The processing of orthography in visual word recognition and its sensitivity to top-down modulation

Jack E. Taylor

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Psychology and Neuroscience
College of Science and Engineering
University of Glasgow

July 2022

Word, kid! Get your ticket from the telepath;
"Wickit, wickit, wickit", on electroencephalograph.

-MF DOOM

Abstract

A vital component of visual word recognition is the decoding of *orthography*, the rules by which language is transcribed from and to visual script. Literate humans demonstrate considerable consistency in the timing and localisation of orthographic processing in the brain, with an early occipitotemporal response showing robust sensitivity to orthographic information as early as 150 - 200 ms post-stimulus. It has been proposed that, consistent with mechanisms involved in other visual perceptual processes, orthographic processing is sensitive to higher-level information provided via top-down inputs. In this thesis, I investigate the degree to which early orthographic processing is modulated by higher-level expectations for word forms over unpredicted word forms that vary in their predictability. I focus on the N1 event-related potential component observed in electroencephalography (EEG). Peaking around 170 ms, this component has shown consistent sensitivity to orthographic information.

I present evidence from two EEG experiments probing the effect of predictions on orthographic processing. In the first of these experiments, I examine the interaction between task (lexical decision, semantic categorisation) and stimulus (category-relevant words, category-irrelevant words, pseudowords, nonwords). I replicate findings of sensitivity to orthography in the N1, and, consistent with previous research, find evidence for a general effect of task on processing during the N1. However, I observe a lack of *selective* sensitivity for category-relevant word forms in the semantic categorisation task, where such a finding would advocate category-level top-down modulation of the N1. I argue that a sensitivity to higher-level predictions in orthographic processing would require a transcoding of information from semantic to orthographic representations, which would be necessarily computationally lossy and entail a loss of specificity in predictions. As a result, selective sensitivity to predicted word forms may only be expected when predictions are more targeted, such that they maximise the specificity of any predictions for orthographic input.

In the second EEG experiment I show that, indeed, when predictions are more targeted, for specific word forms, an effect of prediction is observed in the N1. I employ a picture-word verification paradigm to induce participants to generate strong predictions for upcoming words. I show an interaction between picture-word congruency and predictability, where predictability negatively predicts N1 amplitudes for picture-congruent words, and positively predicts N1 amplitudes for picture-incongruent words. I argue that these findings are inconsistent with typical predictive coding accounts, in which predicted orthographic information is "explained away" such that activity scales with prediction error, but support an account in which top-down modulation results in a "sharpened" sensitivity to predicted orthographic features, such that activity scales with prediction congruency. I suggest that the development and testing of

computational models of orthographic processing can better delineate the specific mechanisms by which top-down contributions influence orthographic processing.

A vital component of any model of orthographic processing is a description of orthography and orthographic similarity. I argue that orthographic similarity is particularly relevant to descriptions of how top-down modulation influences orthographic processing - whether responses are "explained away" or "sharpened", the degree to which predictions modulate neural activity associated with orthographic processing should correlate with the similarity between the predicted and presented word form. Orthographic Levenshtein distance, the current gold-standard measure of orthographic similarity in alphabetic orthographies, by default overlooks sub-character complexities. In work in this thesis, I develop and validate a sub-character measure of orthographic similarity, showing that its performance in predicting behavioural and neural correlates of visual word recognition, including the N1 component in EEG, can elucidate and better explain sensitivity to sub-character features of orthography.

I additionally describe and validate methodological approaches that can improve experimental design and statistical inference in the research area. Specifically, I describe an R package I developed, LexOPS, that provides a formalisation of an item-matching approach that is flexible and reproducible. Such item-wise matching of factorial conditions is a key component of experimental design in visual word recognition research, as well as in other areas of psychological science. I also describe a formalised distribution-wise approach to matching that can be integrated with the item-wise approach implemented in LexOPS. I apply the item-wise and distribution-wise approaches to matching in stimulus design of the experiments reported in this thesis. Another key component of psychological research that I examine is the norming of items on subjective Likert ratings. Work in this thesis demonstrates, via Monte Carlo simulations, that a statistical approach that appropriately accounts for the hierarchical and ordinal nature of rating norming studies' data, drawing inferences from cumulative-link mixed-effects models, can more accurately and meaningfully summarise rating norms. I demonstrate the improvements conferred by this approach on existing datasets, including normed ratings of perceived orthographic similarity.

This thesis combines multiple complementary approaches to provide insight into the processing of orthography in visual word recognition, and the degree to which such processing may be sensitive to top-down contributions. I provide in-depth experimental evidence and methodological developments that can inform and equip future research and computational models of orthographic processing in the brain.

Acknowledgements

There is no way I could have reached the point of submitting this thesis without the continual support provided by wonderful colleagues, friends, and family. I am deeply thankful to my thesis supervisors, Dr Sara Sereno and Dr Guillaume Rousselet, whose passion for science, and whose humour, in equal measure made this thesis so enjoyable to complete, and who never asked, "exactly how does this fit into your thesis?". Sara, your no-nonsense support and mentorship were as indispensable as your acuminate wit and depth of knowledge. Guillaume, your methodological and statistical expertise are galaxy-brain level and consciousness-expanding. I really cannot imagine a better pair for supervising this thesis - thank you both.

I am also indebted to collaborators on the projects that this thesis includes. Thank you to Alistair Beith for lending indispensable knowledge of linguistics, phonology, and R to LexOPS, and for the incredible music suggestions. Thank you to Dr Christoph Scheepers for the insightful contributions to the CLMM paper, and for helpful discussions on matching approaches. Similarly, thank you to Dr Wilhelmiina Toivo for discussing the distance-based method she has used for matching stimuli in her research.

Thank you to everyone in the Methods & Metascience centre, the Centre for Cognitive Neuroimaging, and the Centre for Social, Cognitive and Affective Neuroscience for fostering such supportive and collaborative research environments. Thank you also to Computing Support, without whom most of this data could never have been collected or analysed.

Thank you to all the friends who have put up with me bundling them into an EEG booth over the past few years: Peter, Matt, Jack, Ryan, Olly, Tommaso, Njenga, Carlos, Alistair, Boss, Caitlin, and Rebecca, I owe each of you a drink or several. Indeed, thanks are due to all the participants whose data form the basis of this thesis, for giving up their precious time to let me stick wires on their heads and/or ask them to read long lists of words, nonwords, or even just letters, to no clear avail. Thank you most of all to Laura for her incredible love, support, and mildly disturbing plots of facial action units: I could not have completed this thesis without you. Thank you also to my parents, who have always supported and encouraged me on this journey.

Finally, my sincerest thanks are reserved for the recent global pandemic, for gently encouraging me to spend more time learning the methodological and statistical techniques that now constitute a substantial part of this thesis. 'Rona - we've never met in person but your influence on this thesis has been far-reaching. You are an inspiration.

Declaration

I declare that, with the exception of section 2.3.3 in chapter 2 which was written in collaboration with a co-author (A. Beith), all work in this thesis is the result of the my own work. This work has not been submitted for any other degree at the University of Glasgow or any other institution.

Abbreviations

AIC	Akaike Information Criterion
ASR	Artefact Subspace Reconstruction
CLMM	Cumulative Link Mixed Effects Model
EEG	Electroencephalography
ELPD	Expected Log Probability Density
EOG	Electro-oculography
ERP	Event Related Potential
fMRI	Functional Magnetic Resonance Imaging
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Effects Model
GUI	Graphical User Interface
HDI	Highest Density Interval
ICA	Independent Components Analysis
IPS	Intraparietal Sulcus
LDT	Lexical Decision Task
LMM	Linear Mixed Effects Model
MEG	Magnetoencephalography
MRI	Magnetic Resonance Imaging
OLD	Orthographic Levenshtein Distance
OLD20	Orthographic Levenshtein Distance 20
PPMI	Positive Pointwise Mutual Information
RT	Response Time
SCOLD	Sub-Character Orthographic Levenshtein Distance
SCOLD20	Sub-Character Orthographic Levenshtein Distance 20
SCT	Semantic Categorisation Task
SD	Standard Deviation
SE	Standard Error
SOA	Stimulus Onset Asynchrony
vOT	Ventral Occipitotemporal (Cortex)
VWFA	Visual Word Form Area

Publications

Journal Articles Containing Work Presented in the Thesis

Parts of chapter 2 of the thesis that relate to the R package, LexOPS, have been published in:

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli, *Behavior Research Methods*. *52*, 2372-2382. https://doi.org/10.3758/s13428-020-01389-1

Parts of chapter 3 of the thesis will be published in:

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2022). Rating Norming Studies should use Cumulative Link Mixed Effects Models. *Behavior Research Methods, in press*. https://doi.org/10.3758/s13428-022-01814-7

Journal Articles Related to the Thesis

Buchanan, E., Cuccolo, K. M., Coles, N., [and 136 others, including **Taylor, J. E.**] (2022). SPAML: Semantic Priming Across Many Languages. *Nature Human Behavior, in principle acceptance as Registered Report*. https://doi.org/10.31219/osf.io/q4fjy

Yao, B., **Taylor, J. E.**, & Sereno, S. C. (2022). What can size tell us about abstract conceptual processing? *Journal of Memory and Language, in press*. https://doi.org/10.31234/osf.io/7dnye

Conference Abstracts and Presentations

Taylor, J. E., Rousselet, G. A., & Sereno, S. C. (2019). Category-Level Semantic Top-Down Modulation of the N170. In *Society for the Neurobiology of Language Conference*. Helsinki: Finland.

Taylor, J. E., Rousselet, G. A., & Sereno, S. C. (2020). Top-Down Modulation of the Word N170. In *Language processing and representation (LangProRep): Scottish-German perspectives*. Dundee: United Kingdom.

Taylor, J. E., Beith, A., & Sereno, S. C. (2021). Workshop: Designing Stimuli Reproducibly. In *The Society for the Improvement of Psychological Science Conference*. Remote Conference.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 Introduction

The written word is a visual symbolic representation of language that supports efficient and precise storage and transmission of information. Cognitively, reading and writing necessarily involve flexible coordination of processes spanning a range of human faculties, including visual, language, motor, and attentional domains. One essential process, if readers are to decode written language, is the accurate perception and recognition of words and their sublexical components in incoming visual information, i.e., *visual word recognition*. For words' meanings to be accessed and processed, readers must decode the *orthography* of written language, that is, the rules of the visual script by which language is transcribed from and to visual representations. The processing of orthography can be viewed as a form of expert perception that enables readers to discriminate, for the case of English orthography, between over 20,000 known unique word forms (Brysbaert et al., 2016) with remarkable speed, with normal reading averaging rates of around 240 ms per word (Brysbaert, 2019).

Despite the clear perceptual expertise that reading demands, written transcription of language is a recent human invention. Rather than a sudden innovation, the invention of writing systems likely emerged through a gradual progression from pictorial, ideographic, mnemonic, and mathematical figures (e.g., X. Li et al., 2003; Locke, 1912; Walker, 1987) before scripts were capable of representing natural language. Nevertheless, archaeological evidence for even the simplest proto-writing is confined to the past 10,000 years of human history, while anatomically modern humans have existed for around 200-300 thousand years. On an evolutionary timescale, then, visual word recognition is a very recent cognitive phenomenon, with limited direct survival or reproductive value. It follows that humans are very unlikely to have evolved dedicated neural circuitry for the specific processes involved in visual word recognition (Dehaene & Dehaene-Lambertz, 2016). Contrast visual word recognition with a cognitive process requiring similarly expert visual perception: face recognition. Evidence shows that the timing and localisation of early visual word and face processing is highly similar (Goodale & Milner, 1992; Rossion et al., 2003), and it has been suggested that they are accomplished and supported by similar neural mechanisms (Kay & Yeatman, 2017; Price & Devlin, 2011). However, the evolutionary value of accurate intra-species face recognition in primates extends

far beyond the homo genus, supporting a range of complex social interactions across primates. Chimpanzees, whose last common ancestor with modern humans lived 4-12 million years ago, show neural mechanisms for face processing in the fusiform gyrus homologous to those of humans (Parr et al., 2009). Evidence for some degree of such perceptual organisation being inherent, rather than learned, is seen in the behavioural preferences of both human and monkey newborns, such as a preference for face-like images or patterns of dots (Goren et al., 1975; Kuwahata et al., 2004; Sugita, 2008; Valenza et al., 1996). An analogous innate preference for alphabet-, cuneiform-, or hanzi-like characters, over non-character control patterns, would be very surprising indeed. Writing systems have existed, and have been perceptually relevant, for a small fraction of the time that faces have. Furthermore, unlike faces, that show consistent geometric regularity as ovoids with predictable locations of and distances between features, orthographic characters can vary considerably in their geometry, within and between writing systems, and across time as writing systems develop. In sum, it would be computationally difficult, and evolutionarily implausible, for humans to have evolved innate neural organisation specifically for visual word recognition.

How, then, do humans achieve this feat so successfully? As this introduction will show, evidence suggests that humans are utilising neural substrates which, although they did not evolve for orthographic processing, are conveniently placed for abstracting orthography from sensory input and bridging it with language and higher-level circuitry. An additional feature that may characterise human orthographic processing, and account for the efficiency with which humans can read, is a sensitivity to *top-down modulation*, permitting flexible and fast processing in a context-dependent manner. Here, it is proposed that higher-level attentional, language, and executive functions guide early orthographic processing via general or targeted predictions of upcoming content. I begin by defining orthography and reviewing evidence for where and when orthography is processed, focusing on early occipitotemporal activity. I then examine the degree to which evidence suggests such occipitotemporal processing of orthography may be sensitive to top-down modulation. Finally, I introduce the methodological approaches I have developed and applied throughout this thesis, and present an outline of the thesis' chapters.

1.2 Defining Orthography

Orthography refers to conventions in the representation of language in a written or printed form. The building blocks of orthography are individual *graphemes* that represent language by each encoding one or many sublexical or lexical features, which can be combined to form progressively larger orthographic units, like morphemes, character N-grams, and word forms. The granularity of information that graphemes represent differs across writing systems. For instance, the characters of alphabetic orthographies typically encode spoken language at the level of phonemes, while writing systems like Chinese characters mostly encode information at the level of syllables. Nevertheless, the features encoded by orthography are almost always phonological in nature, even when the graphemes are logographic (i.e., each character represents an individual word), as Chinese characters are. Indeed, although they vary in the transparency of this orthography-phonology relationship (Katz & Frost, 1992), most

orthographies serve to represent spoken language. Writing systems that represent concepts entirely independently of spoken language are very rare, and very limited in their range of expression (Sampson, 2016). In transcribing spoken language, orthographies therefore reproduce many of its features. As an example, one central feature of spoken language, and therefore of written language, is its ordinal nature. Changing the order of characters, word forms, or phrasal components can dramatically alter, or destroy, the intended meaning of language, just as it would for the spoken language counterparts of such linguistic units. Consequently, writing systems must include rules dictating the order in which orthographic components are to be read and combined, such as left-to-right and top-to-bottom, to convey their intended meaning.

To summarise, orthography refers to rules by which phonetic components of spoken language are transcribed into visual word forms as lexical graphemes, or else via sublexical graphemes that can then be combined to produce individual word forms. These word forms can then be further combined ordinally to represent natural language, reproducing features of the spoken language that orthographies transcribe.

1.3 The "Visual Word Form Area" and the N1

Given the implausibility of orthography-dedicated neural substrates arising from evolutionary pressures, it is at first glance surprising that literate humans show a high degree of homogeneity in the neural processing of visual word forms. In particular, an area in the left ventral occipitotemporal cortex (vOT) of the fusiform gyrus, not anatomically distant from other regions implicated in expert visual object perception and recognition (Goodale & Milner, 1992), has been consistently implicated in orthographic processing (Cohen & Dehaene, 2004; Dehaene & Cohen, 2011; McCandliss et al., 2003; Petersen et al., 1988; Price, 2012; White et al., 2019). Containing a region sometimes referred to as the visual word form area (VWFA), vOT shows robust sensitivity to visually presented words (Cohen & Dehaene, 2004; Price, 2012). Furthermore, this region is known to be *functionally* implicated in reading and visual word recognition, as opposed to showing epiphenomenal activation, as demonstrated in studies of participants with lesions to vOT, who consistently show alexia (Cohen & Dehaene, 2004; Turkeltaub et al., 2014; Wilson et al., 2013).

Readers also show striking similarity in the *timing* of orthographic processing. The first negative-going event-related potential (ERP) component observed in electroencephalography (EEG) signals, following word presentation, is consistently associated with the processing of orthographic features of word forms (Bentin et al., 1999; Ling et al., 2019; Maurer, Brandeis, et al., 2005; Maurer, Zevin, et al., 2008; Pleisch et al., 2019). The magnetoencephalography (MEG) counterpart to the N1, the M170, shows similar timing and topography, and a comparable sensitivity to orthographic and morphological features (Hsu et al., 2011; Lewis et al., 2011; Solomyak & Marantz, 2010; Zweig & Pylkkänen, 2009). In EEG research, this largely left-lateralised ERP component has been referred to by two names: as the *N1*, reflecting its ordinal status as the first negative-going component, and as the *N170* (or *M170* in MEG), reflecting the approximate timing of its peak in milliseconds. In this thesis, I refer to this ERP component,

when it is observed in EEG, as the N1, due to variability in the peak's timing across studies, individuals, groups with different reading experience, scalp locations, and stimulus features (Brem et al., 2009; Fan et al., 2015; Maurer, Rossion, et al., 2008, see also results of chapters 4 and 5).

Although the N1 is mostly observed as a left-lateralised occipitotemporal ERP component, some studies report an N1 component with a similar pre-200 ms peak, but with anterior topography. This component has also been referred to as the early left anterior negativity (ELAN; Friederici, 2002; Lee et al., 2012; Neville et al., 1991). Although there are exceptions (e.g., Lau et al., 2006), whereas the posterior N1 is typically observed in studies using global average, the ELAN is typically observed in studies using a mastoid reference (Nieuwland, 2019). One interpretation of the ELAN, given its similar timing to the posterior N1, is that it is the same component as the posterior N1, with its topography altered by the use of a different reference system. A similar effect is observed for the N170 component elicited by faces, whose topography is highly dependent on the referencing method used, observed as a posterior negativity with an average reference, or an anterior positivity with a mastoid reference (Joyce & Rossion, 2005). However, this explanation of the ELAN fails to explain the purported sensitivity of the component to syntax, rather than the sensitivity to orthography associated with the posterior N1 (Friederici & Weissenborn, 2007; Neville et al., 1991), and why the ELAN remains negative rather than reversing in polarity (Joyce & Rossion, 2005). In addition to topographical disparities with the posterior N1, it is of note that the ELAN is more reliably observed in response to auditory, rather than visual, stimuli (Steinhauer & Drury, 2012). While ELAN components are observed in some reading studies, they are rare, and may reflect effects carried over from preceding words' ERPs such as the N400 or P600 (Steinhauer & Drury, 2012). Given the apparent differences between the N1 and ELAN, and the ELAN's lack of sensitivity to orthography, I focus on the posterior, occipitotemporal N1 in this thesis.

It is quite likely that the occipitotemporal N1's neural generator is the area identified as the VWFA, which shows similar sensitivity to orthographic features, and a comparable developmental trajectory (Brem et al., 2006; Pleisch et al., 2019; J. Zhao et al., 2014, c.f. Brem et al., 2009). Indeed, a range of evidence suggests that the N1 and M170 originate in an area of the left occiptotemporal cortex that largely aligns with the purported location of the VWFA in vOT. Such evidence is observed in source localisation of M/EEG (Brem et al., 2009; Maurer, Brem, et al., 2005; Parviainen et al., 2006; Taha et al., 2013; Xiang et al., 2019; Zweig & Pylkkänen, 2009), as well as in studies that have combined M/EEG with fMRI methodologies (Cohen et al., 2000; Dale et al., 2000; Pleisch et al., 2019) or have measured EEG responses of vOT intracranially (Allison et al., 1994; Nobre et al., 1994; Whaley et al., 2016; Woolnough et al., 2021). In sum, research shows that an early, occipitotemporal response to visual word forms is involved in the processing of orthography.

1.4 Neural Recycling and Visual Word Form Specificity

If humans are unlikely to have evolved orthography-specific neural circuitry, why is it that humans show such consistency in the timing and location of orthographic processing? One plausible

explanation is that humans are re-purposing, or "recycling", neural circuitry that has features convenient for the representation and decoding of visual word forms (Dehaene & Cohen, 2007, 2011). According to such an account, sensitivity to orthography arises through experience, utilising neural substrates that would otherwise be involved in processes that are unrelated, though perhaps computationally analogous. There is broad agreement with some version of the neural recycling hypothesis, though the extent to which this renders the VWFA a misnomer, because it is not necessarily visual, or even not necessarily word-form-related, persists as a subject of debate.

