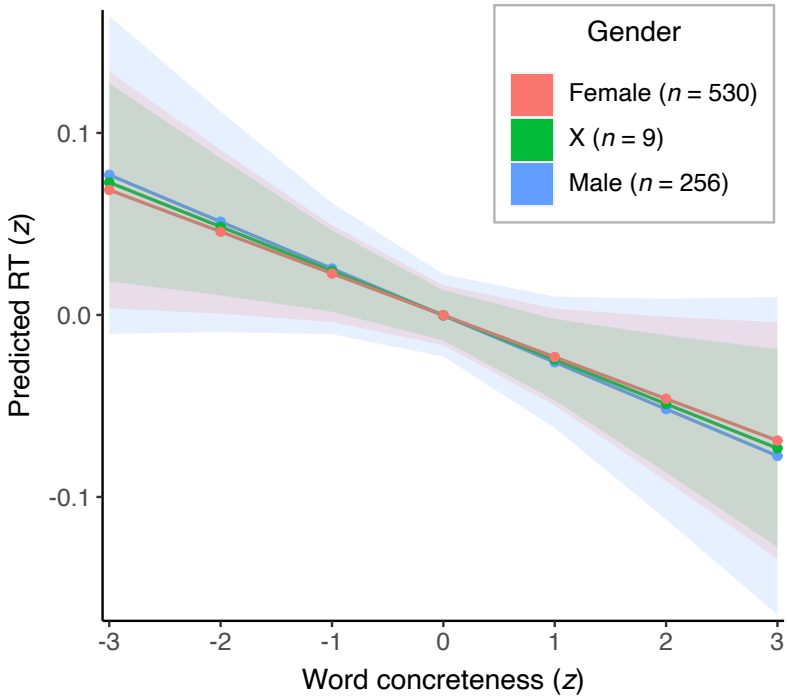

Figure D7

Interaction of word concreteness with gender. Gender was analysed using z-scores, but for clarity, the variable is shown in its basic form here. n *= number of participants contained between sextile values.*

Appendix E: Results from the Bayesian analyses

This appendix presents extended results from the Bayesian analyses, containing a prior sensitivity analysis (Schoot et al., 2021). For each study, three tables are presented that contain the results from the informative prior model ($SD = 0.1$), the weakly-informative prior model ($SD = 0.2$) and the diffuse prior model ($SD = 0.3$). All models had an exponentially modified Gaussian (dubbed 'ex-Gaussian') distribution with an identity link function (for background, see main article and Appendix C). The $\widehat{R}$ value is a convergence diagnostic that should ideally be smaller than 1.01 (Vehtari et al., 2021).

The approach used in this Bayesian analysis is that of estimation (Tendeiro & Kiers, 2019; also see Schmalz et al., 2021). Thus, the estimates were interpreted by considering the position of their credible intervals in relation to the predicted value of RT (z). That is, the closer an interval is to a value of 0 on the predicted RT (z), the smaller the effect of that predictor. For instance, an interval that is symmetrically centred on 0 indicates a very small effect, whereas—in comparison—an interval that does not include 0 indicates a far larger effect (for other examples of this approach, see Milek et al., 2018; Pregla et al., 2021; Rodríguez-Ferreiro et al., 2020).

Study 1: Semantic priming

Table E1 presents the results of the informative prior model, and Table E2 presents those of the diffuse prior model (further to this, the weakly-informative prior model will be added when it finishes running).

Table E1

Informative prior model for the semantic priming study.

	β	*SE*	95% CrI	$\widehat{R}$
(Intercept)	0.00	0.00	[0.00, 0.01]	1.00
Individual differences				
Attentional control	0.00	0.00	[0.00, 0.01]	1.00
Vocabulary size [a]	-0.01	0.00	[-0.01, 0.00]	1.00
Gender [a]	0.00	0.00	[0.00, 0.01]	1.00
Target-word lexical covariates				
Word frequency	-0.11	0.00	[-0.12, -0.11]	1.00
Number of syllables	0.07	0.00	[0.06, 0.07]	1.00
Prime-target semantic relationship				
Word-concreteness difference	0.01	0.00	[0.00, 0.01]	1.00
Language-based similarity [b]	-0.06	0.00	[-0.07, -0.06]	1.00
Visual-strength difference [b]	0.01	0.00	[0.01, 0.02]	1.00
Task condition				
Stimulus-onset asynchrony (SOA) [b]	0.03	0.01	[0.02, 0.04]	1.00
Interactions				
Word-concreteness difference × Vocabulary size	0.00	0.00	[0.00, 0.00]	1.00
Word-concreteness difference × SOA	0.00	0.00	[0.00, 0.00]	1.00
Word-concreteness difference × Gender	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Attentional control	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Attentional control	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Vocabulary size	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Vocabulary size	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Gender	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Gender	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × SOA [b]	0.00	0.00	[0.00, 0.00]	1.00
Visual-strength difference × SOA [b]	0.00	0.00	[0.00, 0.00]	1.00

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.

[b] By-participant random slopes were included for this effect.

Table E2

Diffuse prior model for the semantic priming study.

	β	*SE*	95% CrI	$\widehat{R}$
(Intercept)	0.00	0.00	[0.00, 0.01]	1.00
Individual differences				
Attentional control	0.00	0.00	[0.00, 0.01]	1.00
Vocabulary size [a]	-0.01	0.00	[-0.01, 0.00]	1.00
Gender [a]	0.00	0.00	[0.00, 0.01]	1.00
Target-word lexical covariates				
Word frequency	-0.11	0.00	[-0.12, -0.11]	1.00
Number of syllables	0.07	0.00	[0.06, 0.07]	1.00
Prime-target semantic relationship				
Word-concreteness difference	0.01	0.00	[0.00, 0.01]	1.00
Language-based similarity [b]	-0.06	0.00	[-0.07, -0.06]	1.00
Visual-strength difference [b]	0.01	0.00	[0.01, 0.01]	1.00
Task condition				
Stimulus-onset asynchrony (SOA) [b]	0.03	0.01	[0.02, 0.04]	1.00
Interactions				
Word-concreteness difference × Vocabulary size	0.00	0.00	[0.00, 0.00]	1.00
Word-concreteness difference × SOA	0.00	0.00	[0.00, 0.00]	1.00
Word-concreteness difference × Gender	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Attentional control	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Attentional control	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Vocabulary size	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Vocabulary size	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × Gender	0.00	0.00	[-0.01, 0.00]	1.00
Visual-strength difference × Gender	0.00	0.00	[0.00, 0.00]	1.00
Language-based similarity × SOA [b]	0.00	0.00	[0.00, 0.00]	1.00
Visual-strength difference × SOA [b]	0.00	0.00	[0.00, 0.00]	1.00

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.

[b] By-participant random slopes were included for this effect.

Figure E1 presents the posterior distribution of each effect in each model. The frequentist estimates are also shown to facilitate the comparison.

Study 2: Semantic decision

Table E3 presents the results of the informative prior model, Table E4 presents those of the weakly-informative prior model and Table E5 presents those of the diffuse prior model.

Table E3
Informative prior model for the semantic decision study.

	β	*SE*	95% CrI	$\widehat{R}$
(Intercept)	0.14	0.42	[0.00, 1.72]	1.31
Individual differences				
Information uptake	0.03	0.08	[-0.01, 0.31]	1.31
Vocabulary size [a]	0.18	0.46	[0.00, 1.44]	1.31
Gender [a]	-0.12	0.39	[-1.56, 0.02]	1.31
Lexical covariates				
Word frequency	-0.18	0.31	[-1.34, -0.07]	1.30
Orthographic Levenshtein distance	0.06	0.56	[-1.14, 1.94]	1.41
Semantic variables				
Word concreteness	0.00	0.26	[-0.08, 1.01]	1.30
Word co-occurrence [b]	-0.05	0.23	[-0.87, 0.40]	1.41
Visual strength [b]	-0.20	0.49	[-1.52, -0.01]	1.31
Interactions				
Word concreteness × Vocabulary size	0.02	0.55	[-1.24, 1.83]	1.42
Word concreteness × Gender	0.07	0.40	[-0.31, 1.58]	1.42
Word co-occurrence × Information uptake	-0.06	0.19	[-0.70, 0.02]	1.31
Visual strength × Information uptake	-0.15	0.46	[-1.79, 0.02]	1.30
Word co-occurrence × Vocabulary size	-0.04	0.55	[-1.92, 1.11]	1.42
Visual strength × Vocabulary size	0.15	0.38	[0.00, 1.27]	1.30
Word co-occurrence × Gender	0.00	0.26	[-0.78, 0.68]	1.41
Visual strength × Gender	0.18	0.49	[-0.01, 1.66]	1.30