1.4.1 Visual Processing of Orthography

The location of the VWFA in vOT places it in close proximity to higher-level visual areas, especially those on the ventral visual stream putatively associated with object recognition (Goodale & Milner, 1992). Similarly, the timing of the N1 is consistent with rapid processing of visual information. These spatial and temporal features concord with the visual nature of orthographic processing, which can be viewed as a form of expert visual perception akin to other occipitotemporal processes like face or tool recognition (Grill-Spector & Malach, 2004). Indeed, sensitivity to orthographic features of word forms scales with reading experience and ability, for both vOT activity and N1 amplitude (Brem et al., 2006; Brem et al., 2018; Dehaene et al., 2010; Dehaene-Lambertz et al., 2018; Pleisch et al., 2019; Varga et al., 2020; J. Zhao et al., 2014). An additional feature of the VWFA's location is that it shows projections into (Bouhali et al., 2014), and functional connectivity with (Vogel et al., 2012; W. Zhou et al., 2016), frontal and perisylvian areas associated with language and attention - more so than the regions that surround it. Furthermore, rather than developing during or after literacy acquisition, such projections even exist prior to language acquisition, observable from just a week after birth (J. Li et al., 2020), though literacy acquisition appears to strengthen these connections (López-Barroso et al., 2020; Moulton et al., 2019). Localisation of the VWFA in vOT differs somewhat between individuals, and its specific location in literate children can be predicted from the patterns of connectivity of candidate locations in the fusiform gyrus that exist prior to learning how to read (Saygin et al., 2016). As a result, it has been often proposed to be a combination of the visual input to the VWFA, its location in the ventral visual stream utilised in object recognition, and its functional connectivity to left-lateralised language areas that result in the area developing sensitivity to orthography in the literate brain with such consistency (Behrmann & Plaut, 2013; Dehaene & Dehaene-Lambertz, 2016).

Notably, regions in the ventral visual stream that are related to object recognition show a degree of invariance in their response across colour, location, size, and orientation of objects (Grill-Spector & Malach, 2004; Rust & DiCarlo, 2010). Intuitively, this largely concords with what may be expected of orthographic processing, which must be achieved for words of varying viewing conditions, locations, and typographies, and is ostensibly consistent with vOT's location in the ventral visual stream. Whether the orthographic processing that takes place in vOT and during the N1 is invariant across such dimensions has therefore been a key question for investigations into the area (Dehaene et al., 2005). Location invariance, i.e., consistent

representation regardless of retinal position, was identified as a likely feature of the VWFA in early investigations. For instance, it has been shown that the VWFA's response is invariant to manipulations of which visual hemifield words are presented in (Cohen et al., 2000; Cohen et al., 2002), with orthographic information in the left hemifield likely being conveyed from the right visual cortex to the left vOT via the corpus callosum (Bouhali et al., 2014; Cohen et al., 2000; McCandliss et al., 2003; Molko et al., 2002). Such an interhemispheric account is consistent with the reduced sensitivity, and possibly delayed timing, of vOT's response to orthographic information that is presented in the contralateral hemifield (Rauschecker et al., 2012, though EEG has failed to demonstrate timing or sensitivity differences, Takamiya et al., 2020). However, it has also been shown that some positional information must be present in the area identified as the VWFA. Rauschecker et al. showed that a support vector machine classifier trained on multivariate patterns of VWFA activity, as measured by fMRI, was able to decode a presented word form's position in the visual field, both horizontally and vertically, with accuracy above what would be expected by chance. This finding of retinotopy in the VWFA was reconciled with earlier evidence for location-invariant orthographic processing by suggestions that the VWFA's representations include retinotopic information in the first stages of orthographic processing, that take place in more posterior regions, but that this information is discarded as representations become progressively more abstracted from the visual input (Hannagan & Grainger, 2013; J. S. Taylor et al., 2019), supported by evidence that posterior regions of the VWFA are sensitive to positional information that the more anterior regions are not (Dehaene et al., 2004). As a result, evidence supports the *emergence* of location invariance in vOT, as representations are abstracted from vision.

A second type of invariance that could be expected from the purported VWFA, if it is responsible for orthographic processing, is an invariance to typography. While it is well known that the VWFA and N1 are consistently sensitive to orthography across scripts and languages (Bai et al., 2011; Fan et al., 2015; Gagl et al., 2020; Krafnick et al., 2016), this alone does not necessarily mean that orthographic information is represented in the same manner, or with the same sensitivity, across scripts and languages. Varying the typographic appearance of word forms, while keeping the word, context, script, and language constant, can provide insight into whether such linguistic information is abstracted from visual input in orthographic representations. Indeed, some degree of invariance to typography should be expected for orthography, given the variability that exists across a single writing system which readers of that orthography are routinely required to negotiate. In addition to differences between fonts (e.g., a vs. ɑ), in an alphabetic writing system a single word can be written using all lower- or upper-case letters that bear only limited visual or geometric resemblance to one another (e.g., *barge* vs. *BARGE*). Because upper-case graphemes are so typographically distinct from their lower-case counterparts, while preserving phonemic and lexical identity, manipulation of letter case has been common in studies of typographic invariance, with findings generally demonstrating that the VWFA's response is invariant to it. The VWFA shows activation in response to word forms whether they are case-consistent or mixed-case (i.e., both *window* and *WiNdOw*; Polk & Farah, 2002). Moreover, priming studies using the same word as prime and target, but with varied letter case, show priming effects on the VWFA in both within- (i.e.,

cat-*cat* or *CAT*-*CAT*) and cross-case (i.e., *cat*-*CAT* or *CAT*-*cat*) manipulations (Dehaene et al., 2004; Dehaene et al., 2001), suggesting shared representation. More recent work, utilising multivariate pattern analysis of fMRI to decode the information encoded by VWFA representations, has further supported the hypothesis that the VWFA does not represent the case of a given word, but represents information abstracted from the given typography. Specifically, a classifier trained on the differences in activity patterns observed between lower-case words and letter strings can also discriminate between upper-case words and letter strings, whereas a classifier trained on the differences observed between lower-case and upper-case words was unable to discriminate between lower- and upper-case strings of letters (Lu et al., 2021). However, as with location invariance, typographic invariance may emerge only in the more anterior regions of vOT. Z. Zhou, Vilis, et al. (2019) showed via a repetition suppression paradigm that while the VWFA exhibits a response that is mostly font-invariant, it also shows a limited degree of font sensitivity: whereas more anterior occipitotemporal regions showed a font-invariant response specifically in the left hemisphere, more posterior occipitotemporal regions showed font sensitivity bilaterally. Notwithstanding an additional finding of possible font invariance in early occipital regions (see section 7.5), Zhou et al.'s findings are largely consistent with evidence for posterior-to-anterior emergence of font invariance in vOT. Indeed, a picture emerges of the VWFA as a region that processes visual orthographic information with representations that become progressively more abstracted from visual input as responses propagate anteriorly (Vinckier et al., 2007), in a manner comparable to, and reproducible within, the hierarchical organisation of neural networks (Hannagan et al., 2021).

1.4.2 Meta-Modal Linguistic Processing

Thus far, vOT's and the N1's responses to visual linguistic input have been considered, but can these neural phenomena be considered exclusively visual? In addition to visual word forms, the VWFA also shows activation in response to spoken words in literate participants (Muneaux & Ziegler, 2004; Perre & Ziegler, 2008; Planton et al., 2019; Salverda & Tanenhaus, 2010). Non-visual modalities are very unlikely to preserve the timing of vOT responses, such that an auditory N1, at around 170 ms in vOT, would be very unlikely. Nevertheless, EEG evidence suggests that there are influences of orthography on ERPs elicited by spoken language, within the period of the N400 (Pattamadilok et al., 2011; Perre et al., 2011; Zou et al., 2012) or even earlier (Pattamadilok et al., 2014; Pattamadilok et al., 2011). Although it is unclear whether orthography-phonology interactions in auditory word recognition are employed automatically or strategically (Cutler et al., 2010; Pattamadilok et al., 2014; Pattamadilok et al., 2011; Planton et al., 2019), such evidence has generally been interpreted in terms of a functionally relevant recoding of spoken language into an orthographic code (Dehaene et al., 2002; Madec, Le Goff, Anton, et al., 2016), facilitating auditory language comprehension. Indeed, Dehaene et al. (2010) have shown that literate participants exhibit greater VWFA activation during auditory word recognition than do illiterate participants, supporting the hypothesis that it is orthographic information being activated. Such cortical reorganisation of oral language

processing is plausible, consistent with behavioural evidence and findings of comparable neural reorganisation in other areas that occurs after literacy acquisition (Dehaene et al., 2015), at least when the orthography contains relevant phonemic information (Brennan et al., 2013). However, more recent evidence has suggested that rather than spoken language being recoded into orthographic information to be represented in the VWFA, the VWFA of literate participants may encode spoken language in an auditory modality directly. Specifically, representational similarity analysis suggests that both orthographic and phonological information is represented in vOT (Qu et al., 2022; L. Zhao et al., 2017), with phonological similarities best predicting vOT activity in more anterior regions, where representations are more abstracted from visual input (J. S. Taylor et al., 2019). Investigating auditory representations in vOT more directly, Pattamadilok et al. (2019) had participants complete a lexical decision task using both visual and auditory modalities, with an adaptation to either modality preceding each trial, in a 2 (auditory/visual adaptation) x 2 (auditory/visual stimulus) x 2 (word/nonword lexicality) x 2 (left/right vOT stimulation) design. Here, TMS was applied to vOT during stimulus presentation, with the right vOT stimulation employed as a control condition. When adaptation and stimulus modality matched, TMS to the left vOT had a facilitatory effect on lexical decision response times, as would be expected if the adaptation period depressed the neurons responding to that modality, but facilitation *across* modalities was either much smaller or completely absent. This suggests that distinct populations of neurons, within the area affected by TMS to left vOT, encode language in distinct modalities. As a result, while the VWFA's response to spoken language may arise or increase alongside reading acquisition, it may in part represent the auditory features of language directly in an auditory code. VWFA recruitment while decoding language has also been shown to generalise to the decoding of language in an auditory *script* (Striem-Amit et al., 2012), further highlighting the multimodal flexibility of the area. Moreover, there is reason to believe that auditory information is not the only exception to the VWFA's ostensible visual specificity, as the area also shows sensitivity to language-related information from other sensory modalities.

In addition to responding to visual word forms and spoken language, the area identified as the VWFA appears to be sensitive to tactile linguistic information. Congenitally blind readers of Braille who have been blind their whole lives show greater activation in a region of vOT that overlaps strikingly with the VWFA of sighted readers, in response to real Braille words relative to Braille nonwords (Büchel et al., 1998; Reich et al., 2011). As in sighted individuals, the anatomic location of the VWFA is highly consistent across and within blind Braille readers, and is implicated most specifically in reading processes rather than language or sensory processing more generally (Reich et al., 2011). If the VWFA is assumed to be visual, then its recruitment in Braille reading is particularly puzzling. It could be argued that participants who lost their sight after having acquired visual reading are recoding tactile information into a visual orthographic code, similar to the recoding account proposed to explain VWFA activation elicited by spoken language (Dehaene et al., 2002), yet it is difficult to see how congenitally blind participants who have *never* seen could be recoding tactile information into a visual code. It has therefore been argued that the VWFA is not visual at all, but is rather *meta-modal* (Dehaene & Cohen, 2011; Reich et al., 2011), employed in computations that require the decoding of sensory information,

across modalities, pertaining to shapes and patterns that have linguistic relevance. Such an account can be considered an augmentation of the recycling hypothesis, suggesting that vOT's capacity for decoding linguistically relevant sensory input applies not only to the visual domain but extends to multiple, possibly *any*, sensory modalities. The meta-modal account of vOT is bolstered by evidence that blind readers of Braille represent both tactile and auditory linguistic information in vOT, possibly in discrete neuronal populations (Dzigiel-Fivet et al., 2021), as has been argued to be true of visual and auditory language processing in sighted readers (Pattamadilok et al., 2019). To summarise, while the area of vOT identified as the VWFA is mostly implicated in visual processing for sighted individuals, the region is seemingly sensitive to language-relevant information across multiple modalities.

1.4.3 Non-Linguistic Processing

In addition to sensory exclusivity, it should also be examined whether the VWFA, and the related N1, are only sensitive to information that is language-relevant, or whether they are also functionally implicated in representation and processing of non-linguistic information. It has been suggested that, indeed, reading acquisition may cause competition in occipitotemporal regions between the perception of words, and the perception of objects or faces (Dehaene & Cohen, 2007). The representation of objects or faces in the VWFA, in the literate or preliterate brain, is certainly plausible: nearby regions are implicated in the perception of such stimuli (Goodale & Milner, 1992), which require visual expertise that is arguably comparable to that required for word recognition, and expert perception tasks like face perception result in ERPs with similar topography and timing (Rossion et al., 2003). Furthermore, rather than vOT responding preferentially to grapheme-like visual input (i.e., two-dimensional monochromatic geometric patterns), the region is even involved in decoding linguistic information when the linguistic units are images of objects such as houses (Martin et al., 2019).

Commonly cited as evidence for competition between word form processing and that of objects or faces is the finding that literacy acquisition affects the lateralisation of face perception in occipitotemporal regions. In both fMRI (Dehaene et al., 2010) and EEG (Dundas et al., 2014), it has been found that literacy acquisition causes face perception to become more right-lateralised, often interpreted as evidence that the development of left-lateralised sensitivity to visual word forms prevents these neural populations from responding to faces (Dehaene et al., 2015), as it is proposed that they may have done prior to literacy acquisition. Similarly, Centanni et al. (2018) found that, in children, the size of the region of the left fusiform cortex which is sensitive to face stimuli (i.e., the left fusiform face area) is negatively correlated with the size of the fusiform region sensitive to letters, indicative of competition leading to pruning of the left fusiform face area. Feng et al. (2022) did not replicate this finding, though they did observe an increase in right-lateralisation of the fusiform face response that occurred alongside an increase in reading expertise. Direct evidence for *online* competition between face and word processing has also been found. Fan et al. (2015) reported that the amplitude of N1 ERPs observed in response to images of faces is reduced when the face stimuli are presented concurrently alongside Chinese characters, with no such reduction observed when faces are presented next

to unidentifiable Chinese characters. This implies some overlap and competition between the neuronal populations responsible for processing faces and word forms.

Suggestions that literacy acquisition causes word form perception to compete with other forms of perception for neural resources have been challenged more recently, however. For instance, findings of literacy acquisition causing a redistribution of hemispheric organisation in processes like face perception have been questioned, with Rossion and Lochy (2022) pointing out that right-lateralisation of face processing emerges in infancy long before reading acquisition, and arguing that there is only limited evidence for reading acquisition influencing the degree of this lateralisation. Furthermore, if literacy acquisition does precipitate neural competition with face and object processing, it does not seem to impede such processing. In a study tracking the effect of literacy acquisition longitudinally, Dehaene-Lambertz et al. (2018) found that while the preliterate vOT is clearly implicated in the perception of non-linguistic objects, such as tools, houses, and faces, literacy acquisition does not reduce the area's sensitivity to such objects. Rather, literacy acquisition appears to cause vOT to increase sensitivity to visual word forms, while maintaining its preliterate category-specific sensitivity to other objects. This concords with behavioural (van Paridon et al., 2021) and ERP findings (Pegado et al., 2014) showing that literacy acquisition does not impede object recognition performance or sensitivity, but is in fact associated with small *improvements*.

If the VWFA is implicated in the processing of non-linguistic objects, what features does it represent? There is some evidence that the region of vOT that the VWFA emerges in is involved in processing dynamic motion of inanimate objects (Jastorff & Orban, 2009; Whitney et al., 2019). Whitney et al. (2019) argue that the VWFA emerges in a region seemingly sensitive to object motion because visual input in the early stages of reading acquisition, wherein single-letter saccades cause word forms to make progressive movements in the visual field, require processing that is computationally analogous to the perception of objects in dynamic motion. This account is not necessarily inconsistent with suggestions of competition for neural resources between orthographic processing and the processing of other objects, as many forms of object perception could utilise and benefit from dynamic object motion representations as orthographic processing may, consistent with vOT's robust sensitivity to non-linguistic visual categories (Dehaene-Lambertz et al., 2018; Whitney et al., 2019). It has also been suggested that the VWFA region of vOT is functionally implicated in attentional processes. In addition to robust connections with language networks (Bouhali et al., 2014; W. Zhou et al., 2016), there is evidence that the VWFA has strong connections to the fronto-parietal attention network (Bouhali et al., 2014; Vogel et al., 2012), which strengthen with reading acquisition (López-Barroso et al., 2020). Exploiting the high resolution and large sample size of the Human Connectome Project, L. Chen et al. (2019) examined the structural and functional connectivity between the VWFA and these networks in more detail. Notably, this included robust connectivity with fronto-parietal regions, markedly with parts of the intraparietal sulcus, which is implicated in visual attention and the coordination of perception and motor responses like eye movements (Culham et al., 2006; Grefkes & Fink, 2005). Chen et al. also showed robust connectivity between the VWFA and the middle temporal visual area (i.e., V5/MT+) heavily implicated in motion perception, perhaps consistent with Whitney et al.'s account of vOT

as an area that processes inanimate object motion. Moreover, Chen et al. demonstrated that the VWFA's connections to language and attentional networks were distinct and independently functionally relevant, predicting competency in behaviours associated with these networks in a double dissociation. Specifically, connections to the language network predicted word reading abilities and picture naming vocabulary, while connections with the attentional network predicted competency in tasks utilising visuo-spatial attention, and attention connections did not predict language abilities or vice-versa.

Findings suggesting that vOT is implicated in visual attentional processes or dynamic object perception are not necessarily mutually exclusive, and could provide insight into why vOT seems to be functionally relevant in language-irrelevant processing, such as making manual manipulations (e.g., twisting vs. pouring) in response to images of action-relevant objects (Phillips et al., 2002) and dynamic motion perception (Whitney et al., 2019). Such findings may further augment the neural recycling hypothesis, as such attentional and visual processing is likely to be functionally convenient for key aspects of reading behaviour and processing, such as the control of eye movements and representation of orthography as a collection of visual objects that consequently move across the retina (Whitney et al., 2019).

1.4.4 Summary of vOT and its Word Form Specificity

To summarise, vOT shows, in sighted readers, robust sensitivity to the orthographic features of visual word forms (Cohen et al., 2002; McCandliss et al., 2003; Price, 2012), with a corresponding sensitivity in the N1 component (Bentin et al., 1999; Maurer, Brandeis, et al., 2005). Moreover, consistent with the neural recycling hypothesis (Dehaene & Cohen, 2007), there is strong evidence that sensitivity to orthographic features in vOT develops with literacy acquisition, rather than existing as an innate preference for grapheme-like patterns. Nevertheless, it is clear that the region is much more flexible than the VWFA name would suggest, showing sensitivity to information from non-visual modalities, and possibly non-linguistic information. The involvement of vOT in a range of faculties, and across multiple sensory modalities, has long led to calls for the role of the area, and accounts for its involvement in visual word recognition, to be reconsidered (Price & Devlin, 2003). Although the original conceptualisation of the neural recycling hypothesis, focusing on linguistic processing of visual shapes, was somewhat limited, the general principle of cortical reuse appears to be well-supported. Evidence for sensitivity to non-visual and non-linguistic information can be used to inform a more flexible neural recycling hypothesis, in which the area is meta-modal, and in which literacy acquisition does not necessarily lead to impairments in non-orthographic vOT processes (Dehaene-Lambertz et al., 2018). Indeed, there have been recent calls to describe vOT in terms of its general processing mechanisms and the forms of computation that it supports, rather than its specificity for a single task or type of stimulus (Kay & Yeatman, 2017; Vogel et al., 2014), to provide an account consistent with the area's flexibility. An emerging account of vOT that satisfies these criteria considers the area to be a site of interplay between *bottom-up* and *top-down* contributions (Price & Devlin, 2011).

1.5 Top-Down Modulation of Orthographic Processing

It has been suggested that orthographic processing in vOT is underpinned by top-down modulation of bottom-up sensory information processing. Such an account is consistent with evidence that vOT's representations are flexible and meta-modal; it has been suggested that top-down modulation could underlie findings of flexibility and meta-modality in vOT such as its sensitivity to phonological information (Dehaene & Cohen, 2011; Fischer-Baum et al., 2017; Pattamadilok et al., 2011; Planton et al., 2019; S. Wang et al., 2022). In the interactive account of vOT's contributions to reading, Price and Devlin (2011) argued that the area is an interface between bottom-up and top-down information, wherein abstraction of orthographic features from sensory input is guided by top-down predictions. Accounts like the interactive account of vOT (Price & Devlin, 2011), that permit top-down influences within a processing hierarchy (Rauss & Pourtois, 2013), exist within a predictive coding framework. According to such a framework, the brain utilises higher-level information to build, maintain, and continually update a generative model (or hierarchical series of generative models) of sensory information (Friston, 2010; Rao & Ballard, 1999). Such generative models are proposed to propagate their predictions to lower-level areas to compare internally generated predictions to externally generated sensory input. In the case of orthographic processing in vOT, this would comprise predictions of visuo-spatial features in graphemes and word forms (Gagl et al., 2022; Price & Devlin, 2011), with neural activity scaling with the size of the prediction error, defined as the difference between backward-propagated predictions and forward-propagated sensory information (Gagl et al., 2020; J. Zhao et al., 2019). According to the interactive account of vOT, therefore, motivated to minimise prediction error (A. Clark, 2013; Friston, 2010; Walsh et al., 2020), the reading system may learn to use higher-level processes to generate predictions of upcoming content, and provide these to lower-level orthographic processes via top-down contributions (Price & Devlin, 2011).