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

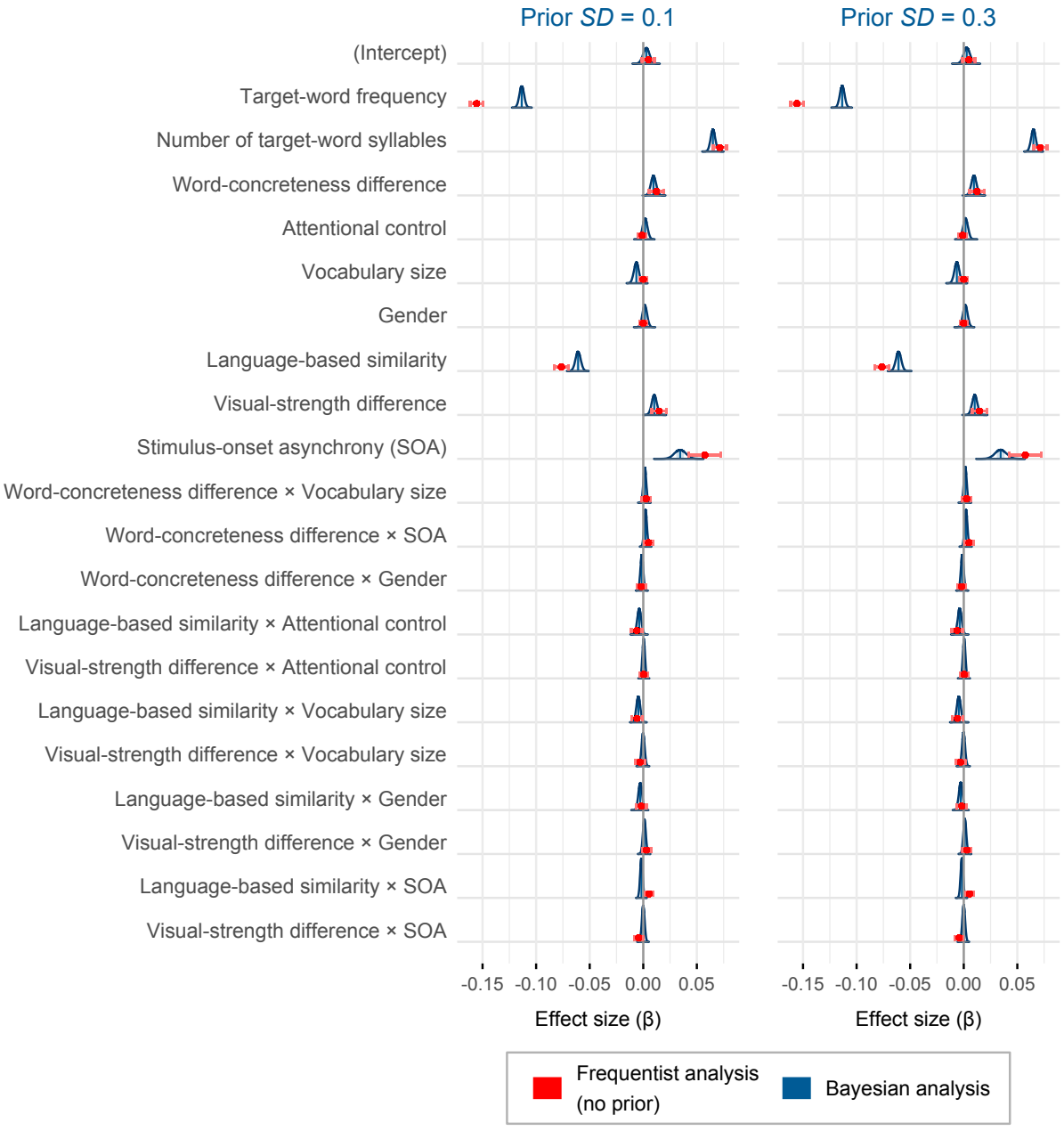

Figure E1

Estimates from the frequentist analysis (in red) and from the Bayesian analysis (in blue) for the semantic priming study, in each model. The frequentist means (represented by points) are flanked by 95% confidence intervals. The Bayesian means (represented by vertical lines) are flanked by 95% credible intervals, in light blue (in some cases, the interval is covered up by the bar of the mean).

Table E4

Weakly-informative prior model for the semantic decision study.

	β	SE	95% CrI	$\hat{R}$
(Intercept)	0.14	0.42	[0.00, 1.72]	1.31
Individual differences				
Information uptake	0.03	0.08	[-0.01, 0.31]	1.31
Vocabulary size [a]	0.18	0.46	[0.00, 1.44]	1.31
Gender [a]	-0.12	0.39	[-1.56, 0.02]	1.30
Lexical covariates				
Word frequency	-0.18	0.31	[-1.34, -0.07]	1.31
Orthographic Levenshtein distance	0.06	0.56	[-1.14, 1.94]	1.40
Semantic variables				
Word concreteness	0.00	0.26	[-0.08, 1.01]	1.30
Word co-occurrence [b]	-0.05	0.23	[-0.87, 0.40]	1.41
Visual strength [b]	-0.20	0.49	[-1.52, -0.01]	1.31
Interactions				
Word concreteness × Vocabulary size	0.02	0.55	[-1.24, 1.83]	1.41
Word concreteness × Gender	0.07	0.40	[-0.31, 1.58]	1.42
Word co-occurrence × Information uptake	-0.06	0.19	[-0.70, 0.02]	1.31
Visual strength × Information uptake	-0.15	0.46	[-1.79, 0.02]	1.31
Word co-occurrence × Vocabulary size	-0.04	0.55	[-1.92, 1.11]	1.42
Visual strength × Vocabulary size	0.15	0.38	[0.00, 1.28]	1.30
Word co-occurrence × Gender	0.00	0.26	[-0.78, 0.68]	1.41
Visual strength × Gender	0.18	0.49	[-0.01, 1.66]	1.31

Note. β = Estimate based on z-scored variables; SE = standard error; CrI = credible interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.

[b] By-participant random slopes were included for this effect.

Table E5
Diffuse prior model for the semantic decision study.

	β	*SE*	95% CrI	$\hat{R}$
(Intercept)	0.14	0.42	[0.00, 1.72]	1.31
Individual differences				
Information uptake	0.03	0.08	[-0.01, 0.31]	1.30
Vocabulary size [a]	0.18	0.46	[0.00, 1.44]	1.30
Gender [a]	-0.12	0.39	[-1.56, 0.02]	1.31
Lexical covariates				
Word frequency	-0.18	0.31	[-1.34, -0.07]	1.30
Orthographic Levenshtein distance	0.06	0.56	[-1.14, 1.94]	1.41
Semantic variables				
Word concreteness	0.00	0.26	[-0.08, 1.01]	1.30
Word co-occurrence [b]	-0.05	0.23	[-0.87, 0.40]	1.41
Visual strength [b]	-0.20	0.49	[-1.52, -0.01]	1.31
Interactions				
Word concreteness × Vocabulary size	0.02	0.55	[-1.24, 1.83]	1.41
Word concreteness × Gender	0.07	0.40	[-0.31, 1.58]	1.41
Word co-occurrence × Information uptake	-0.06	0.19	[-0.70, 0.02]	1.31
Visual strength × Information uptake	-0.15	0.46	[-1.79, 0.02]	1.30
Word co-occurrence × Vocabulary size	-0.04	0.55	[-1.92, 1.11]	1.41
Visual strength × Vocabulary size	0.15	0.38	[0.00, 1.27]	1.30
Word co-occurrence × Gender	0.00	0.26	[-0.78, 0.68]	1.42
Visual strength × Gender	0.18	0.49	[-0.01, 1.66]	1.30

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates. Some interactions are split over two lines, with the second line indented.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Figure E2 presents the posterior distribution of each effect in each model. The frequentist estimates are also shown to facilitate the comparison.

Study 3: Lexical decision

Table E6 presents the results of the informative prior model, Table E7 presents those of the weakly-informative prior model and Table E8 presents those of the diffuse prior model.