To investigate effects of top-down modulation, it is important to define it. Top-down modulation is a potentially broad term (Rauss & Pourtois, 2013). For instance, components *within* orthographic processing, progressing from location- and typography-sensitive visual information, to graphemic, and morphological levels of representation, may transfer information locally via feedforward and feedbackward connections, as proposed by connectionist models (e.g., McClelland & Rumelhart, 1981). Findings have supported the notion that such interactive processing hierarchies exist and function as predicted by a predictive coding account. For instance, Gagl et al. (2020), have shown that orthographic prediction error, calculated as the pixel-wise distance between a presented word form and the orthography's *average* word form, predicts N1 amplitude, and fMRI activity in multiple regions including an area close to vOT. Similarly, J. Zhao et al. (2019) showed that in developing readers of Chinese script, bottom-up orthographic regularity interacts with their ability in reading and lexical classification to predict N1 amplitude, with amplitude in the N1, interpreted as indicative of top-down modulation, becoming less extreme as word recognition ability improves. Such findings provide support for a predictive coding account of orthographic processing, and a likely influence of top-down modulation. However, it has been argued that such findings should not be considered evidence

for top-down modulation unless they also demonstrate an influence of processes outside of the sensory domain or brain region in question (Barlow, 1997; Rauss & Pourtois, 2013). The average word forms used by Gagl et al. (2020), and the lexical classification ability of developing Chinese readers tested by J. Zhao et al. (2019), were context-irrelevant, and so the finding of sensitivity to orthographic error does not necessarily support the notion that predictive coding in orthographic processing is sensitive to top-down modulation. Instead, it is possible that predictions are formed and tested entirely within lower-level orthographic processing mechanisms, without higher-level input. As a result, top-down modulation of orthography can be defined as the direct influence of higher-level *non-orthographic* processes on lower-level orthographic processing, bypassing intermediary stages in the processing hierarchy (Barlow, 1997; Rauss & Pourtois, 2013). Examples of such higher-level non-orthographic information include the broader semantic or task context that a word form is presented within.

How plausible is it, then, that orthographic processing is sensitive to top-down modulation? One aspect that should be considered is whether the brain's anatomical connections could support top-down modulation. Notably, vOT shows robust anatomical and functional connectivity with frontoparietal areas (Bouhali et al., 2014; L. Chen et al., 2019; Vogel et al., 2012), including attention networks and prefrontal regions causally implicated in the top-down modulation of sensory processing (Gilbert & Li, 2013). Findings also suggest that these connections between vOT and frontoparietal areas influence the early processing necessary for word recognition: evidence from MEG and intracranial EEG suggests that frontoparietal regions influence vOT activity in some manner earlier than 200 ms after stimulus presentation (Whaley et al., 2016; Woodhead et al., 2014), within the timeframe of the N1. The plausibility of top-down modulation of such early occipitotemporal processing is further supported by evidence for top-down modulation in correlates of visual perceptional processes comparable in latency and localisation. For instance, research has shown sensitivity to top-down modulation in the N1 (N170) associated with face processing (Blau et al., 2007; Dou et al., 2021; Mattavelli et al., 2013; Rousselet et al., 2011; Wieser & Brosch, 2012), vOT activity during face perception (Kay & Yeatman, 2017), and the N1 observed during object recognition (Maier & Abdel Rahman, 2019; Rose et al., 2005). The notion that readers maintain a generative model of upcoming content capable of influencing multiple levels of representation has also been argued to be well supported by research on predictability (Kuperberg & Jaeger, 2016), though the nature and scope of such a predictive system is widely debated (Altmann & Mirković, 2009; Huettig, 2015; Pickering & Gambi, 2018). Arguments against such predictive processes influencing reading have often contended that naturalistic text is not predictable enough to support predictions of it (Huettig & Mani, 2016). However, while average predictability in naturalistic texts tends to be low, highly biasing contexts do occur rarely, but consistently, with robust facilitative effects on reading behaviours like eye movements (Bianchi et al., 2020; Luke & Christianson, 2016, 2018; Rayner, 1998; Staub, 2015). Further, when specific word forms cannot be predicted, morphological features may still be highly predictable, and may also benefit from orthographic predictions (Lopukhina et al., 2021; Luke & Christianson, 2015, 2018), while non-linguistic information such as visual aids, existing knowledge of the text, and task demands other than reading for comprehension (e.g., skimming a text for specific words

or phrases; Rayner et al., 2016) may provide additional context to reading processes beyond the direct predictability of the text.

The extent to which higher-level, non-orthographic processes can modulate early, lower-level orthographic processing is the focus of this thesis. While research suggests that top-down modulation of orthographic processing is certainly plausible, the extent to which early orthographic processing is causally influenced by such contributions is not unequivocal. To demonstrate causality in effects of predictability, researchers must manipulate the degree of top-down modulation that participants employ in reading processes, while controlling for other factors. In some domains of cognitive research, researchers can interfere with processing in regions involved in top-down processes to demonstrate a causal role, such as with transcranial magnetic stimulation (e.g., Feredoes et al., 2011; Mattavelli et al., 2013; Zanto et al., 2011). Such an approach would make demonstration of causality in language processing difficult: the frontoparietal network involved in attentional and executive processes covers a broad network of regions of the brain (Gilbert & Li, 2013), and regions close to or within this network are likely involved in language processes unrelated to top-down modulation of orthographic processing. As a result, researchers instead tend to investigate the causal role of top-down modulation on orthographic processing with methodological paradigms that intend to differentially bias participants' orthographic predictions, while controlling for bottom-up features. Causal interpretation of such evidence additionally requires consideration of the kinds of insights that can be gained from the selected neuroimaging method.

1.5.1 Questions Permitted by Temporal and Spatial Perspectives

Top-down modulation of orthographic processing has long been proposed to be a vital component in cognitive models of reading (Neisser, 1967), allowing acquired knowledge about written language to inform perceptions of it (Rumelhart, 1977). Indeed, the interactive account of vOT shares key features with connectionist models of word recognition like the interactive activation model (McClelland & Rumelhart, 1981), such as a continual interaction between processing levels (or brain regions) to permit feedforward and feedbackward synthesis. Such cognitive models, although often inspired by principles of neural organisation (Rumelhart, 1989), were principally developed to account for behavioural observations like lexical decision response times (Norris, 2013). With neuroimaging techniques like M/EEG and fMRI, however, researchers can delineate the neural dynamics of such top-down modulation, restricting the search to specific time frames or brain regions. This allows the development and testing of temporally and spatially constrained models of specific neural processes involved in visual word recognition (e.g., Gagl et al., 2022; Kay & Yeatman, 2017; Price & Devlin, 2011). Whether a given investigation provides insight into spatial or temporal aspects of orthographic processing depends largely on which neuroimaging method is applied. For instance, the millisecond-level temporal resolution afforded by M/EEG allows researchers to delineate the timing of cognitive processes, whereas the superior spatial resolution of fMRI can provide insight into their location.

A spatial perspective can provide insight into the spatial dynamics of top-down modulation.

For instance, fMRI could demonstrate that a low-level region shows sensitivity to higher-level information that cannot be inferred from a stimulus' low level features. Analyses could also probe the representational nature of this sensitivity (Kriegeskorte et al., 2008), to examine which features of higher-level information are encoded in lower-level regions, and the extent to which lower-level representations encode the same information as higher-level regions. Furthermore, by analysing functional connectivity, a spatial perspective can delineate the dynamics of how information is communicated between regions to permit top-down modulation (e.g., L. Chen et al., 2019), including identifying which regions the higher-level information is projected from. However, the haemodynamic response, underlying the blood-oxygen-level-dependent signal that fMRI records, is so slow (peaking several seconds post-stimulus) as to provide poor insight into the *timing* of such interactions (S. G. Kim et al., 1997). Although sub-second resolution can sometimes be achieved with fMRI through a variety of methodological techniques (e.g., Buckner et al., 1996; Posse et al., 2012), it still provides far poorer resolution than techniques like MEG and EEG, and interpretation is further obfuscated by variability in the timing of the haemodynamic response across brain regions (Pfeuffer et al., 2003; Siero et al., 2011). Functional near-infrared spectroscopy partially ameliorates such problems, but is still temporally imprecise because of its reliance on the haemodynamic response. Activity indexed by the haemodynamic response could potentially amalgamate multiple distinct events, temporally discrete but spatially proximate. With such coarse temporal resolution, it is difficult to ascertain the behavioural relevance of any top-down influence, as any representation of higher-level information could emerge long after bottom-up orthographic processing has occurred and may actually be contingent on word recognition having already been achieved. Considering the speed at which humans can read and identify words (Brysbaert et al., 2019; Hauk et al., 2012; Keuleers et al., 2012; Sereno & Rayner, 2000), the orthographic processing necessary for word recognition must occur very quickly (Sereno et al., 1998), such that late effects of higher-level information on orthographic processing, in regions like vOT, could be irrelevant to initial word recognition (Hauk, 2016). Indeed, intracranial EEG recordings of responses to words do suggest that, in addition to initial, early orthographic processing in vOT, sensitivity to features like lexicality also emerge in the same (and in proximate) regions hundreds of milliseconds later (Woolnough et al., 2021), possibly related to later feedback from higher-level regions (Woodhead et al., 2014). Furthermore, because of the flexible and meta-modal nature of the region, sensitivity to higher-level information in vOT could, in later periods, even reflect the direct processing of higher-level non-orthographic information, like the presented word's phonology (Pattamadilok et al., 2019), rather than a modulation of orthographic processing. Consequently, while a spatial perspective can provide insights into the localisation of, and connective mechanisms underlying, top-down modulation, direct, online evidence for behaviourally relevant top-down modulation of early orthographic processing requires higher temporal resolution than that afforded by techniques like fMRI.

In contrast to the haemodynamic response recorded by fMRI, EEG measures changes in the electrical currents of the brain, while MEG records the magnetic fields those currents produce. Changes in these electrical currents reflect more directly the activity of neuronal populations, affording far superior temporal resolution. With the millisecond-level resolution

afforded by M/EEG, it is possible to discriminate between effects of top-down modulation on early orthographic processing and later patterns of activity that arise in the same area but are unnecessary for word recognition. Indeed, the ability of M/EEG to temporally disentangle distinct neural events, that would be amalgamated by methods like fMRI, may in part explain some findings indicating distinct activity patterns, or developmental trajectories, in the N1 and vOT (Brem et al., 2009), despite consistent localisation of the N1 in vOT. These disparities may reflect fMRI's failure to isolate early transient activity from later activations in the same region. To distinguish top-down modulation of early processes from later, post-lexical processes that utilise the same regions, I focus on M/EEG methods in this thesis. A spatial perspective can provide useful insight into which regions top-down modulation may influence and originate from, but a temporal perspective is vital to demonstrate an impact on the early processes that instigate initial word recognition.

1.5.2 Biasing Predictions to Causally Investigate Top-Down Modulation of Orthographic Processing

Studies investigating the causal nature of top-down modulation of occipitotemporal orthographic processing have principally manipulated readers' expectations of specific visual word forms or orthographic features and have examined the resulting variations in the N1's amplitude and latency, and in vOT activity. Typically, such biasing of expectations is achieved via linguistic contexts, where text preceding the target word form provides a context that can vary in how predictable it makes the target word form. An alternative approach, meanwhile, biases expectations via non-linguistic cues, such as cross-modal contexts and manipulation of task demands.

Biasing Word Form Predictions via Linguistic Cues

Readers' predictions are typically manipulated via linguistic cues. A common approach is to present a sentential context before a target word to bias its semantics. In these studies, a word's predictability is determined in a pre-experiment norming study, operationalised via Cloze probability: participants read a sentential context and explicitly predict what the next word will be. An alternative approach relies on the readers' adherence to grammatical rules, with the assumption that highly constraining grammatical rules induce strong predictions for the identity or part-of-speech category of an upcoming word form. A key feature common to studies that use linguistic cues to bias predictions is that they bias predictions using information more high-level than orthography, rather than orthography itself. As an illustration, an approach to biasing predictions without necessitating higher-level predictions or processing may be to use something like a repetition priming paradigm, where a target word form is preceded by an identical prime. For example, Eisenhauer et al. (2022) showed that identical targets preceded by an identical prime showed facilitated orthographic and visual processing. Such an effect is interesting, highlighting a role of orthographic preactivation in word recognition. However, as with the predictive coding findings reported by Gagl et al. (2020) and J. Zhao et al. (2019), although the effect reported by Eisenhauer et al. (2022) is suggestive of top-down modulation,

its explanation does not necessitate top-down modulation, as orthographic preactivation could occur entirely within an orthographic processing system or module. Rather than identical primes and targets, then, stimuli should be orthographically distinct, primed (or preactivated) by higher-level information. Indeed, studies with orthographically distinct but semantically related prime-target pairs have demonstrated some effect of semantic relations on vOT activity (e.g., Devlin et al., 2006). However, effects of priming between semantically associated words may reflect higher-level knowledge, but may not necessitate online top-down modulation. For instance, vOT may have learned to automatically preactivate word forms that often co-occur with observed word forms without input from higher-level representations. This relates to an historic argument made about the need to avoid direct semantic associativity, between semantic contexts and target word forms, to conclude that effects result from top-down or interactive effects rather than intralexical preactivation (Fodor, 1983; Forster, 1979). Similarly, if linguistic cues are used to provide a biasing context to interrogate top-down modulation of orthographic processing, these linguistic cues must bias predictions using only word forms that are orthographically unrelated, in terms of orthographic features but also word form co-occurrence, to ensure that the effect of predictability is not intra-orthographic. This is why studies have largely used sentential contexts, which can bias predictions using higher-level semantic and discourse processes, avoiding direct orthographic priming or effects of semantic associativity.

ERP investigations that have manipulated sentential context have generally demonstrated effects in the N1, although the pattern of effects across studies is inconsistent (for a review, see Sereno et al., 2019). Such studies have also typically varied word frequency, with the assumption that an interaction of predictability with word frequency would provide evidence for top-down influences on lexical access. While effects often extend to earlier and later components, I focus here on those effects most relevant to this work, involving predictability within the N1 window. Except where noted, sentences were displayed word-by-word, although different word presentation rates or stimulus onset asynchronies (SOAs) were used. Importantly, the studies cited here all broadly examined effects of prediction on the N1, but differ somewhat in the exact windows they analysed: these timing differences are illustrated in Figure 1.1. The review of studies utilising sentential contexts to examine the effect of predictability is presented chronologically. Sereno et al. (2003), using a 450 ms SOA, manipulated predictability (low, high) and word frequency (low, high) and found an interaction of these factors in the N1 (132-192 ms) across posterior and anterior sites (comprising their first factor in a spatial factor analysis). Specifically, they reported an effect of predictability, leading to less negative amplitudes at higher predictability. This effect was observed for low but not high frequency words. Penolazzi et al. (2007), using a longer SOA of 700 ms, manipulated predictability (low, high), word frequency (low, high), and word length (four or six letters). In a 170-190 ms window, they found that high predictability conditions showed a more negative-going amplitude over centro-parietal sites than low predictability conditions, but unlike Sereno et al. (2003), found no interaction with word frequency. In a German study, Dambacher et al. (2012) varied predictability (low, high) and word frequency (low, high) in three experiments using different SOAs, with results possibly suggesting that the disparity in SOA may account for differences in whether a predictability-frequency interaction is observed. At the shortest SOA

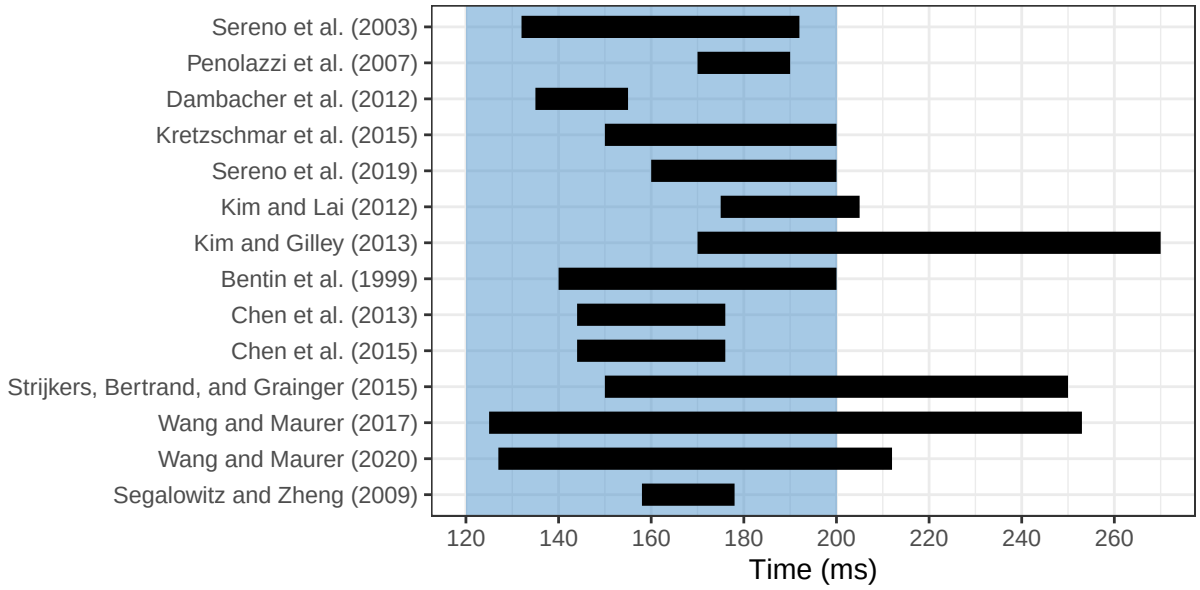


Figure 1.1: The timing of the N1 windows in predictability studies. Studies are listed in order of mention in this section. For reference, the blue region reflects the period defined as the as the N1 window in the experiments in this thesis, as reported in chapter 4 and chapter 5.

of 280 ms, but not at SOAs of 490 or 700 ms, they found an interaction of predictability and frequency in the early portion of the N1 (135-155 ms). For high predictable words only, there was a frequency effect, with low frequency words showing a more negative-going amplitude than high frequency words over posterior sites. In a study measuring eye movements and EEG simultaneously during normal reading, Kretzschmar et al. (2015) also manipulated items' predictability (low, high) and frequency (low, high). Testing only bilateral centroparietal electrodes, their fixation-related potentials (FRPs) demonstrated a main effect of predictability in a window from 150-200 ms, with high predictable words showing a more positive-going amplitude than low predictable words, but did not find an interaction with frequency. Finally, Sereno et al. (2019) manipulated both predictability (low, high) and frequency (low, high). While the first, context sentence was presented in full, the second sentence containing the target word was presented word-by-word, with a short, 300 ms SOA. Sereno et al. (2019) found a predictability-frequency interaction in the N1 (160-200 ms). A predictability effect emerged only for high frequency words. Amplitudes to low predictable words, in comparison to those to high predictable words, were more positive-going over left-hemisphere sites, but more negative-going over right-hemisphere sites. In sum, while these studies using sentential contexts have demonstrated predictability effects in the N1 window, it is clear that the timing and topography of effects, and interactions with frequency are not consistent.

A related study, also using sentential contexts to bias expectations, was conducted by A. Kim and Lai (2012). In contrast to studies cited above that examined the effect of predictability, however, all sentences were designed to be high-cloze. Using a 550 ms SOA, A. Kim and Lai (2012) presented participants with target word forms that were either: the predictable word, a pseudoword orthographically similar to the predictable word, a pseudoword orthographically dissimilar from the predictable word, or a consonant-string

nonword. Consistent with an orthographic explanation for prediction effects in the N1, it was found that, relative to control words, N1 (175-205 ms) amplitude was more negative-going for orthographically dissimilar pseudowords and nonwords. Orthographically similar pseudowords, meanwhile, elicited N1 components more similar in amplitude to control words.

Another method of providing sentential context utilises grammatical manipulations. In such a study, A. E. Kim and Gilley (2013) demonstrated effects of syntactic anomaly on the N1. Sentences leading to a strong prediction for the determiner “the” were presented unchanged or with the determiner replaced with a preposition (e.g., “the thief was caught by *the*/*for* police”). Left-lateralised occipitotemporal N1s (170-270 ms) elicited by the target word were more negative-going when the determiner was replaced with a syntactically anomalous preposition. As the authors point out, the effect of the syntactic anomaly on the N1 is unlikely to be evidence for sensitivity to syntax per se. Rather, in light of evidence suggesting the N1 is sensitive to orthographic features, it is probably more accurate to interpret this early sensitivity to syntactic differences as support for prediction of orthographic features, eliciting less negative-going N1 components when these predictions are confirmed.

That Kim and Gilley’s manipulation simultaneously altered both orthography and syntax reflects an inherent issue with the use of sentential contexts to investigate early processing of visual words. Namely, one cannot alter the visual word form without also altering the semantics, syntax, or plausibility of the sentence or wider discourse the word appears in. Methodological issues also arise, such as that ERPs elicited by the target word can become difficult to disentangle from ERPs elicited by preceding words, especially if the delay is short or unjittered. I argue that while sentential contexts, optionally with realistic presentation times, are useful for demonstrating that early modulation by predictive processes extends to naturalistic reading, their application is not necessary to demonstrate such modulation can occur. It is also of note that in a recent review of ERP studies using sentence- and discourse-level contexts to examine top-down contributions to visual word form processing, Nieuwland (2019) concluded that findings thus far have been weak, inconsistent, and in need of more replication attempts. Most studies so far were additionally not preregistered and used inappropriate models that did not account for measurement variability, raising questions about false positives in that literature.