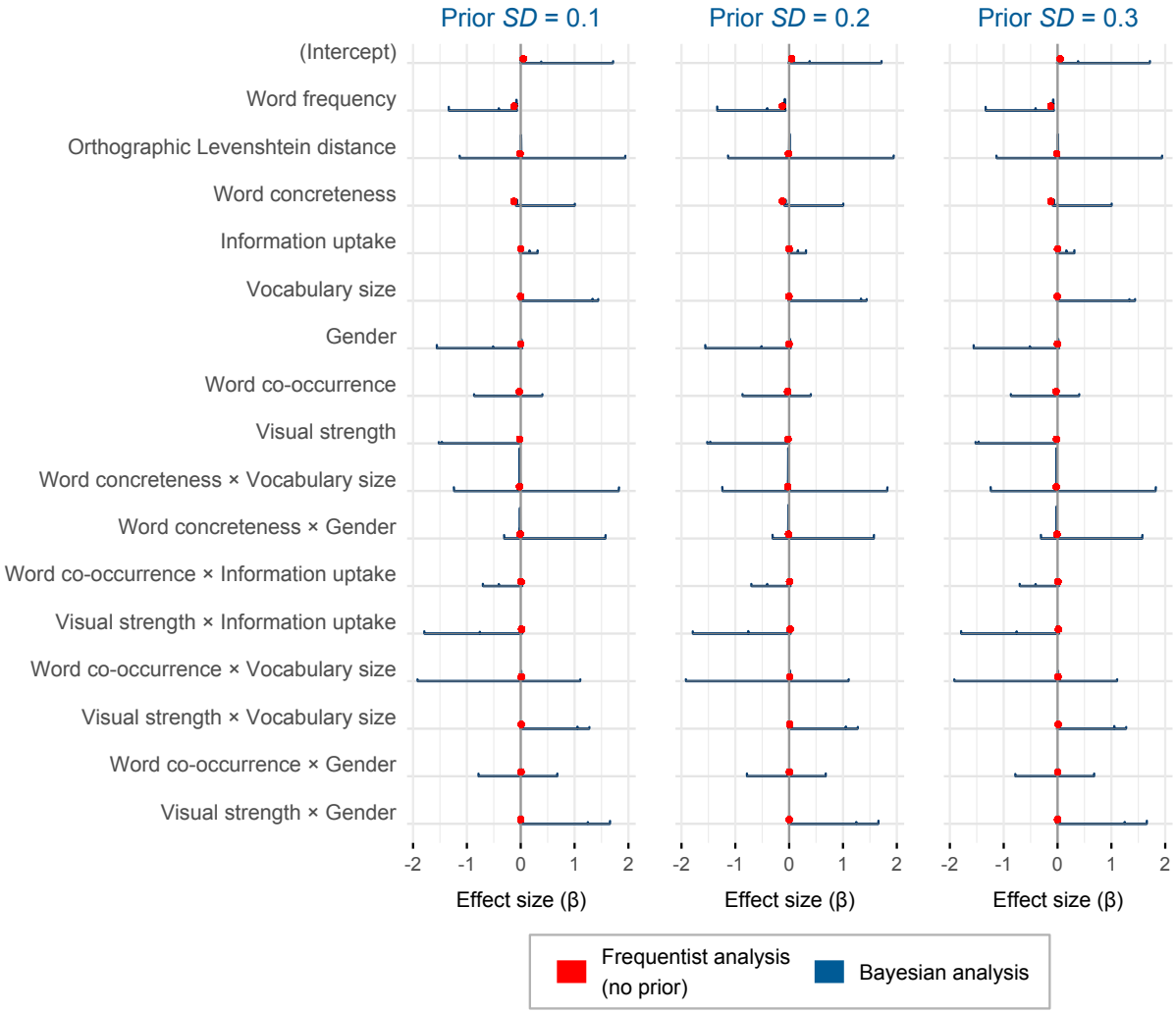

Figure E2

Estimates from the frequentist analysis (in red) and from the Bayesian analysis (in blue) for the semantic decision study, in each model. The frequentist means (represented by points) are flanked by 95% confidence intervals. The Bayesian means (represented by vertical lines) are flanked by 95% credible intervals, in light blue (in some cases, the interval is covered up by the bar of the mean).

Table E6

Informative prior model for the lexical decision study.