Biasing Word Form Predictions via Non-Linguistic Cues

As an alternative to relying on linguistic contexts to bias participants’ predictions, effects of top-down modulation may be investigated using paradigms that modulate non-linguistic features of tasks and stimuli. In one approach, task instructions are altered while stimuli are unchanged or designed to be equivalent, exerting an influence on task demands and on participants’ task sets. For instance, a task which explicitly requires participants to judge the lexicality of words may be more likely to lead participants to predict and show sensitivity to lexical variables than a task which requires judgements on a non-lexical dimension, such as the word’s colour. In one such study, Bentin et al. (1999) presented words in a lexical decision task (word vs. nonword), semantic categorisation task (concrete vs. abstract), and rhyme task (rhymes vs. does not rhyme with "-ail"). Results revealed a task-stimulus interaction on the N1 (140-200 ms), with

the difference between orthographically plausible and implausible stimuli (nonwords eliciting N1 components with more negative-going amplitudes) being larger in tasks requiring lexical or semantic processing than in the rhyme task. Y. Chen et al. (2013) also presented target words compared lexical and semantic decision, but included a condition with minimal task demands requiring only silent reading of the words. They identified an effect of task on the N1 (144-176 ms), with a more negative-going N1 for words observed for lexical decision and semantic decision than for silent word reading. In a similar study, Y. Chen et al. (2015) further suggested that the degree to which variables like frequency and imageability affect activity in the N1 (144-176 ms) is task-dependent. For instance, the effect of word frequency on source-space activity in the N1, where less activity is observed as frequency increases, was larger in lexical decision than in semantic decision or silent word reading. Using a related paradigm, Strijkers et al. (2015) similarly reported that ERP amplitude in a period including the N1 (150-250 ms) is more sensitive to word frequency (with more negative amplitudes for higher frequency words) during a semantic categorisation than a colour categorisation task. F. Wang and Maurer (2017) applied a similar paradigm to Chinese symbols, finding that the sensitivity of Chinese-reading participants' N1 (125-253 ms) components to the difference between familiar Chinese characters and stroke-matched, unfamiliar Korean symbols (with more negative amplitudes for Korean symbols than Chinese characters) was greater in delayed naming and colour categorisation tasks than in a repetition detection task. This effect was specifically observed in the N1's offset period of 172-253 ms, where onsets and offsets are defined respectively as the periods in the component's time window which precede or succeed its peak. Other non-sentential approaches to biasing participants' word form predictions include a related attempt to alter expectations for different types of script, with the finding that native Mandarin speakers' sensitivity in the N1 (onset 127-162 ms; offset 162-212 ms) to differences between unfamiliar Korean (more negative offset) and familiar Chinese characters (less negative offset) was greater when participants were led to expect Chinese characters (F. Wang & Maurer, 2020). As task modulation only requires single word presentation, it avoids the need to disentangle neural responses to successive words that often exists with approaches that use linguistic contexts. Correspondingly, the paradigm has been applied in fMRI studies. Comparing neural activity between implicit (silent reading) and explicit (word naming) reading tasks, Qu et al. (2022) showed via representational similarity analysis that representations of both orthographic and phonological information in vOT were sharpened during explicit word naming.

In addition to general effects of task, it is also possible that top-down modulation could affect orthographic processing of *targeted* representations - readers may construct predictions of specific words or categories of words from semantic contexts. Indeed, it has been shown that semantic information can be decoded from vOT activity in a manner that is modulated by task. Examining representational similarity between semantic dimensions (taxonomic and thematic) and VWFA activity during tasks when participants were required to categorise Chinese characters on taxonomic or thematic dimensions, X. Wang et al. (2018) showed that relevant semantic information could be decoded from fMRI patterns in a task-dependent manner. More taxonomic semantic information could be decoded during a taxonomic categorisation task, and more thematic information during a thematic categorisation task. Nevertheless, as

has been argued, such fMRI findings could reflect both (or either) early and late interactions between semantics and orthography, while less temporally coarse approaches like EEG and MEG may be more appropriate for demonstrating an influence of predictions on initial word recognition. In an ERP study that used a task manipulation paradigm, Segalowitz and Zheng (2009) presented words and pseudowords in either a lexical decision task or lexical semantic task. The latter task was here identical to the lexical decision task, except that words were drawn from a single category (e.g., *animals*). Segalowitz and Zheng reported an interaction between stimulus type and task in the N1 (158 - 178 ms), wherein, after the main effect of task had been accounted for, N1 amplitudes for words differed between the tasks, but N1 amplitudes for pseudowords did not. This finding may suggest that knowledge of semantic category membership affected processing during the N1 component, indicating a sensitivity to lexical or post-lexical information in the N1, possibly resulting from top-down modulation. However, it is unclear from this study whether the effect indeed reflects sensitivity to category relevance, or a general effect of task demands on early sensitivity to lexicality or lexical features like frequency, consistent with the research cited above. While the finding that Segalowitz and Zheng report is interesting, further research could improve on the implementation of this paradigm, disentangling interactions between task and lexicality, from interactions between task and category relevance. Although not fully describing the interactions between both task and lexicality, and task and category relevance, evidence has supported the notion that both effects may exist. Using a similar paradigm to that used by Segalowitz and Zheng, Hauk et al. (2012) compared ERPs in lexical (word vs. pseudoword) and semantic (living vs. non-living) decision tasks, showing that in both tasks, stimulus differences were observed as early as 166 ms (data were analysed continuously, with no N1 window definition). This finding suggests, consistent with the findings of Segalowitz and Zheng (2009), an early sensitivity to category relevance during the N1 which, given the N1's robust sensitivity to orthography, is likely to reflect a top-down influence of higher-level predictions on orthographic processing.

To summarise the results of published studies that have used task manipulations to investigate top-down modulation of the N1, task manipulations do seem to alter sensitivity to orthographic features of words, suggesting a contribution of higher-level strategic processes. Less research has examined whether task demands can influence sensitivity to word forms that belong to expected semantic categories, and a key question remains whether such sensitivity reflects a task interaction with category relevance or of lexicality. Paradigms that manipulate task demands in this way have several advantages. Firstly, they remove the need for word forms to be presented within sentences or wider discourse. This makes it much easier to avoid issues like intralexical or intra-orthographic priming which can otherwise obfuscate the interpretation of effects as top-down. Furthermore, manipulating task while presenting identical stimuli across tasks mitigates effects of the target word form that may inadvertently alter bottom-up processing. Indeed, as previously mentioned, no two words differ only in semantics, but will necessarily vary in orthography, as well as in a range of dimensions such as frequency, age of acquisition, familiarity, and phonology. Such features are likely to impact the bottom-up processing of words, requiring careful and precise matching of items. By altering task demands, while presenting identical or tightly controlled stimuli, researchers can alter

top-down contributions but ensure that bottom-up information is comparable or equivalent.

In a related attempt to modulate top-down expectancy without linguistic context, Dikker and Pylkkanen (2011) implemented a picture-word verification task, in which images containing a target object alone, or a target object among other related objects, were followed by written noun phrases (article + noun) denoting the target object. They manipulated congruency (match or mismatch of the noun phrase to an object in the image) and predictability (low, with the target object as one of several objects, or high, with the target object presented alone). Although they did not examine effects in the MEG equivalent of an N1 window, they did find effects of congruency only in the high predictive condition in preceding and succeeding temporal windows. Dikker and Pylkkanen's stimuli were designed such that orthographic similarity between congruent and incongruent pairs of stimuli was minimised, suggesting that the authors anticipated that any early sensory effect of predictability may be related to orthographic processing. It is possible that the study (N=7) lacked the sample size necessary to identify such an effect in an N1-like window. Indeed, in a related paradigm, Kherif et al. (2011) recorded fMRI while picture and word prime-target pairs were presented in a repetition priming design. The stimulus types of prime-target pairs were either matching (*word-word*, varying typography, and *picture-picture*) or non-matching (*word-picture* and *picture-word*). Kherif et al. (2011) showed cross-stimulus priming effects in vOT for the *picture-word* condition. Assuming that picture identity is not directly processed in the orthographic processing system, these findings suggest that higher-level processes link the identity and content of pictures to orthographic representations of word forms. However, Kherif et al.'s use of fMRI obscures the interpretation of the timing of such effects - mapping of picture content to orthographic representations could occur so late as to be irrelevant to initial word recognition processes.

As with the task manipulation approach, picture-word tasks like those used by Dikker and Pylkkanen (2011) and Kherif et al. (2011) have several advantages over paradigms that rely on linguistic cues. Like task manipulations, the picture-word paradigms can avoid intralexical or intra-orthographic priming effects that might sometimes confound approaches like sentential contexts or word-word priming to bias predictions. An additional advantage of picture-word paradigms is that the researcher can control and manipulate variables like predictability and specificity of the picture-word relation. This was demonstrated in the design of the task used by Dikker and Pylkkanen (2011), where the picture preceding the target word could clearly bias participants towards expecting one word form, with an image of one clearly identifiable object, or towards expecting multiple possible word forms, with an image of multiple candidate items. Such a manipulation is comparable to the use of cloze probability in sentential contexts. Nevertheless, unlike task manipulations, picture-word paradigms must use stimuli that are carefully matched between picture-congruent and picture-incongruent word forms to ensure that effects do not arise from bottom-up differences.

1.5.3 Summary of Top-Down Modulation of Occipitotemporal Orthographic Processing

In sum, findings suggest there may be effects of predictability on early occipitotemporal responses to word forms, indexed by the N1, though studies thus far have largely relied on sentential contexts, and have been inconsistent in effects, in need of further research and replication (Nieuwland, 2019). Two promising, non-linguistic approaches to biasing participants' expectations of upcoming word forms are paradigms that manipulate task demands, and paradigms that use cross-stimulus semantic relationships such as picture-word verification tasks. One key question concerns the specificity of prediction effects, as most research has only presented high and low predictability items, rather than varying predictability continuously to delineate more precisely the point at which higher-level information impacts early orthographic processing. This review also highlights the importance of methodological considerations, such as precise close matching of items.

1.6 Methodological Considerations

Throughout this introduction, several methodological issues have been highlighted, such as the need for precise stimulus matching and appropriate modelling approaches. These issues are considered in detail in this thesis, and potential solutions are proposed.

1.6.1 Controlling for Confounding Variables

As highlighted in this introduction, no two words differ on only one dimension; words are multi-faceted, varying continuously on a range of dimensions from orthography and phonology to frequency and semantic associations. With regards to research on top-down modulation of word recognition processes, this means that words differing in their top-down relevance, either congruous or incongruous with the reader's expectations, also differ in terms of their bottom-up features. The problem is often solved in psycholinguistics research by selecting items that differ in the relevant dimension, but are precisely matched on dimensions that may confound the effect of interest. For instance, in research on the top-down modulation of orthographic processing, features like frequency, length, and orthographic neighbourhood should be matched between prediction-congruent and -incongruent conditions. Such matching techniques are are often applied manually, and are difficult to reproduce. However, with the growing importance of replication, reproducibility, and preregistration in Psychological science (Munafò et al., 2017; Nosek et al., 2018; Nosek et al., 2022), there is an increasing need to formalise matching methodologies computationally. In this thesis, I explicitly formalise two approaches for stimulus matching, respectively presenting a new R (R Core Team, 2021) package I developed, LexOPS, and suggesting a novel application of existing statistical tools.

1.6.2 Consideration of Statistical Approaches

Variables used to match items in stimulus design are often derived from norming studies, where participants rate items on specific dimensions. Such variables are also widely used as predictors in statistical models, either as independent variables or as covariates to mitigate the effects of confounds. The norming approach to calculating variable features is most common when variables are difficult to formalise computationally, or cannot be derived from corpora of texts (e.g., the emotional valence of a word). However, such norms are usually calculated as simply the average rating on the Likert scale. Such an approach, when the norms are treated as continuous variables, is inappropriate for a variable that is fundamentally ordinal: there is no reason to assume that the step between *1* and *2* on a Likert scale is equal in magnitude to the step between *2* and *3*. Inappropriate treatment of ordinal ratings as continuous variables limits the quality of inferences that can be made about them, and is the cause of misleading artefacts in norming studies (e.g., Pollock, 2018). Further, because of the size of norming studies, especially for large-scale norming studies of words (e.g., Brysbaert et al., 2014), ratings are calculated from subsets of participants rating subsets of items, which introduces hierarchical complexities into norming data. I argue that with more appropriate statistical approaches, like explicitly ordinal, hierarchical models with random (varying) effects, researchers will be able to more accurately and meaningfully norm items.

One case where such an approach may not improve inferences is when the normed variable is binned into ordinal categories. For example, researchers may compare the 100 words that are rated highest on the normed dimension to the 100 lowest words. However, such binning into ordinal categories considerably reduces the quality of inferences that can be gained from data (MacCallum et al., 2002; Royston et al., 2006) discarding meaningful differences between observations. One possible reason for ordinal binning of continuous variables being so widespread is that it has often been necessary to satisfy the requirements of statistical methods. For instance, *t*-tests and Analyses of Variance (ANOVAs) compare means of groups, while accounting for variability within groups. Traditionally, including continuous variables in analyses would have required the use of regression approaches that fail to account for repeated measurements within groups. However, with the rise of mixed effects (hierarchical/multilevel) approaches to modelling it is possible to include multiple random effects in a model with continuously varying fixed effect predictors, more accurately describing how the data were generated (Baayen et al., 2008; Barr et al., 2013; Pinheiro & Bates, 2000; Yarkoni, 2022). In addition to using random effects to calculate norms, in this thesis I apply mixed effects models to more accurately describe behavioural and EEG data. Such approaches can additionally benefit from the use of link functions more appropriate to the probability distribution of the dependent variable, in *generalised* linear mixed effects models (Bürkner & Vuorre, 2019; Lo & Andrews, 2015), and from descriptions of changes in the full distribution rather than simply its central tendency, in distributional models (e.g., Heathcote et al., 1991; Staub et al., 2010).

1.6.3 (Re-)defining Orthographic Similarity

Just as statistical models can be selected that more fully describe data, I argue that fuller computational descriptions of orthography can improve the quality of insight that can be gained from cognitive models of visual word recognition. The most widely applicable measure that can be derived from an orthographic description is *orthographic similarity*, quantifying the distance between word forms. Orthographic similarity has played a significant or central role in models of visual word recognition (e.g., Adelman, 2011; Davis, 2010; Gagl et al., 2022; Gomez et al., 2008; Norris, 2013; Norris & Kinoshita, 2012). However, measures of orthographic similarity have mostly limited their scope to the level of graphemes, implicitly ignoring the more fine-grained elements from which characters are composed. Indeed, the mostly widely used measure of orthographic similarity, orthographic Levenshtein distance, and its neighbourhood complement, OLD20 (Yarkoni et al., 2008), are calculated under the implicit assumption that characters are functionally equivalent and interchangeable in orthography. Standard Levenshtein distance neglects the complexities of individual graphemes, and similarities between graphemes. I argue that calculating orthographic similarity from a pixel-based implementation of orthography can yield more powerful and sensitive metrics of orthographic similarity. Such measures can be used to further research on areas like typographic sensitivity, since it permits the calculation of font-specific similarity estimates. I further argue that such a measure can contribute meaningfully to the development of more specific hypotheses regarding orthographic processing, particularly its sensitivity to top-down modulation, as models of top-down modulation make specific predictions about the relationship between top-down modulation of orthography and orthographic similarity between observed and predicted word forms.

1.7 Thesis Layout

In chapter 2, I present an R package I developed, LexOPS, to enable reproducible and precise item-wise stimulus design. I also describe and evaluate an alternative approach using existing tools to match stimuli in a distribution-wise manner, and compare this to, and integrate it with, the item-wise approach implemented in LexOPS. In chapter 3, I argue that norming studies that use Likert rating paradigms should calculate their norms from hierarchical ordinal models. I specifically focus on Cumulative Link Mixed Effects Models (CLMMs), and show via a series of Monte Carlo simulations and re-analyses of existing datasets (including of subjective orthographic similarity ratings), that CLMMs solve many statistical issues that exist for traditional norming methods. Approaches outlined in chapters 2 and 3 are applied in research presented in later chapters, including stimulus design and hierarchical analyses of rating data. In chapter 4, I expand on the paradigm introduced by Segalowitz and Zheng (2009) and examine whether the occipitotemporal N1 shows sensitivity to category-level expectations of word forms, manipulating task by presenting the same items in a lexical decision task and a semantic categorisation task. I present evidence that stimulus-task interactions emerge later than the N1. In chapter 5 I hypothesise that the lack of effect in chapter 4 was due to the lack

of word form specificity permitted by category-level predictions. In a picture-word verification task, I vary predictability continuously to identify the relationship between predictability and top-down modulation. I find an interaction between picture-word congruency and predictability that is consistent with top-down modulation of activity during the N1. However, I argue that the pattern of results that I find is inconsistent with a simplistic implementation of predictive coding, instead supporting a "sharpening" of neural responses to predicted orthographic information. In light of these results, I consider that a full description of how semantic information interacts with orthographic processing requires a computational description of orthography, from which orthographic similarities and neighbourhoods can be inferred. In chapter 6 I implement a pixel-based measure of orthographic similarity (SCOLD) that could be integrated into models of orthographic processing to provide a more powerful framework for describing hypotheses of orthographic processing and its sensitivity to top-down modulation. I compare the predictive power of SCOLD to that of standard Levenshtein distance in models of behavioural and neural correlates of orthographic similarities and neighbourhood densities. Finally, in chapter 7, I consider and evaluate the conclusions that can be drawn from this thesis, relating the work to existing literature and future work.

Chapter 2

LexOPS: An R Package and User Interface for Stimulus Selection

2.1 Introduction

A fundamental dilemma inherent to psycholinguistics research is that there exists a finite set of real words. By necessarily sampling from a finite population of stimuli, psycholinguists are limited to only drawing inferences about the effects of variables from locations within these variables at which words exist. Were a feature space created from all visual, auditory, and semantic variables which might impact humans' perceptions of words, words would populate this space rather sparsely, and there would be few cases of any two words differing in terms of only one variable. This stands in contrast to forms of stimuli for which unique items can be generated by sampling values from continuous parameters which define them. At the extreme, stimuli randomly generated from these parameters can be applied in approaches like reverse correlation to infer isolated effects of, and interactions between, relevant variables. Examples include attempts to delineate the perception of features in orthographic characters (Gosselin & Schyns, 2003; Ling et al., 2019), faces (Jack & Schyns, 2017; Mangini & Biederman, 2004) or voices (Ponsot et al., 2018). Similarly, researchers wishing to modulate an independent variable while holding confounding variables constant can typically generate, from such a system, items which differ in one variable but are matched exactly on other variables, even if no such stimulus exists in the population of real stimuli that the system emulates. While such an approach may be applied in psycholinguistics to generate novel plausible pseudowords which are perceptually similar (orthographically or phonologically plausible) to real words (such as with an N-gram chain method of generating pseudowords; Keuleers & Brysbaert, 2010), its application restricts researchers to *only* examining and considering effects of stimuli's sublexical features. This reduction would neglect the effects of higher-level variables like semantics and frequency of exposure which are present in real words, and which the researcher may even be specifically interested in, as a novel pseudoword has no meaning and a frequency of zero. Instead, researchers in psycholinguistics largely rely on presenting real words varied and matched on respective variables within some reasonable tolerance, or else using data-driven approaches attempting to implement controls statistically on the limited number of observations available.

Whether designing controlled experiments or estimating effects of variables in data-driven

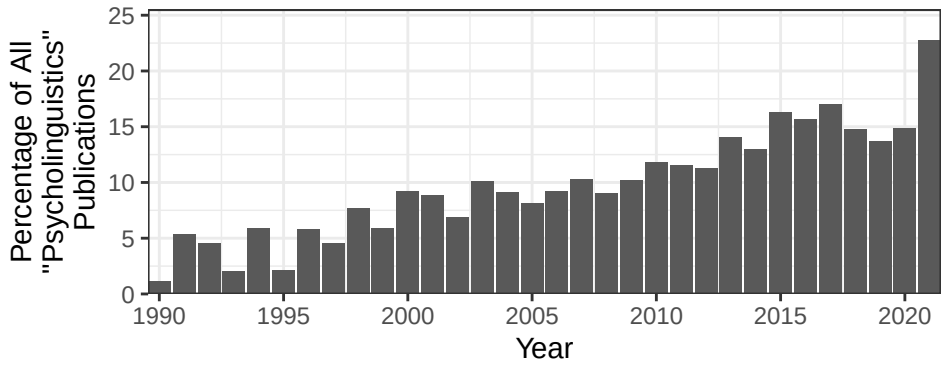


Figure 2.1: The percentage of documents on Scopus published each year in the period 1990-2021 containing the term "psycholinguistics" in the title, abstract, or keywords, which also contains the term "corpus", "database", or "norms".

approaches, psycholinguists require a large amount of data on variables of interest to identify suitable candidate words and to implement controls. Reflecting this demand, the number and size of psycholinguistic corpora that have been created and employed in research have greatly increased in recent years. Figure 2.1 shows the growing proportion of psycholinguistic research over the past three decades that provides or cites databases related to various properties of words. Indeed, the use of large datasets has been made considerably more feasible as a result of the internet and an increase in computing power. Although such large-scale databases of psycholinguistic features, with interfaces for querying and downloading contents, have existed for many years (e.g., Balota et al., 2007; Coltheart, 1981), few tools currently exist to aid in adapting these datasets to generate suitably controlled stimuli, and these are often greatly limited in their capabilities, requiring considerable supervision and manual inputs from researchers. This makes the generation of controlled word stimuli currently time-consuming, labour-intensive, and difficult to reproduce.

One tool that has been provided and is widely used to control for factors in stimulus design is *Match*, a command line tool written in C++ (van Casteren & Davis, 2007). *Match* supports item-wise matching of factorial designs on numeric variables. Here, it uses Euclildean distance, with equal weighting given to all variables, to identify optimal matches across conditions. For instance, if matching two conditions of word stimuli by word frequency and length, it will iteratively try combinations of word pairs to minimise the overall sum of Euclidean distances between all pairs of words. The final stimuli provided will comprise the list which, of those tried, has the smallest overall sum of distances between pairs. *Match* has been widely used to generate stimuli, especially in psycholinguistic research. However, the program has key limitations when compared to how researchers often match words. For instance, matching by unweighted Euclidean distance assumes that researchers consider all variables as equally important in matching. In fact, researchers may be more concerned with matching variables like word frequency, that account for large proportions of variance in word recognition behaviour (Brysbaert et al., 2011; van Heuven et al., 2014), than they would be for matching variables likely to be less impactful, such as a word's number of synonyms. Furthermore, the researcher may wish to match items using variable-specific tolerances (e.g., match word length exactly,

but match frequency within a certain tolerance), combine such variable-specific approaches with distance matching, or match by multidimensional item-wise variables like orthographic similarity. In this chapter I present a tool I have developed to address such limitations of existing software, providing a much more flexible solution to the problem of matched stimulus design.

This chapter presents LexOPS, a flexible R (R Core Team, 2021) package I developed to offer a comprehensive range of capabilities relevant to the selection of psycholinguistically controlled word stimuli. The appellation, 'LexOPS', is derived from four types of word properties commonly recognised in psycholinguistics: Lexical, Orthographic, Phonological, and Semantic. The most noteworthy feature of LexOPS is that it can produce suitably controlled word stimuli for any possible user-specified factorial design, specified in a readable and fully reproducible pipeline of code. To further support readability and interpretability, the package features an easy-to-use graphical user interface (GUI) in the form of a Shiny app (W. Chang et al., 2018), which provides multiple interactive visualisations and summaries of available word properties, as well as how stimuli LexOPS has generated relate to these properties, and can translate selected GUI options into reproducible code. Another novel feature of the package is that it can work with any database of variables for a finite set. This means that the user is not limited to built-in variables or words, but can design stimuli according to any numerically or categorically defined properties, for words from any language. Nevertheless, several useful psycholinguistic variables are included from a range of datasets to illustrate the capabilities of LexOPS. These also serve as a template demonstrating the expected format of the data if users wish to run LexOPS on their own databases. Given that LexOPS can work with any suitably formatted data, and the ease with which new datasets can be downloaded and combined, the built-in dataset included with LexOPS is explicitly not exhaustive in its coverage.

This chapter first provides an overview of the package's functionality in generating well-controlled stimuli. I then describe the variables native to LexOPS, citing sources for the data and explaining the processes by which original variables were calculated. Using variables drawn from the built-in dataset, I then provide illustrative examples of possible applications for LexOPS. Following this, an introduction to the package's accompanying Shiny app is presented. I also report the results of a validation analysis, comparing the stimuli used in several well-controlled experiments to examples generated with the package. Implications for reproducibility and replicability are discussed. Finally, a distribution-wise approach to matching stimuli is introduced and briefly contrasted to the explicitly item-wise appraoch of LexOPS. I demonstrate that the item-wise and distribution-wise approaches are not mutually exclusive, but can be combined to flexibly generate stimuli matched using both approaches. Although I provide an overview of the LexOPS package's functionality, and alternative computational approaches to matching, this chapter is not intended to be read as a tutorial. Detailed instructions on how to install and use the package are available in the LexOPS walkthrough: https://JackEdTaylor.github.io/LexOPSdocs/.

2.2 Functionality Overview

LexOPS is designed to support two main methods of stimulus generation: a fully automated grouping of items into factorial cells according to specific constraints (with the "generate

pipeline"), and a more bespoke matching of stimuli from several candidates (with the *match_word()* function). Example practical applications are provided with code later in this chapter.

2.2.1 The Generate Pipeline

The "generate pipeline" consists of three main functions: (1) *split_by()*, for specifying independent variables; (2) *control_for()*, for specifying variables that should not differ between conditions; and (3) *generate()*, for running the algorithm that generates lists of stimuli. Factorial designs with any number of main or interactive effects can be generated by calling the *split_by()* function once for each variable which consists of a main effect. The factorial designs specified by these functions can adopt any number of word properties, expressed either numerically (e.g., concreteness) or categorically (e.g., part of speech), as independent variables with user-defined levels. Similarly, the user can define any number of control variables with multiple calls to *control_for()*, with tolerances of any size. The *generate()* function employs options defined in calls to *split_by()* and *control_for()* to create a stimulus list, with the requested number of items, that fit the specified options.

The *generate()* function creates lists of stimuli in the following way. Firstly, the boundaries of each factorial cell are identified, and the factorial cell to which items should be matched (i.e., the "match-null" condition), is assigned. By default, this is done pseudo-randomly, such that each factorial cell is used as a match-null with equal frequency, and in a random order. If the number of stimuli requested is not divisible by the number of factorial cells, the match-nulls will, by default, be allocated as equally as possible across conditions, with over-represented conditions selected randomly. Other (non-default) options for assigning conditions as match-nulls to each item include assigning a single factorial cell as the match-null for all items, assigning match-nulls completely randomly (i.e. no attempt is made to balance assignment across conditions), and designating that the tolerances should be treated "inclusively" such that every factorial cell is within each tolerance of every other factorial cell. The *generate()* function then iteratively identifies suitable combinations of stimuli. On each iteration, a word is randomly selected that fits the current match-null condition's specifications (e.g., a word with a low valence rating that is a noun). Possible matches that fit the other conditions' specifications (e.g., high valence nouns, low and high valence verbs), but that are matched to the word selected from the match-null condition on control variables (e.g., within ±.2 Zipf frequency and of equal length), are then identified for each condition, from a pool of unused words. One word is randomly selected from this pool for each condition. If it is not possible to generate a match from each condition for the word from the match-null condition, the function will discard the result of this iteration, and randomly select another word that has not yet been tried, from the same match-null condition. Words that are successfully generated for each condition are stored, and the function will attempt to generate another matched set for the next match-null condition. This will be repeated until as many stimuli are generated as was requested, or until the function fails to generate new stimuli. In addition, the user can elect to generate as many stimuli as possible. If this is specified, the function will generate items until it can no longer generate a matched set across

all conditions.

2.2.2 Generating more Complex Experimental Designs

The two key functions of the generate pipeline detailed in the previous section, *split_by()* and *control_for()*, are suitable for most applications, but are insufficient for more complex experimental designs. Additional functions have been created to allow the *generate()* function to accommodate these commonly used methods of matching stimuli. These functions can be combined with *split_by()* and *control_for()* and can be similarly chained multiple times within a single pipeline to allow the user to define any number of splits or controls.

Creating Conditions Unrelated to the Stimuli

A central assumption of the *split_by()* and *control_for()* functions is that the experimenter wishes to control for specific variables across conditions which differ according to features of the stimuli. In many cases, however, this is not the case. The experimenter's manipulation may instead be unrelated to the stimuli. For instance, the experimenter may wish to present tightly controlled stimuli in two distinct tasks, or varying whether the participant receives genuine or sham brain stimulation such as Transcranial Magnetic Stimulation (TMS), but avoiding presenting the same words to each participant twice. The experimental design may additionally examine whether the effect of this variable interacts with one or many variables which could be defined more straightforwardly via *split_by()*.

The *split_random()* function was written to allow for stimuli for such designs to be created with LexOPS. The function allows the user to specify a split in the data which is random (though reproducible with a seed), with any number of levels, matched item-wise using the specified controls. If a pipeline is run which uses both *split_by()* and *split_random()*, this will result in a factorial split similar to that produced by a pipeline using multiple calls to *split_by()*.

Higher-Order Matching

Similarly, an assumption of the *control_for()* function is that controlled-for variables can be expressed, and are stored as, vectors containing a single number for each candidate item. While this is true of most variables (and of all variables in the inbuilt dataset), some variables can only be fully expressed in multidimensional arrays. As an example, consider similarity values. The experimenter may wish to control for orthographic, phonological, or semantic distance values, constraining the extent to which matched items have similar appearances, sounds, or meanings to one another. Such values could only be stored in a matrix of size N^2. Alternatively, if the variable may be computationally costly to calculate for all possible cases, the user may wish to only calculate the variable for items which are actually considered as candidate matches by the *generate()* function.

To allow the generate pipeline to work with such multidimensional or computationally costly variables, the *control_for_map()* function was written. As the "map" suffix suggests, this is a higher-order function allowing the user to specify a function that should be called to return a

given iteration's values for the variable, given the identifier for the current iteration's item and the candidate item it is being matched to. The user-provided function may either index an existing array of data if all possible values have already been calculated, or alternatively may itself calculate only the values necessary for a single iteration.

Matching by Measures of Distance

While specifying specific tolerances for each numeric variable in *control_for()* is easily interpretable, and is more likely to reflect researchers' practice in Psycholinguistics when matching manually, an alternative approach exists which is more flexible. Rather than matching items in separate one-dimensional arrays of values, the researcher may wish to match items by measures of distance in a multi-dimensional space derived from these variables. As an example, *Match* (van Casteren & Davis, 2007) generates stimuli according to Euclidean distance calculated from an arbitrary number of variables. An advantage of matching by Euclidean distance in this way is that it penalises candidates which are extremely distant from the target in one dimension, by favouring or only permitting matches which are then also closer in other dimensions. Additionally, such a Euclidean space could be weighted so that the extent to which variables contribute to the distance between two items is proportional to their relative importance. This relative importance may be decided by experience, for example giving more weight to word length and frequency based on the prior knowledge that these variables explain the most variance in lexical decision tasks (LDTs). Alternatively, these weightings could be data-driven, drawn, for instance, from standardised Beta values of a linear model.

To support this approach to stimulus matching in LexOPS, and its combination with more variable-specific approaches to matching, the *control_for_euc()* function was written. This user-friendly wrapper for *control_for_map()* allows the user to provide a list of variables, and an optional list of weights, to control for variables in terms of Euclidean distances. If weights are not provided, the variables are by default weighted equally. This function can also be arbitrarily combined with the *control_for()* function such that a single pipeline can generate stimuli matched in a combination of controls calculated in Euclidean space and raw one-dimensional values. In addition, multiple calls to *control_for_euc()* can be used to match in multiple Euclidean spaces simultaneously. This latter use could be applied, for example, in cases where the researcher wishes to match by multiple estimates of the same features. For instance, two calls to *control_for_euc()* could be used to match word frequency as a combination of objective word frequency estimates and subjective familiarity ratings via one call to the function, while also matching for multiple estimates of word concreteness using different Likert scales via another call to the function.

2.2.3 Matching Individual Words

Finally, LexOPS permits more bespoke stimulus generation with the *match_item()* function. This function suggests possible matches for a given string (or item), within tolerances for any number of variables specified by the user. This is useful for cases when the automatic stimulus generation detailed above is unsuitable. For instance, experiments presenting

stimuli within sentences often require that matched controls for target words are semantically plausible replacements within a given sentential context, which can be difficult to quantify. The *match_item()* function will return a list of possible matches ordered by Euclidean distance (calculated from all numerical matching variables). The user can then easily select the best match that is a suitable replacement for the target word.

2.3 Inbuilt Variables

While the package can generate stimuli from any dataset provided by the user, LexOPS has a dataset already inbuilt. This dataset is not exhaustive, but is an amalgamation of several variables useful for generating word stimuli. These variables can be broadly sorted into five categories: (1) lexical, (2) orthographic, (3) phonological, (4) semantic, and (5) behavioural. Some variables were taken directly from freely available published corpora, whereas others were calculated indirectly from such sources. All built-in variables are for English words only. The package will work with variables from any language, but these need to be provided by the user.

The built-in dataset was filtered, such that word entries were excluded based on the following criteria: (1) they contained non-alphabetic characters; (2) they were longer than 28 characters; or (3) they were only observed once out of all of the word frequency corpora that were used. This left a total of 262,532 unique word strings.

2.3.1 Lexical Variables

Built-in lexical variables include word frequency and part of speech. Word frequency corpora comprise the SUBTLEX-US corpus (Brysbaert & New, 2009), the SUBTLEX-UK corpus (van Heuven et al., 2014), and the British National Corpus ("The British National Corpus, version 3 (BNC XML Edition)", 2007). Frequencies are available in LexOPS in two standardised measures: in frequency per million words (fpmw), or in the Zipf scale, calculated as Zipf = log10(frequency per billion words) (van Heuven et al., 2014). The Zipf scale is a log-normalised measure of word frequency bounded between 1 and 8, which in the context of LexOPS makes it easier to visualise and implement as an independent variable or control variable than fpmw or log(fpmw) (Brysbaert et al., 2018). The BNC frequencies were calculated by parsing the tagged xml of the latest version of the BNC. LexOPS additionally separates the written and spoken sources in the BNC, though the combined frequency across these modalities is also available.

The part of speech for a given word in LexOPS is defined as its most commonly identified part of speech within a specific corpus. Part of speech is available as a categorical variable, according to SUBTLEX-UK, the BNC, and the English Lexicon Project (ELP; Balota et al., 2007).

2.3.2 Orthographic Variables

Inbuilt orthographic variables comprise length (number of characters), bigram probability, and orthographic neighborhood size.

Character bigram probability was calculated using the word frequency corpora listed in the previous section. For each word frequency corpus, the probability of each possible character bigram (from *aa* to *zz*) was calculated by counting the number of times each bigram appears, weighted by the frequencies of the words it appeared in, in fpmw. These bigram frequencies were then scaled from 0 to 1 to get the respective probabilities of all bigrams. A word's bigram probability could then be calculated as the mean probability of all its constituent bigrams (i.e., both overlapping and non-overlapping).

Orthographic neighborhood size is available in two measures. The first is Coltheart's *N* (Coltheart et al., 1977), defined as the number of words at a Hamming distance of 1 (i.e., a one-character substitution) from a given word.. The second is Orthographic Levenshtein Distance 20 (OLD20; Yarkoni et al., 2008), defined as the mean Levenshtein distance between a given string and its 20 closest Levenshtein neighbors, where Levenshtein distance is the minimum number of character insertions, substitutions, or deletions between two strings. The OLD20 measure is generally preferable to Coltheart's *N*, as it allows for distance calculation between strings of different lengths and better accounts for behavioural correlates of orthographic neighbourhood density (Yarkoni et al., 2008). Both of these measures were calculated using the R package, "vwr" (i.e., "visual word recognition"; Keuleers, 2013).

2.3.3 Phonological Variables

The inbuilt phonological variables of LexOPS comprise the following: number of phonemes, number of syllables, number of pronunciations, rhyme, and phonological neighborhood size. The phonological features were calculated using phonetic transcriptions from two different sources: the eSpeak speech synthesiser's ("eSpeak version 1.48.15", 2015) standard British English pronunciations of all entries in the database; and the Carnegie Mellon University (CMU) Pronouncing Dictionary of American English (Weide, 2014).

The transcription system adopted by eSpeak (Kirshenbaum phonetic encoding) uses one-character ASCII representations for individual phonemes, but two-character representations for affricates and diphthongs. The affricates /ʧ/ and /ʤ/ (as in the beginnings of *char* and *jar*, respectively) are encoded with one-character ASCII representations. The CMU transcriptions are represented by an ARPAbet transcription system for American English, and are represented as either two-letter or one-letter ASCII characters. For example, the word *how*, containing the diphthong /aʊ/, is represented as 'HOW' in the two-character system or as 'hW' in the one-character system. Similarly, the word *China*, containing the affricate /ʧ/, is represented as 'CH-AY-N-AE' (with phonemes separated by hyphens) in the two-character system, or as 'CYN@' in the one-character system.

Number of phonemes is simply a count of how many phonemes a word contains. The number of syllables was calculated by simply counting the number of vowel phonemes that occurred in the transcription. The number of pronunciations is a variable only available for the

CMU Pronouncing Dictionary, calculated by counting how many possible pronunciations are listed for each entry. This includes differences in both pronunciation and stress patterns.

Rhyme is represented as a categorical variable consisting of a transcription of all phonemes from the final vowel phoneme until the end of the word (i.e., the final syllable's 'rime'). For instance, eSpeak's British English pronunciation of *partake* is represented as /pɑteɪk/ in the International Phonetic Alphabet (IPA) and, as such, belongs to the rhyme category of /-eɪk/, which it shares with entries such as *steak* and *opaque*.

Phonological neighborhood size is available in terms of the phonological Coltheart's N and Phonological Levenshtein Distance 20 (PLD20), calculated similarly to the orthographic neighborhood measures, using the "vwr" package for R (Keuleers, 2013).

2.3.4 Semantic Variables

Semantic features which LexOPS has built-in mostly come from norming studies in which participants provide ratings for a particular semantic aspect of a word on a Likert scale. A summary of the available semantic features is presented in Table 2.1.

2.3.5 Behavioural Variables

Behavioural variables consist primarily of lexical decision response time and accuracy from the ELP (Balota et al., 2007) and the British Lexicon Project (BLP; Keuleers et al., 2012).

Behavioural variables also include measures of proportion known (the proportion of people who know a given word) and word prevalence (probit-transformed proportion known), taken from (Brysbaert et al., 2019). Brysbaert et al. (2019) demonstrate that proportion known and word prevalence have advantages over variables such as word frequency, age of acquisition, and familiarity (which have traditionally served as proxies to gauging word difficulty) since these two measures more directly operationalise word difficulty.

2.4 The Shiny App: An Interactive User Interface

LexOPS features a GUI in the form of a Shiny app (W. Chang et al., 2018), which provides an interactive front-end to the package's functions. For instance, tolerances for independent variables can be specified via a slider (i.e., a moveable graphical button on an analogue scale), and are then visualised as shaded areas in a plot of a variable's density. Figure 2.2 presents such an example for defining experimental conditions in the *split_by()* function. The "generate pipeline" is accessible through a "Generate" tab in the sidebar, while the *match_word()* function is accessible through a "Match Word" tab. Interactive functionality is also provided for querying the LexOPS dataset (through the "Fetch" tab), and for integrating custom variables or datasets into the app (through the "Custom Variables" tab). The Shiny app's GUI is likely to be more accessible for users unfamiliar with R, as it can be run with a minimal amount of R code with the *run_shiny()* function, though the speed and ease with which it allows for stimulus generation

Table 2.1: Summary of the sources and semantic features used in LexOPS

Source and Semantic Feature	Scale	*N* Words	Observations/Word[a]
Scott et al. (2019)			
FAM	1-7	5553	30.58 (3.71)
AOA	1-7	5553	33.7 (3.72)
CNC	1-7	5553	32.71 (3.85)
AROU	1-7	5553	32.71 (3.74)
VAL	1-9	5553	33 (3.76)
DOM	1-9	5553	32.6 (3.78)
IMAG	1-9	5553	32.6 (3.8)
SIZE	1-7	5553	32.78 (3.84)
GEND	1-7	5553	33.33 (4.03)
J. M. Clark and Paivio (2004)			
FAM	1-7	2311	16
IMAG	1-7	2311	47-49
Kuperman et al. (2012)			
AOA	ages 1-25	30124	18-22 for most items
Brysbaert and Biemiller (2017)			
AOA[b]	ages 2-14[c]	43991	Around 200
Brysbaert et al. (2014)			
CNC	1-5	37058	$\geq$25
Warriner et al. (2013)			
AROU	1-9	13915	22.97 (23.73)
VAL	1-9	13915	21.81 (23.44)
DOM	1-9	13915	24.32 (25.07)
Engelthaler and Hills (2018)			
HUM	1-5	4997	32.93 (5.64)

For each source, the relevant semantic feature(s), scale, number of words, and observations per word are specified.
AROU arousal; *VAL* valence; *DOM* dominance; *CNC* concreteness; *IMAG* imageability; *FAM* familiarity; *AOA* age of acquisition; *SIZE* semantic size; *GEND* gender association; *HUM* humor.

[a] Where the number of observations for each word was available, the mean, and standard deviation in parentheses, are presented; otherwise, summary statistics are reported. [b] This measure is test-based, not from a rating study. [c] Age estimates cover ages 2, 4, 6, 8, 10, 12, 13, and 14.

make it a convenient feature for all users. Furthermore, the Shiny app automatically translates the user's selections into reproducible R code that can then be run as a stand-alone R script.