	β	SE	95% CrI	$\widehat{R}$
(Intercept)	0.00	0.01	[-0.01, 0.01]	1.00
Individual differences				
Vocabulary age [a]	0.00	0.01	[-0.01, 0.02]	1.00
Gender [a]	0.00	0.01	[-0.01, 0.01]	1.00
Lexical covariate				
Orthographic Levenshtein distance [b]	0.15	0.01	[0.13, 0.17]	1.00
Semantic variables				
Word concreteness [b]	-0.03	0.01	[-0.05, -0.02]	1.00
Word frequency [b]	-0.14	0.01	[-0.16, -0.12]	1.00
Visual strength [b]	-0.01	0.01	[-0.02, 0.01]	1.00
Interactions				
Word concreteness × Vocabulary age	0.01	0.01	[-0.01, 0.03]	1.00
Word concreteness × Gender	0.01	0.01	[-0.01, 0.03]	1.00
Word frequency × Vocabulary age	0.00	0.01	[-0.02, 0.02]	1.00
Visual strength × Vocabulary age	0.00	0.01	[-0.02, 0.01]	1.00
Word frequency × Gender	-0.01	0.01	[-0.03, 0.01]	1.00
Visual strength × Gender	-0.01	0.01	[-0.02, 0.01]	1.00

Note. β = Estimate based on z-scored variables; SE = standard error; CrI = credible interval. Shaded rows contain covariates.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Table E7

Weakly-informative prior model for the lexical decision study.

	β	*SE*	95% CrI	$\widehat{R}$
(Intercept)	0.00	0.01	[-0.01, 0.01]	1.00
Individual differences				
Vocabulary age [a]	0.00	0.01	[-0.01, 0.02]	1.00
Gender [a]	0.00	0.01	[-0.01, 0.01]	1.00
Lexical covariate				
Orthographic Levenshtein distance [b]	0.15	0.01	[0.13, 0.17]	1.00
Semantic variables				
Word concreteness [b]	-0.03	0.01	[-0.05, -0.02]	1.00
Word frequency [b]	-0.14	0.01	[-0.16, -0.12]	1.00
Visual strength [b]	-0.01	0.01	[-0.02, 0.01]	1.00
Interactions				
Word concreteness × Vocabulary age	0.01	0.01	[-0.01, 0.03]	1.00
Word concreteness × Gender	0.01	0.01	[-0.01, 0.03]	1.00
Word frequency × Vocabulary age	0.00	0.01	[-0.02, 0.02]	1.00
Visual strength × Vocabulary age	0.00	0.01	[-0.02, 0.01]	1.00
Word frequency × Gender	-0.01	0.01	[-0.03, 0.01]	1.00
Visual strength × Gender	-0.01	0.01	[-0.02, 0.01]	1.00

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Table E8
Diffuse prior model for the lexical decision study.

	β	*SE*	95% CrI	$\widehat{R}$
(Intercept)	0.00	0.01	[-0.01, 0.01]	1.00
Individual differences				
Vocabulary age [a]	0.00	0.01	[-0.01, 0.02]	1.00
Gender [a]	0.00	0.01	[-0.01, 0.01]	1.00
Lexical covariate				
Orthographic Levenshtein distance [b]	0.15	0.01	[0.13, 0.17]	1.00
Semantic variables				
Word concreteness [b]	-0.03	0.01	[-0.05, -0.02]	1.00
Word frequency [b]	-0.14	0.01	[-0.16, -0.12]	1.00
Visual strength [b]	-0.01	0.01	[-0.02, 0.01]	1.00
Interactions				
Word concreteness × Vocabulary age	0.01	0.01	[-0.01, 0.03]	1.00
Word concreteness × Gender	0.01	0.01	[-0.01, 0.03]	1.00
Word frequency × Vocabulary age	0.00	0.01	[-0.02, 0.02]	1.00
Visual strength × Vocabulary age	0.00	0.01	[-0.02, 0.01]	1.00
Word frequency × Gender	-0.01	0.01	[-0.03, 0.01]	1.00
Visual strength × Gender	-0.01	0.01	[-0.02, 0.01]	1.00

Note. β = Estimate based on z-scored variables; *SE* = standard error; CrI = credible interval. Shaded rows contain covariates.

[a] By-word random slopes were included for this effect.
[b] By-participant random slopes were included for this effect.