In addition to providing an interface to LexOPS functions, the Shiny app also provides an interface in its "Visualise" tab for interactive visualisation of relationships between variables, and the distribution of generated stimuli across variables. Here, users can select variables to plot on x- and y- axes, and can optionally elect to plot variables on a z-axis or color scale. LexOPS will generate an interactive scatter plot of all words which have a value for all requested variables, where each point represents a single word. By hovering with the cursor over a given point, the user can query the word visualised at that location as well as its specific values (coordinates) across the plotted variables.

Whereas axes can only be used to visualise numerical values, color scales can be used to visualise the distributions of variables which are either numerical or categorical. For instance, the user can select to view the distributions of different parts of speech by means of differential coloring of the defined levels of this variable. The user can also have the app visualise distributions of stimuli produced by the Generate tab, as shown in Figure 2.3, as well as suggested matches produced by the Match tab, or words uploaded to the Fetch tab.

2.5 Example Applications

2.5.1 Psycholinguistic Stimuli

Fully Automated Word Selection

As an example, a user could define a 2 x 2 design to investigate the interaction between character bigram probability, according to SUBTLEX-UK, and concreteness ratings, according to Brysbaert et al. (2014). The user could also specify that stimuli should be controlled across conditions for word frequency within ±.2 Zipf according to SUBTLEX-UK, as well as exact word length. The dataset that stimuli are generated from can be additionally filtered, for instance according to word prevalence reported by Brysbaert et al. (2019) such that the generated stimuli consist entirely of words that at least 90% of people know. The following R code will generate 50 words per factorial cell (200 in total) that fit these specifications. The variables used in this example have all been drawn from the inbuilt dataset described in the previous section to make the code more easily readable and reproducible.

```
stim <- lexops %>%
  subset(PK.Brysbaert >= 0.9) %>%
  split_by(BG.SUBTLEX_UK, 0:0.003 ~ 0.009:0.013) %>%
  split_by(CNC.Brysbaert, 1:2 ~ 4:5) %>%
  control_for(Length, 0:0) %>%
  control_for(Zipf.SUBTLEX_UK, -0.2:0.2) %>%
  generate(n = 50)
```

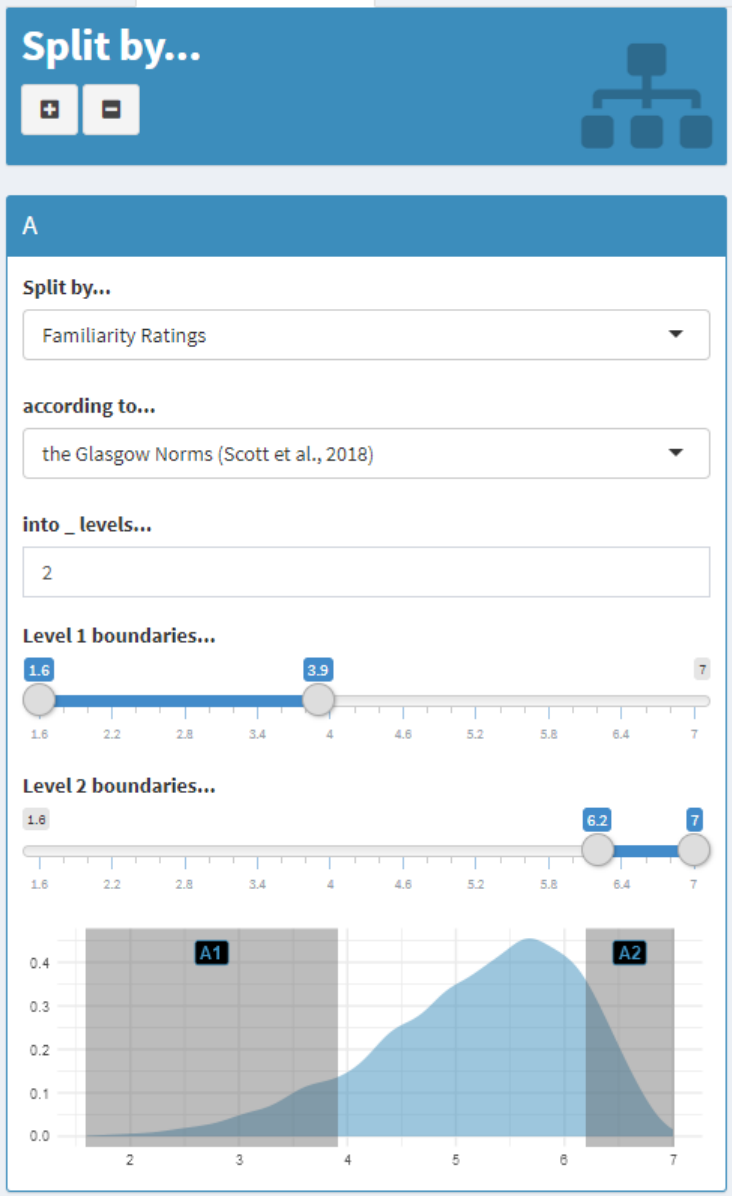
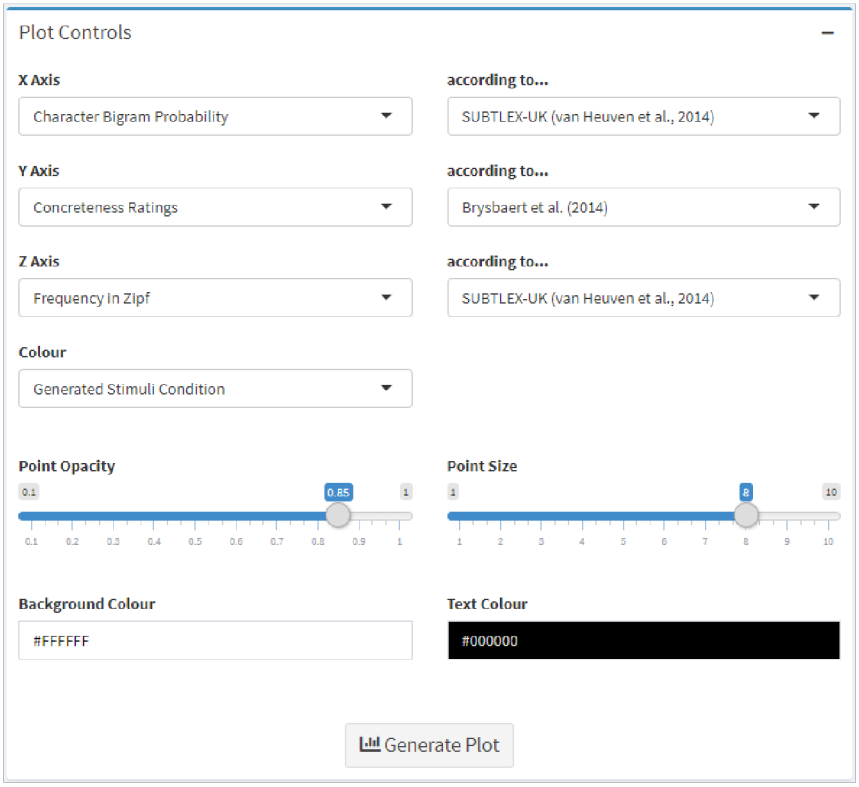


Figure 2.2: An example box for specifying the levels of an independent variable in the Shiny app. Here, two levels (A1, A2) are being specified for the variable of Familiarity from the Glasgow Norms (Scott et al., 2019). In this case, the density plot shows that the distribution is skewed towards words rated as more familiar, with far fewer words rated as less familiar. As such, it might make sense to use a wider range or bin for a low familiarity condition, to ensure there are enough candidate words. Similar boxes are used for specifying controls and filters. Such boxes can be added to or removed from the design specification with the plus and minus buttons, respectively.

A



B

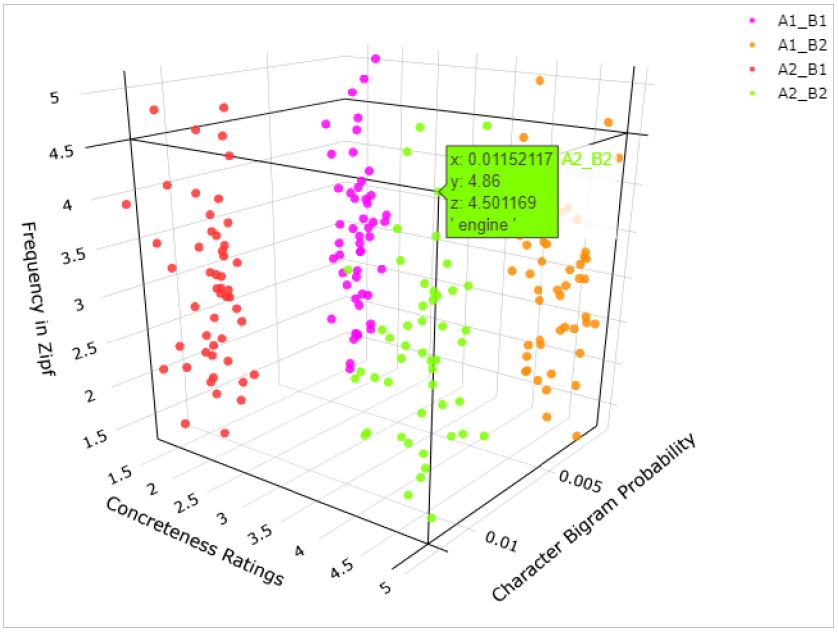


Figure 2.3: Example showing (A) user interface options and (B) resulting interactive plot produced by the Visualise tab, for stimuli generated by the "generate pipeline" specified by the code in the Example Applications section (2 x 2, character bigram probability by concreteness design, controlling for length and frequency). Each point corresponds to one word, which can be queried by the user by moving the cursor directly over that point. In the example, the user has queried a word from condition A2_B2, corresponding to the word, "engine".

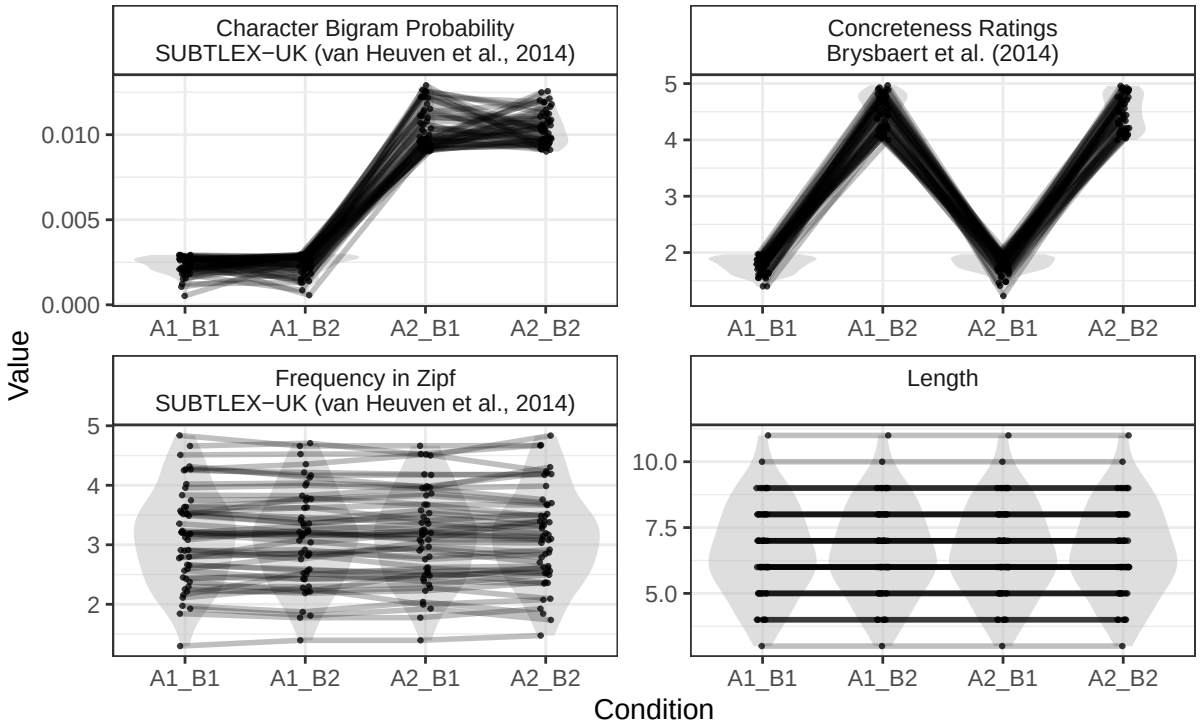


Figure 2.4: An example figure generated by the *plot_design()* function, for a stimulus list generated by the example code, consisting of 200 words split into four factorial cells: A1_B1 (low bigram probability, low concreteness), A1_B2 (low bigram probability, high concreteness), A2_B1 (high bigram probability, low concreteness), and A2_B2 (high bigram probability, high concreteness). In this example, words are controlled in terms of frequency (within ±.2 Zipf), and length (exactly). When words are more closely matched on a variable, the distributions of control variables appear more similar, and the slopes of lines between matched items are less steep. The differences between conditions in character bigram probability and concreteness ratings (sought by the user) are reflected in the upper two plots.

The distributions of generated stimuli on relevant numerical variables can be readily examined using the *plot_design()* function. Figure 2.4 presents an example figure generated by the *plot_design()* function for a stimulus list generated by the code above. This function produces a multi-faceted figure showing the distributions (in violin plots) of all numeric independent or control variables used for each generated condition. Within each distribution, individual words are visualised as points, joined by lines to other words (points) from the same matched set (i.e., that share the same match-null). Such a figure can be a convenient way to check that LexOPS has generated stimuli as expected. For instance, excessive differences between generated conditions in the distributions of control variables may indicate that more restrictive tolerances might be appropriate.

In addition, the user may be interested in how representative their stimuli are in the variables they are interested in. To support visualisation of this, the *plot_sample()* function compares the generated stimuli's distributions on all variables in the design of the generated stimuli to those in the pool of possible candidates from which they are drawn. An example is presented in Figure 2.5.

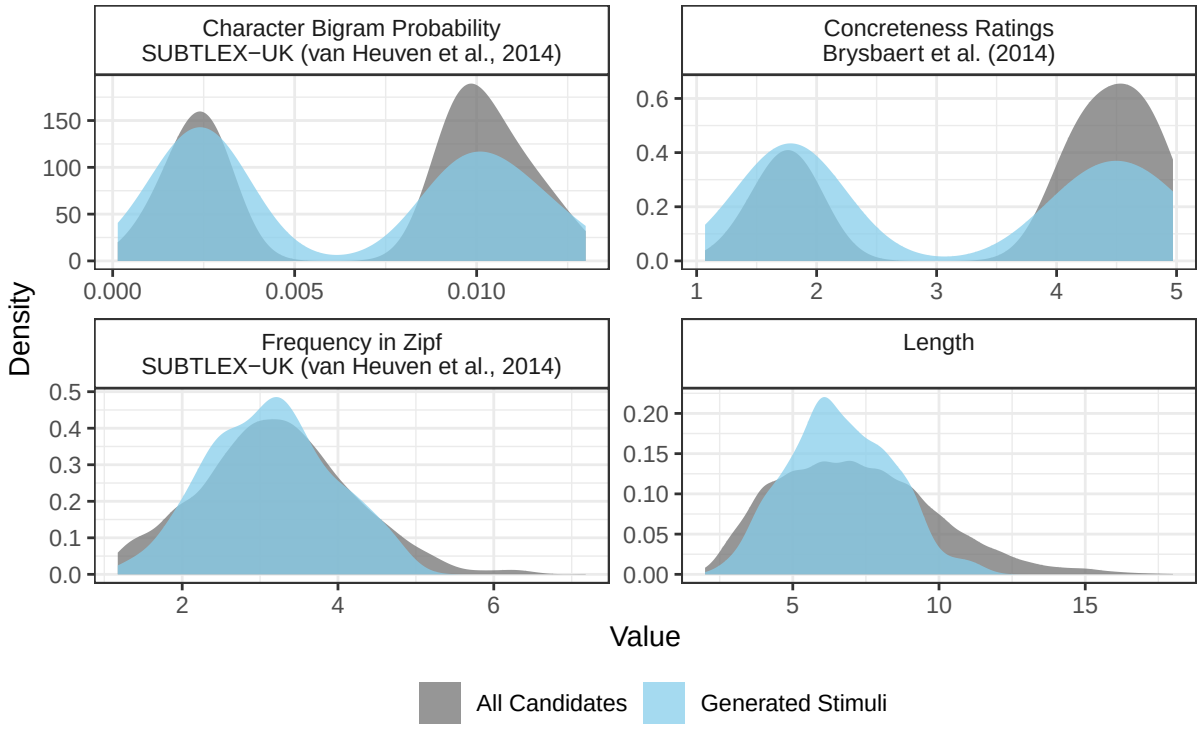


Figure 2.5: An example figure generated by the *plot_sample()* function, for a stimulus list generated by the example code. The distributions of the 200 generated items are presented in blue, while the distributions of all candidate words available to the LexOPS algorithm (which includes those generated) are presented in grey.

Finally, the user can examine the algorithm's performance by looking at the output of the *plot_iterations()* function. This function generates a plot showing the cumulative number of items generated for each iteration of the algorithm. As LexOPS begins to exhaust the pool of possible candidates available for the specified design, this plot's line will typically "flatten-out" and begin to show a logarithmic relationship. Example output from the *plot_iterations()* function is presented in Figure 2.6.

Supervised Word Selection

The *match_item()* function is convenient in cases where matches need to be controlled for factors that would be difficult to operationalise as numeric or categorical variables, such as maintaining sentence plausibility when a target word is replaced, such that an expert needs to supervise the word selection with these rules in mind. As a practical example, imagine an experiment where the researcher wants to replace target words in existing sentences with words having a later age of acquisition. Suppose they also want the words to be controlled for length, frequency, concreteness and part of speech (according to the written texts of the BNC). If the researcher wanted to find a suitable replacement for the word "butterfly" in the sentence, "The man looked up - he had never seen such an enormous butterfly before", they could use the following code to identify a suitable match. Again, all the variables used have been drawn from the inbuilt dataset for readability.

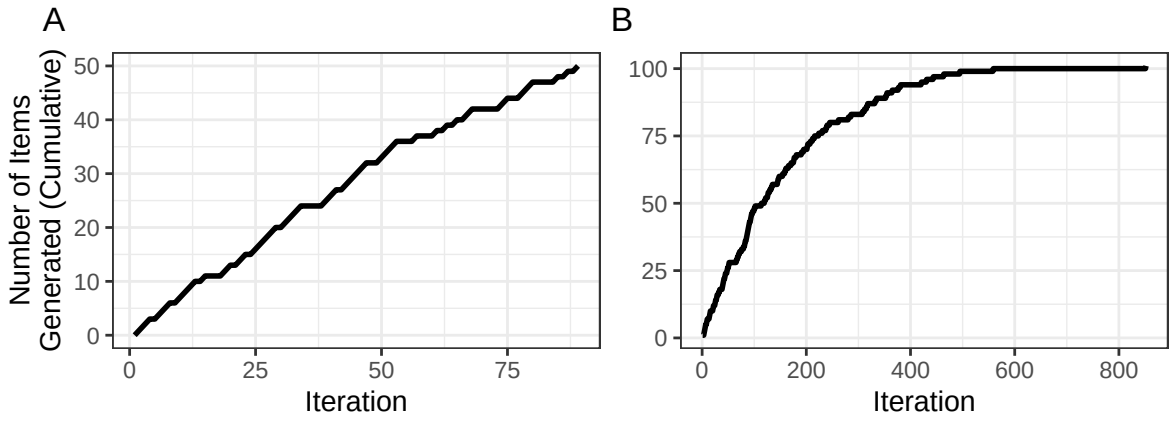


Figure 2.6: An example figure generated by the *plot_iterations()* function, for a stimulus list generated by the example code. To highlight how the plot may be interpreted, panel A shows the results from the example stimuli generated, while panel B shows results from a pipeline with the same design, but generating as many stimuli as possible (i.e. until the pool of candidate matches is exhausted).

```
stim <- lexops %>%
  match_item(
    "butterfly",
    Length,
    Zipf.SUBTLEX_UK = -0.2:0.2,
    CNC.Brysbaert = -0.25:0.25,
    PoS.BNC.Written
  ) %>%
  subset(AoA.Kuperman >= 9)
```

This would return a data frame containing four possible matches, ordered by Euclidean distance in the matching variables: "satellite", "orchestra", "champagne", and "machinery". Of these, the researcher would probably select the word "satellite", as the closest match that is a plausible replacement for "butterfly" in the example sentence.

2.5.2 Applications Beyond Psycholinguistic Stimuli

Although LexOPS was developed primarily for experiments employing word stimuli, the package can also be used to generate stimuli in any experimental domain for which there is a finite set of possible stimuli, having properties that have been coded numerically or categorically. For example, the Chicago face database (Ma et al., 2015) is a resource that specifies both objective and subjective measures of a set of faces. LexOPS could be used on this database to generate stimuli to investigate, for example, a possible effect of attractiveness on face recognition processes. Analogous to its functionality with words, LexOPS could easily be adapted to define levels of facial attractiveness, while controlling for variables such as the race, gender, and luminance of individual faces.

Furthermore, in addition to generating matched stimuli from any database, LexOPS could feasibly be applied to generate matched items of any form of entry in a database. An example, applicable to a large portion of scientific research using human participants, is participant selection. A common requirement of between-subject designs is that participants are matched across conditions on relevant variables such as age, sex, and socioeconomic status. LexOPS could easily be run on a database of participants to generate lists of participants matched across conditions for such variables, within desired tolerances.

2.6 Validation

To demonstrate that the package is a valuable tool for generating word stimuli, I tested whether LexOPS could produce stimulus sets comparable to those of previous studies that employed well-controlled word stimuli. Four studies were selected based on the following criteria: the experimental design was unambiguously presented (e.g., with clear definitions and/or boundaries of conditions); the characteristics of stimuli (e.g., concreteness, valence) were taken from freely available published norms; the stimuli across conditions were matched on an item-by-item basis; and the complete set of stimuli was provided. The first study, by Kousta et al. (2011), examined concreteness (high/low), using 38 words per condition, and controlling for 12 different psycholinguistic variables. The second study, by Scott et al. (2009), investigated the interaction between word frequency (high/low) and emotional valence (negative/neutral/positive), using 40 words per each of the six conditions, and controlling for word length and frequency. The third study, by Sereno et al. (2015), employed a similar frequency (high/low) by emotion (negative/neutral/positive) design, with a different set of 40 words per condition, and similarly controlled for word length and frequency. Finally, B. Yao et al. (2018) examined the interaction between concreteness (high/low) and emotion (negative/neutral/positive), using 45 words per factorial cell, and controlling for word length and frequency.

For each study, I used LexOPS to generate the same number of stimuli according to the original constraints that had been specified. I used the same databases that were detailed within the studies with one exception (the norms for one of Kousta et al.'s control variables, context availability, were obtained locally for that study and not made freely available). In all cases, LexOPS was able to generate stimuli that fit within the boundaries of the original conditions, which were matched at least as closely on all control variables. In many cases, it was found that closer tolerances on many variables were possible than those implemented in the original studies. To encapsulate the comparison between the original stimuli and those generated by LexOPS, for both lists the Euclidean distance in all numeric control variables (scaled by standard deviation for comparability) was calculated between each word in the list, and each word it should be matched to. As the controls were implemented item-wise, this resulted in $n\frac{k(k-1)}{2}$ observations of Euclidean distance for each stimulus list, where n is the number of items per factorial cell, and k is the number of factorial cells. The calculated values are presented in Figure 2.7.

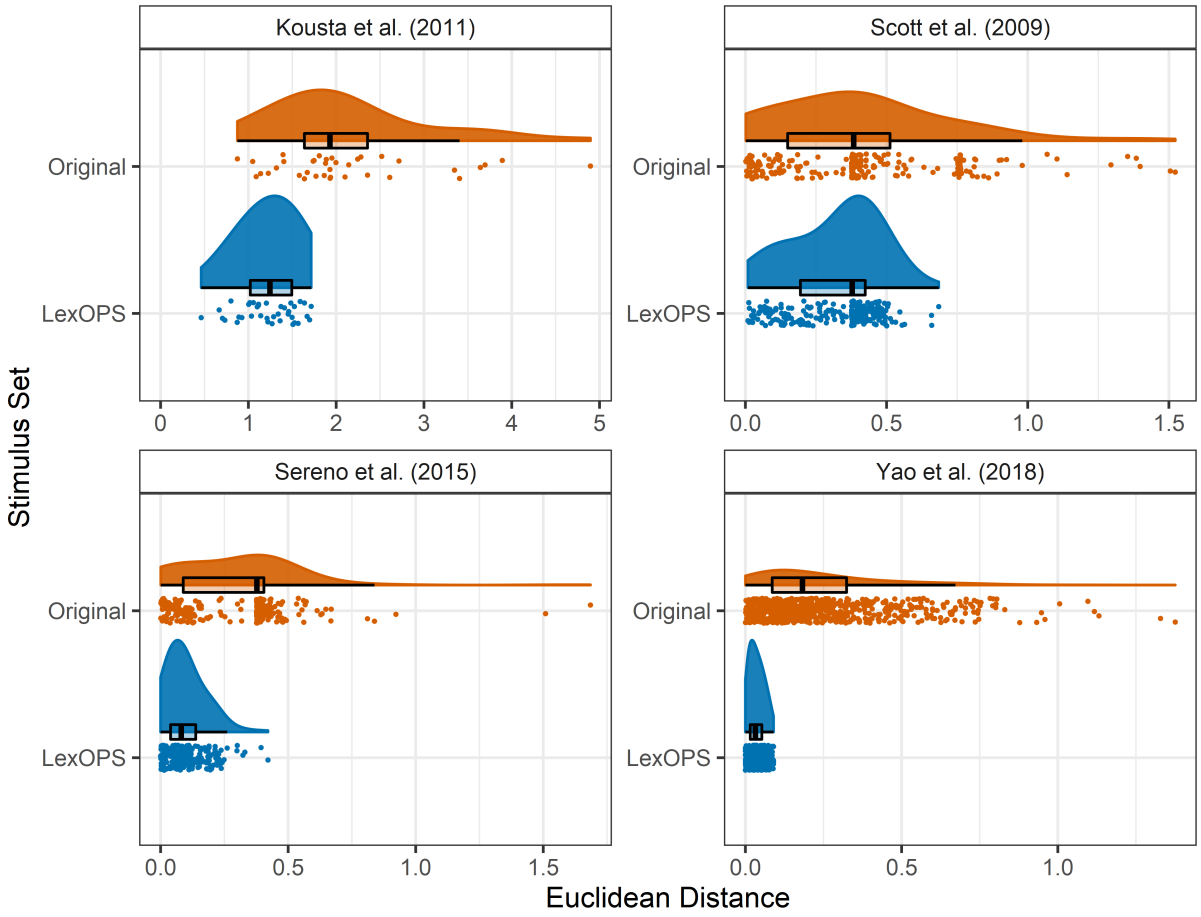


Figure 2.7: Summary of the results from the validation analysis of LexOPS. The Euclidean distance values between each matched pair of words in the four studies, for the original study (*in orange*) and the stimuli generated by LexOPS (*in blue*). Each point represents a single value of distance, while the density plot above depicts the shape of the distribution. The overlaid boxplots present summary statistics of the median (*central, dark vertical line*), first and third quartiles (*the left- and right-most ends of the boxes*) and the range of the values, bounded to within a distance of 1.5 times the interquartile range from the boxes (*the whiskers*). The bands of points seen in the values for Scott et al. and Sereno et al. reflect that stimuli from these studies were allowed to differ in length. The bands are absent in the distance values for the LexOPS stimuli generated for Sereno et al., as these stimuli were matched for length exactly.

2.7 Contributions to Replicability and Reproducibility

LexOPS offers a valuable contribution to research in terms of reproducibility and replicability. By sharing LexOPS code, for example in existing repositories such as the Open Science Framework and GitHub, researchers can provide the exact specifications, in readable code, used to generate stimuli lists that were found to produce a given effect. Moreover, the code can include a random seed that allows other users to reproduce a specific stimulus list. If a random seed is not set, or is set to a different value, a given pipeline will generate a different set of stimuli each time it is run. This means that an experimental design can be replicated, with the same relationships between variables, and same precision in matching tolerances, but consisting of different stimuli. Other users can also modify shared code to see how such changes in the experimental design might alter a reported effect, for instance, by modifying the cut-off values of a variable's levels or the tolerances of control variables, or by including additional control variables.

2.8 An Alternative Approach: Distribution-Wise Matching

The method of matching in LexOPS is exclusively item-wise; each item from factorial cell x is matched on a controlled variable, within a specified tolerance, with one item from factorial cell y. Distributional similarity in variables that are matched between factorial cells is a necessary by-product of item-wise matching, with the degree of distributional similarity dictated by the stringency of the item-wise tolerance. Item-wise matching has the additional advantage of providing items for each condition which are directly comparable. Another approach to matching, however, could centre on distribution-wise matching, maximising distributional similarity directly, without generating item-wise matches.

2.8.1 Parametric Distribution-Wise Matching

In a simple case, a split which maximises the distributional similarity between two conditions on a normally distributed variable could be identified by minimising the difference between the conditions in the two parameters that define the normal distribution: μ (the mean) and σ (the standard deviation). A simple way of maximising similarity in these parameters could be to randomly perform the split using a large number of random seeds and recording each split's values for each distribution. The best split could then be identified as that with the minimum distance between the generated conditions' mean and standard deviation values. This approach could be applied to any possible distribution. For example, similarity could be maximised in terms of the Weibull distribution by minimising distance in the parameters of β (shape), η (scale), and γ (location).

There are, however, some clear issues with matching distributions by their parameters. Variables rarely conform perfectly to these parametric assumptions, and their distributions can be highly dependent on the filtering criteria used for other variables (although researchers may decide that artificially imposing parametric assumptions is desirable for their stimuli, e.g.

Solomyak & Marantz, 2010). Even if a variable conforms strongly to distributional assumptions in a population, random samples drawn from this population can by chance differ greatly from their population and from each other, especially if samples are very small relative to the population. Moreover, matching by distributional parameters requires identification of a distribution that describes well each variable that is to be matched, while some variables' distributions are so unusual that they would require considerable time and effort to tailor mathematical parametrisation (as, for example, in the multimodal distribution of OLD20; see chapter 6).

2.8.2 Assumption-Free Distribution-Wise Matching

An ideal solution, therefore, would allow the identification of suitable splits by maximising a measure of distributional similarity that makes no parametric assumptions about the distributions being compared. Suggested measures satisfying this description include the Kolmogrov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1948) and the Q statistic (Wilcox & Muska, 1999). A recently proposed method of measuring distributional similarity in an assumption-free manner is outlined by Pastore and Calcagnì (2019) in the form of the overlapping index. Implemented in the R package, *overlapping* (Pastore, 2018), the overlapping index makes it possible to quantify the degree of overlap between empirical distributions.

As various existing solutions for calculating distributional similarity, including the *overlapping* R package, can be applied to stimulus matching with a minimal amount of code, and since the scope of LexOPS is focused on item-wise matching, no functionality is provided within LexOPS for generating stimuli with distribution-wise matches. Nevertheless, code maximising distributional similarity could be combined with the item-wise approach of LexOPS with relative ease. To demonstrate this, and to highlight differences between item-wise and distribution-wise matching, the following LexOPS pipeline was written. The imagined study uses 200 matched pairs of abstract and concrete words, controlled for length and frequency item-wise.

```
stim <- lexops %>%
  split_by(CNC.Brysbaert, 1:2 ~ 4:5) %>%
  control_for(Zipf.SUBTLEX_UK, -0.1:0.1) %>%
  control_for(Length, 0:0) %>%
  generate(200)
```

In addition to the item-wise matching native to LexOPS, additional distributional controls may be implemented in these stimuli. As a demonstration, I matched the design specified above by three additional variables in a distribution-wise manner: age of acquisition (Kuperman et al., 2012), character bigram probability in SUBTLEX-UK, and OLD20. To achieve this, I used a Monte-Carlo approach, generating 3000 unique stimulus sets with the above LexOPS pipeline, providing unique random seeds in the *seed* argument of the *generate()* function for each set. To ensure no words were selected with missing data for the distribution-wise matches,

the inbuilt dataset was filtered at the start of the pipeline such that all candidate items had observations for each distribution-wise control. After generating each unique stimulus set, I calculated the overlapping index between the concrete and abstract conditions' distributions for the three distribution-wise controls in each stimulus set. To maximise distributional similarity in the desired variables, I could then select the stimulus set with the largest total overlap by summing (or, equivalently, averaging) over each iteration's overlapping index values for each distribution-wise control.

The distributions on the independent variable, and on the item-wise and distribution-wise controls, are presented for an example stimulus set in Figure 2.8. To further highlight the difference between item-wise and distribution-wise controls, the correlations between the matched pairs' values, in each of these variables, are presented in Figure 2.9. These two plots demonstrate that whereas item-wise controls show high similarity between both distributions and matched items, distribution-wise controls show similarity in distributions, but less similarity between matched items. A noticeable exception is OLD20, which shows a moderately strong correlation between matched pairs' values even though OLD20 was only controlled distribution-wise. This item-wise similarity in orthographic density most likely arose from the item-wise matching of length, as word length and orthographic neighbourhood density are highly correlated (see Figure 2.10), since character addition and subtraction are costly operations in the calculation of Levenshtein distance.

To evaluate the reproducibility of the method combining LexOPS' item-wise approach with distribution-wise controls, the pipeline was run 12 times. The overlapping index values between the abstract and concrete words' values for the distribution-wise controls, for each of the 3000 iterations, in each of the 12 runs, are presented in panel A of Figure 2.11. As mentioned, controlling for multiple variables in a distribution-wise manner requires that the stimulus set with the best overall or average distributional overlap is selected. As panel B of Figure 2.11 demonstrates, this introduces more variability between runs of the algorithm than maximising distributional similarity in a single variable would, and can decrease the maximum overlap value that may be achieved in any single distribution-wise control. This is due to a combination of an increase in the degrees of freedom (as more parameters are being optimised), and the constraints imposed by non-orthogonalities between the variables in the population (e.g. words with earlier age of acquisition may have systematically higher character bigram probabilities). Nevertheless, it is shown that reasonably high distributional similarity, within the constraints of the experimental design and item-wise matching specifications, can be achieved when matching by multiple variables within only a few thousand iterations. Due to the aforementioned changes in the number of parameters and the non-orthogonalities between them, the number of iterations required to elicit similar performance in such an algorithm will increase with each additional distributional control that is added.

To summarise, an alternative approach to matching stimuli matches features in variables' distributions. This can be achieved while making minimal assumptions about distributional shape. Although LexOPS is explicitly item-wise in its approach, it can be flexibly combined with distribution-wise approaches to produce more carefully matched stimulus sets.

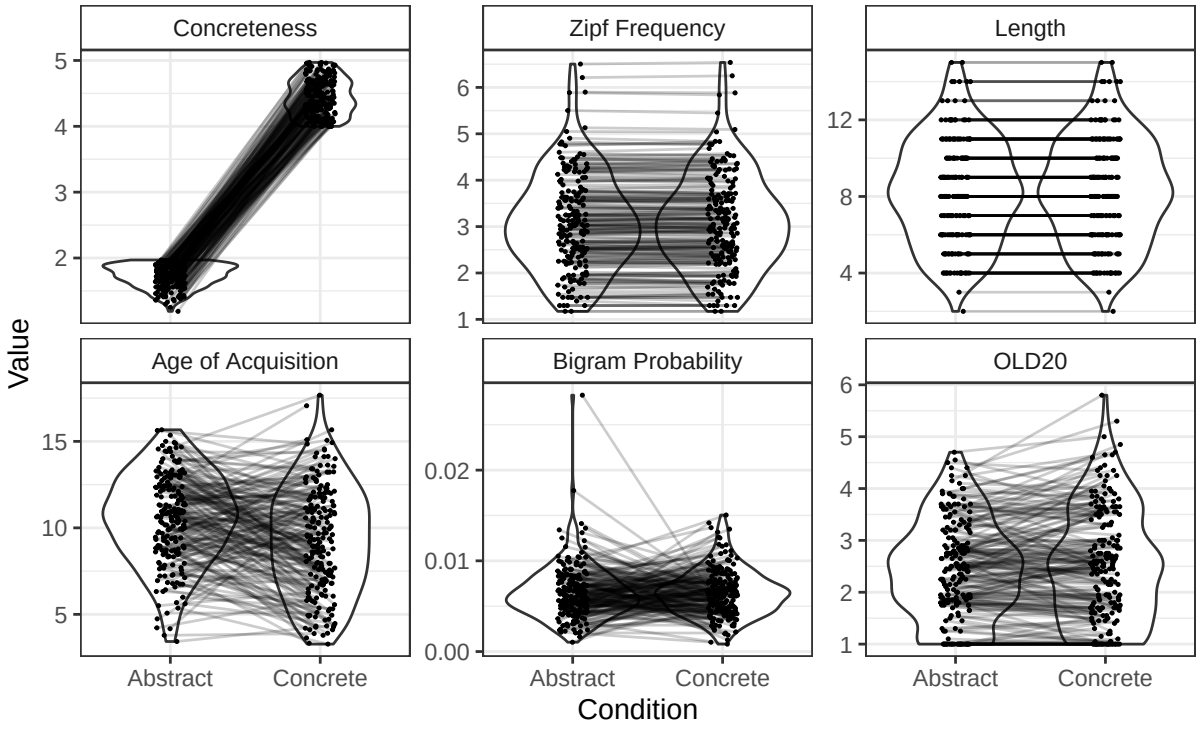


Figure 2.8: The distributions on relevant variables of an example stimulus set, generated from one run of the combined item-wise and distribution-wise matching algorithm, in the abstract and concrete conditions. The top three panels depict the distributions in concreteness (the independent variable), and frequency and length (which were matched item-wise). The bottom three panels depict the distributions in the three variables which were matched distribution-wise. Violin plots depict variables' densities, trimmed to the maxima and minima. Matched items are depicted as points joined by lines, with position on the x axis jittered randomly for visibility.

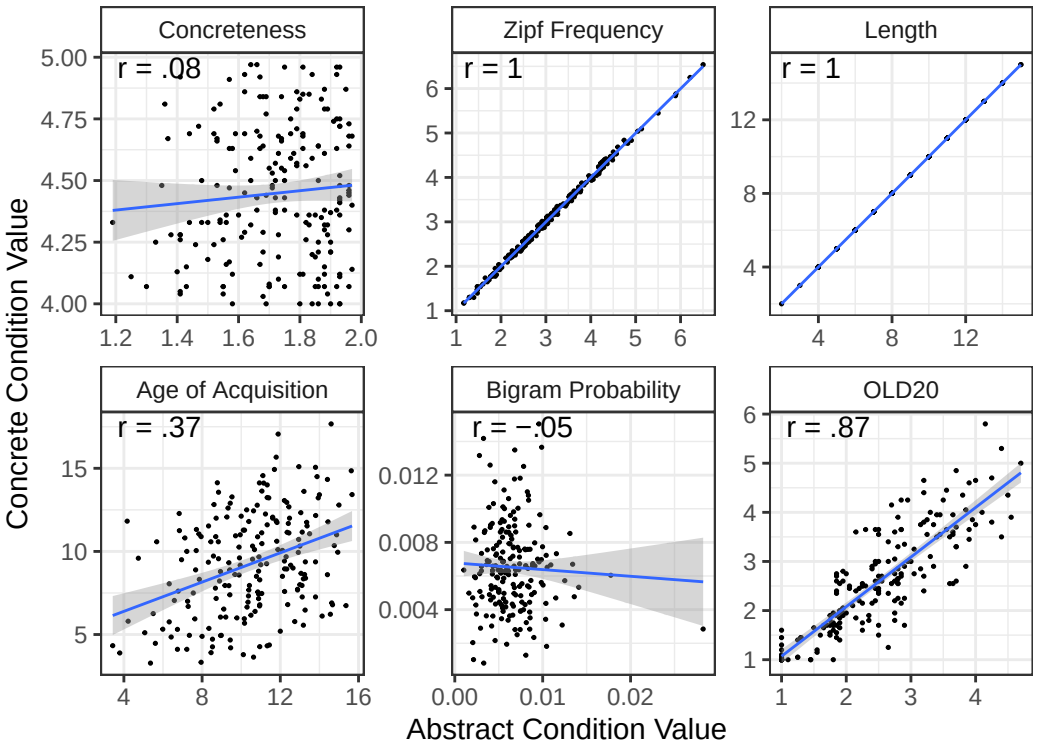


Figure 2.9: The correlation between matched concrete and abstract items' values in relevant variables of the example stimulus set depicted in Figure 2.8. Individual observations are depicted as points. Pearson's *r* values are presented in the top-left corner of each plot. The respective linear relationships are depicted blue lines, with grey shaded areas depicting 95% confidence intervals.

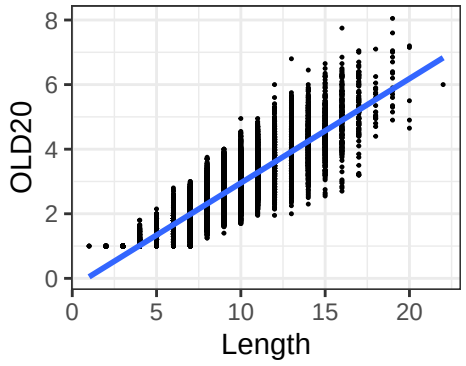


Figure 2.10: The relationship between length and OLD20 for all words in the Brysbaert et al. (2019) concreteness norms (this filter was applied as all words in the generated stimuli were drawn from this corpus). The blue line depicts the positive linear relationship (*r* = .89).