Figure E3 presents the posterior distribution of each effect in each model. The frequentist estimates are also shown to facilitate the comparison.

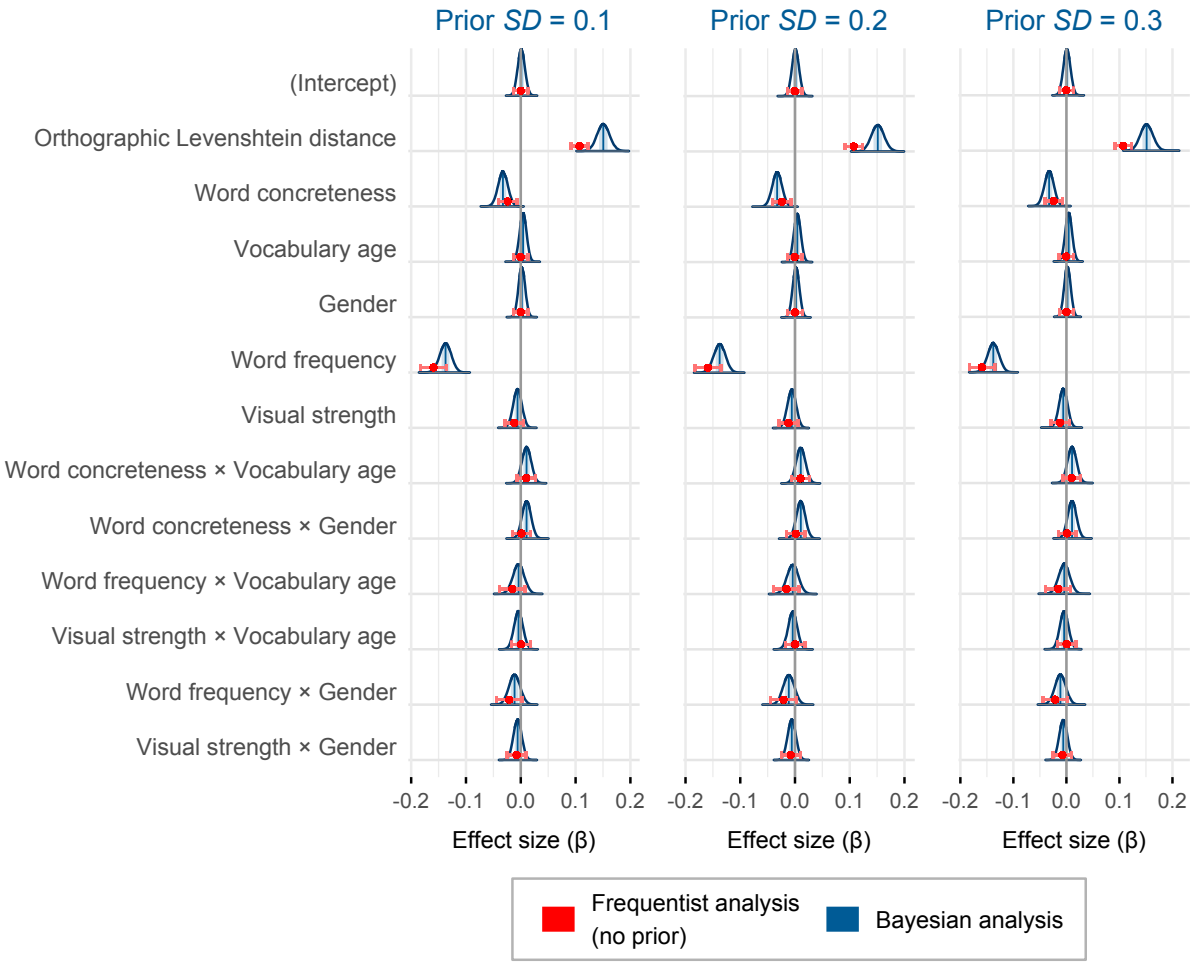

Figure E3

Estimates from the frequentist analysis (in red) and from the Bayesian analysis (in blue) for the lexical decision study, in each model. The frequentist means (represented by points) are flanked by 95% confidence intervals. The Bayesian means (represented by vertical lines) are flanked by 95% credible intervals, in light blue (in some cases, the interval is covered up by the bar of the mean).