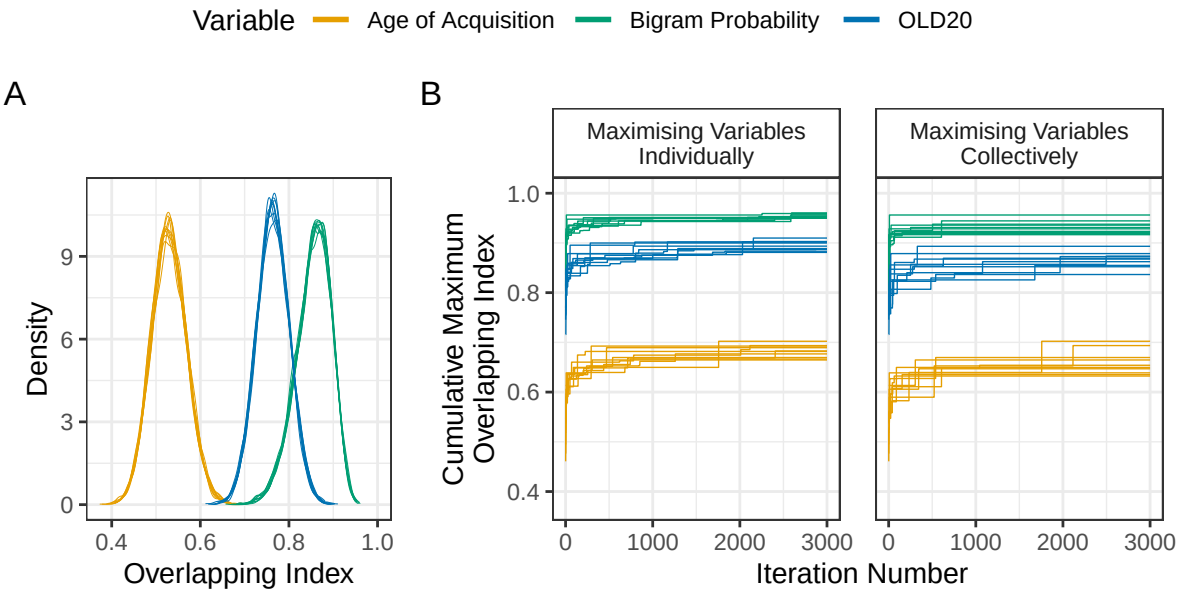


Figure 2.11: The overlapping indices generated from 12 runs (superimposed) of the combined item-wise and distribution-wise matching algorithm, with 3000 iterations (observations) in each run. Each iteration generated a unique random list of stimuli fitting the specified LexOPS pipeline. The left panel (A) shows the densities of overlapping index distributions observed from all iterations on each run, for each variable. The two right panels (B) shows the cumulative maximum value overlapping index over iterations for each variable on each run. The final value of this latter variable for each run indicates the overlapping index which would then be observed in the selected stimulus set. Results for panel B are split into (left) the values observed when maximising overlap for variables individually, and (right) the values observed when maximising total overlap across all three variables. For the former, three stimulus sets are generated for each run, controlling distributionally for one variable only. For the latter, a single stimulus set is generated controlling for all three distribution-wise controls concurrently.

2.9 Discussion

LexOPS is a valuable resource to researchers who use matched sets of stimuli, providing a method for flexible and controlled generation of items, with the added value of intuitive interfaces. In addition, LexOPS facilitates the reproducibility and replicability of experiments, allowing specific stimulus lists to be recreated, and providing an easy method for generating novel stimulus lists for the purposes of replication or validation. Furthermore, its flexibility allows its algorithm to be combined with other methods of matching stimuli, such as implementations of distribution-wise matching.

One point that should not be overlooked is that both the item-wise matching of LexOPS and distribution-wise matching suggested as a less constrained alternative are inherently limited by the nature of the variables, tolerances, and condition boundaries that are used. For instance, some variables have entries for relatively few words (e.g., the familiarity rating norms from J. M. Clark and Paivio (2004) include ratings for only 2311 words), and there is often limited overlap of items between different corpora. This means that if variables from small corpora, or from multiple corpora with little overlap, are used as independent or control variables, the pool of possible stimuli will be greatly reduced. Similarly, variables are often highly correlated, as, for instance, imageability and concreteness are (Scott et al., 2019). It would be difficult to generate stimuli, matched item-wise, for designs probing interactions between such highly correlated variables, or for those in which independent variables and control variables are highly correlated. Finally, the precision of control variables' tolerances, and the positioning of independent variables' boundaries relative to the variables' density distributions, will also modulate the number of possible stimuli that can be generated.

A similar caveat exists for other methods of generating matched stimuli, such as distribution-wise matching. This is a point well demonstrated by the matched design outlined in the section on assumption-free distribution-wise matching. The strong correlation between words' concreteness and age of acquisition ratings (Scott et al., 2019), likely reflecting the later age of acquisition of abstract words relative to concrete words, means that the overlap observed between concrete and abstract words' age of acquisition ratings will always remain low regardless of how many iterations are run. The lack of any consistent relationship between concreteness and character bigram probability, meanwhile, means that the overlapping index values observed by chance are likely to be relatively high in any iteration. As with the item-wise matching used in LexOPS, results will also be similarly constrained by relationships between controls.

While the features of LexOPS detailed here are unlikely to change, work will continue on the package, and it is very likely that I will add extra functionality to LexOPS in the future in response to users' requests. Similarly, the inbuilt database may be expanded to include further variables if they are likely to be of use to many researchers. Any such additions or changes will be described in the package's documentation and in the LexOPS walkthrough.

To conclude, I have developed and made freely available a flexible and intuitive tool for the controlled generation of matched stimuli, with a focus on word stimuli. This R package allows researchers to generate robustly lists of matched stimuli for factorial designs in a reproducible

and replicable manner. The approach is flexible, and can be easily combined with other approaches to controlling for confounding variables. LexOPS has potential to be of great benefit to a broad range of researchers, particularly those who use word stimuli.

Chapter 3

Rating Norms should be Calculated from Cumulative Link Mixed Effects Models

3.1 Introduction

In a typical rating norming study, participants are asked to rate features of stimuli on Likert scales (e.g., on a scale from *1* to *7*). These ratings are used to estimate how participants perceive these features. Such estimates may be used to validate stimuli for an existing experiment, design new stimuli, or correlate with observations of behaviour or neural activity. For the latter two purposes, the estimates are often made public for use by other researchers alongside dedicated publications. Examples include but are not limited to ratings on various dimensions, for stimuli as diverse as words (Brysbaert et al., 2014; Scott et al., 2019; Warriner et al., 2013), orthographic characters (Simpson et al., 2013), photographs of objects (Brodeur et al., 2014) or faces (Ma et al., 2015), and melodies (Belfi & Kacirek, 2021). Such norming studies are typically summarised via per-item statistics of means and standard deviations (*SD*s) of the ratings for each item. In this chapter, I argue that ordinal models can provide more robust measures of item norms, quantifying dimensions more meaningfully, for purposes of statistical analysis and for application in different methods of stimulus design (e.g., chapter 2). I focus on cumulative link mixed effects models (CLMMs), showing that they can yield summary statistics analogous to the traditional estimates of means and *SD*s, but disentangled from artefacts of nonlinearities in participants' response patterns.

Datasets of norms typically report, for each individual item, the mean of the Likert ratings, the *SD* of the Likert ratings, and the number of observations. These reflect, respectively, estimated central tendencies of ratings, variability in these central tendencies, and sample size from which the summary statistics are calculated. These simple metrics are intuitive and easy to calculate, and can be used to rank items on the rated dimension. However, the use of means and *SD*s to accurately estimate distances between normed items would require that Likert scales are continuous, with an equal step size between each successive option. In fact, while the dimension participants are judging may scale continuously when measured objectively (e.g., age of acquisition), and while a Likert scale may be presented to participants with equal steps between options (e.g., via radio button inputs), there is no reason to assume that judgements on the target dimension are graduated linearly. Instead, Likert scales are examples of ordinal

scales, with responses scaling in one direction (i.e., 1<2<3<4<5. . .), but not necessarily in equal steps. At the very least, the true relationship between ratings and the dimension(s) they are supposed to measure remains underspecified. By norming items on an ordinal variable via their means and *SD*s, researchers produce estimates which can be distorted by nonlinearities in the scaling of Likert judgements (Liddell & Kruschke, 2018). If researchers were only interested in ranking items, summaries like the mean would be sufficient. However, it is often useful to accurately know the relative distances between items in the target dimension. For instance, item norms are frequently included in statistical analyses as continuous variables or predictors (e.g., Fernandino et al., 2016; Goh et al., 2016; Hollis & Westbury, 2016; Khanna & Cortese, 2021; Perry et al., 2018; Pexman et al., 2019; Scott et al., 2019; Vejdemo & Hörberg, 2016). Instances where researchers dichotomise a rated feature to compare the *N* highest- and lowest-rated items may be less impacted by distortions in averages of Likert ratings, as the comparison is still essentially ordinal. However, such dichotomisation will result in an unnecessary loss of statistical power and precision if continuous alternatives are available (MacCallum et al., 2002; Royston et al., 2006).

An alternative approach to summarising Likert judgements is to assume that a latent continuous distribution underlies the ordinal scale, allowing any given ordinal response to be converted into possible latent values (Figure 3.1). This is the approach implemented in cumulative link models (CLMs), where ordinal dependent variables are mapped onto ordered regions of a latent distribution (Bürkner & Vuorre, 2019; McCullagh, 1980). Responses are commonly modelled via probit- or logit-link functions which, respectively, assume that the latent variable is normally or logistically distributed. The model estimates the locations of ordered thresholds demarcating the borders between regions of the latent distribution associated with each response, while other coefficients can estimate a constant shift in the location of the distribution associated with changes in the values of predictors (i.e., slopes). The CLM approach can be extended to account for multilevel data in the form of CLMMs, which allow the researcher to estimate not only the values of population-level intercepts and slopes (i.e., fixed effects), but also how these intercepts and slopes differ across members of distinct populations which are sampled in the data (i.e., random, or "varying", effects). For instance, a CLMM can estimate how the mean latent value associated with each individual participant or item differs from that of the population average.

The need for ordinal models such as CLMs and CLMMs to appropriately model ordinal responses is already commonly recognised in the analysis of typical experiments (Liddell & Kruschke, 2018). Correspondingly, several tools currently exist, and are already widely used, to fit CLMMs, such as the *ordinal* (Christensen, 2020) package for the R programming language (R Core Team, 2021). When these models are applied, however, they are typically used to estimate the effects of experimental manipulations (i.e., fixed effects); when CLMMs are applied, random effects are typically included to account for participant and item variability, thereby improving accuracy of fixed effect estimation, but are rarely examined in any detail beyond a cursory glance at summary statistics like random effects variances. Estimating a CLMM with by-item random effects could, however, also be used to norm items in a manner which is not distorted by participants' response patterns. Indeed, random effects in such models are per-unit (per-item,

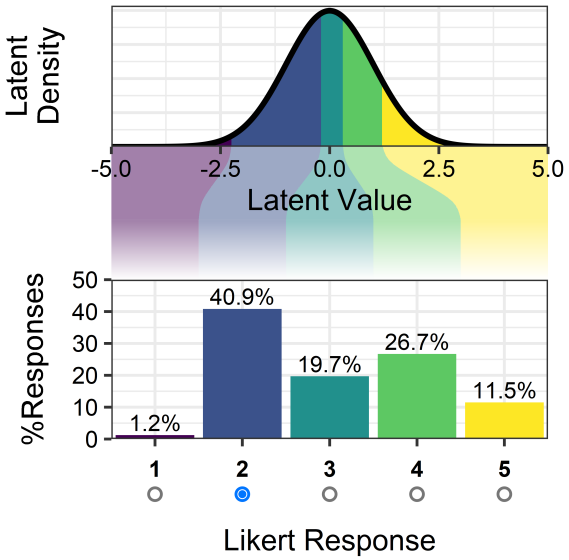


Figure 3.1: The assumed relationship between a continuous latent distribution and ordinal Likert responses (here, on a 1-5 scale). Each Likert response corresponds to a region of the latent distribution. The probability of observing any Likert response is the probability of a value being drawn from the latent distribution which is between the lower and upper bounds of that Likert response's region. In the example illustrated, the latent distribution is assumed to be normal (as is the case for a probit-link function). The nonlinearities in this example response pattern result in the most likely response being *2*, while the responses *1* and *5* would be comparably rare.

per-participant, etc.) estimates of each unit's most likely deviation from the corresponding fixed effect, in link units. CLMMs and related ordinal models assume the overall mean of the latent distribution (i.e., what would be the fixed-effect intercept in a linear model) to be equal to zero, for identifiability. In the case of a CLMM with per-item random effects, therefore, the extracted random effects will represent estimates of the latent mean associated with each item. In R, these values are stored within a fitted CLMM object, and can be extracted, for example, via the generic R function *ranef()*. Norming items via random effects in this way confers additional benefits, such as improvements in accuracy associated with shrinkage (where outlying, unlikely values, are appropriately pulled towards more likely estimates) and the concurrent estimation of additional sources of variability (such as per-participant random effects). In this article, I argue that CLMMs are well-suited to calculating norms from Likert responses, and solve key issues associated with more traditional analyses of norming studies.

One issue with traditional analyses of norming studies centres around the finding that heterogenous relationships are frequently observed between means and SDs. Notably, Pollock (2018) highlighted a common relationship in ratings of word concreteness (Figure 3.2), whereby the lowest SDs are observed at the extremes of a Likert scale, while items towards the centre of the scale show much higher SDs. Such heterogeneity should be expected to some degree for any scale which has lower and upper bounds. However, Pollock showed that although it was common for participants to agree on ratings at the extremes of the scale (*1* and *5*), such inter-rater reliability was exceedingly rare for ratings at midpoints in the scale. Pollock interpreted this finding as evidence that participants' judgements on dimensions showing this

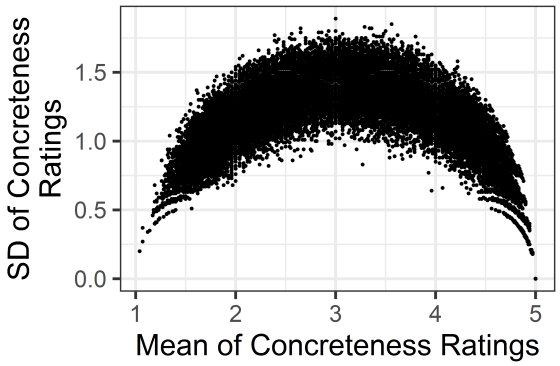


Figure 3.2: The relationship between the mean and *SD* of items' Likert ratings (1-5 scale) in word concreteness, from Brysbaert et al. (2014). The pattern suggests that responses are most consistent at the extremes of the Likert scale, but that items with averages at the midpoints of the Likert scale elicit less consistent responses.

overall pattern are largely dichotomous. Such a view is directly relevant to theories concerned with how concreteness is represented, standing in stark contrast to perspectives that suggest concepts like concreteness exist on a continuum (e.g., Gentner & Asmuth, 2019). It was argued that Likert scales are inappropriate for norming items on variables with dichotomous responses, and that averages at the centre of the scale merely reflect polarisation in responses, rather than a meaningful estimate. For instance, if half of all responses for a single item were *1*, and half were *5*, this would result in an average Likert response of *3*, even though no participant gave this response. This inconsistency would also be reflected by a high *SD* of >2.

Pollock's argument has been criticised by Neath and Surprenant (2020), who examined whether a concreteness effect in a serial word recall task differs between words with low or high *SD*s in Likert judgements. If mid-scale responses are less meaningful, they may be expected to predict effects of concreteness less well. Neath and Suprenant showed, however, that the effect of word concreteness was estimated as a consistent effect size when average Likert responses are used as the predictor, regardless of how large the *SD*s of Likert ratings are for the presented items. Further to this, I argue that Pollock's interpretation of the mean-*SD* relationship suggests that Likert responses are expected to be continuous, rather than ordinal. When Likert responses are instead viewed as ordered regions of a latent continuous variable, a unimodal latent distribution can lead to an apparent dichotomy in Likert responses, and responses can appear inconsistent even when there are meaningful differences in the latent distribution. Such a pattern could arise from any response pattern where the lowest and highest Likert responses (e.g., *1* and *5* on a five-point scale) account for large portions of the latent distribution, increasing the likelihood of any given latent value being mapped onto an extreme Likert response, while responses at the scale's centre (e.g., *2*, *3*, and *4*), account for much less, making these mid-scale responses comparably less likely. Importantly, even though responses could appear dichotomous in such cases, changes in the relative likelihood of the different Likert responses would still track meaningful shifts in the central tendency of the latent distribution. Furthermore, lower *SD*s at the extremes of a scale may reflect floor and ceiling effects, rather than agreement among raters. There may be meaningful differences between items that share the minimum

or maximum possible average rating, which are nonetheless undetectable within the limited bounds of the rating scale.

When dichotomous response patterns are explained with reference to ordered regions of a latent distribution, it is clear that many other response patterns should also be possible, and that these would result in distinct patterns in the mean-*SD* relationship of Likert responses (Brainerd et al., 2021). In any pattern, items whose average is closer to regions that participants are biased towards should be more likely to show greater consistency in responses, and thus have lower *SD*s, while items further from these regions will be more likely to have higher *SD*s. Figure 3.3 shows the mean-*SD* relationships observed in the Likert judgements of words on three different semantic variables from the Glasgow Norms (Scott et al., 2019): Dominance, Familiarity, and Gender. Each of these variables shows a qualitatively different mean-*SD* relationship distinct from that identified by Pollock (2018). For Dominance, the lowest and highest *SD*s, respectively reflecting the greatest and least consistency, are at the centre of the Likert scale, and no items are observed at or close to the scale endpoints. This suggests that, for this sample, judgements of words' dominance are biased towards a mid-point response or are dichotomous, and that there was never any consensus among raters for items having extreme Dominance values. For Familiarity, in contrast, responses are most consistent at the upper end of the Likert scale, with lower SDs observed as average Familiarity increases. Further, the average Likert response never reaches lower than 1.5, suggesting that for this sample of items participants rarely consistently agree that a word is unfamiliar. In the case of Gender, three separate regions of the Likert scale show the lowest SDs, with intervening responses never showing such consistency. These three regions may suggest that participants were biased towards three different responses: *1* for highly male, *4* for gender neutral, and *7* for highly female. It is important to note, however, that the highest *SD*s are also observed at the gender-neutral centre of the scale, suggesting that the average Likert response for some words may index polarisation in responses, with dichotomous ratings as either highly male or highly female. An example of such a word is *bridegroom*, a compound word which technically refers to a man, yet consists of two highly, yet oppositely, gendered words, *bride* and *groom*. The inconsistency observed for words like *bridegroom* stands in contrast to the consistent gender neutrality observed for words whose gender ratings also average to 4, but which result in low *SD*s (e.g., the words *impaired*, *name*, *occurrence*). This highlights that relative differences in the variance of Likert judgements can reflect meaningful differences, such as an items' ambiguity or discriminability. If overall variance is calculated on raw Likert responses, however, these meaningful differences in variance will be entangled with differences in response consistency that result from the overall response pattern.

If variance reflects meaningful differences among items, how can it be estimated without distortion from response patterns? One solution could be for researchers to estimate both a latent mean and latent *SD* for each item that is normed. Although CLMMs traditionally assume homogeneity of variance, the framework provided by CLMMs may be extended to simultaneously describe meaningful differences in both the central tendency and spread of a latent distribution. Just as latent means are analogous to raw means, estimates of latent *SD*s are analogous to raw *SD*s, similarly disentangled from response patterns. While most

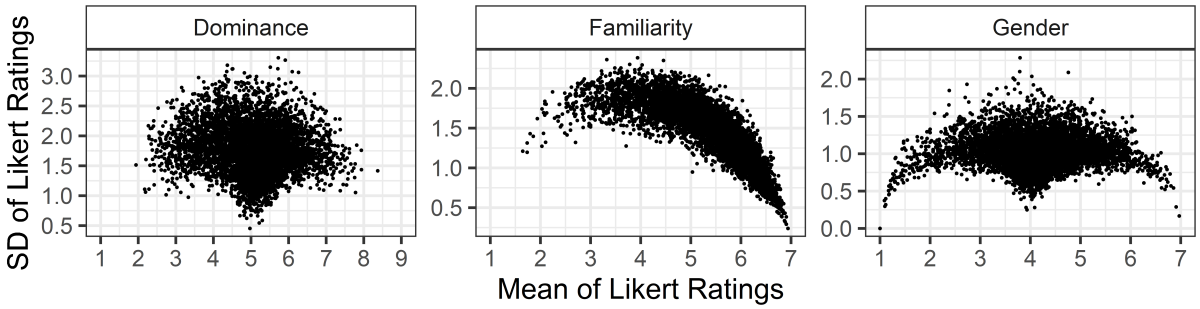


Figure 3.3: Mean-*SD* relationships for judgements of words on three semantic variables in the Glasgow Norms (Scott et al., 2019). Dominance was judged on a 1-9 Likert scale, while Familiarity and Gender were judged on 1-7 scales.

CLMMs, including those fit by the *ordinal* package (Christensen, 2020), exclusively model changes in the central tendency of a latent distribution (assuming homogenous variance across observations), it is possible to fit a model which concurrently describes changes in both the variance and central tendency of the latent distribution. The *brms* package (Bürkner, 2018) for R, an interface to STAN (STAN Development Team, 2021), provides an accessible solution to fit such models. Here, in addition to multi-level changes in the mean of the latent distribution, a discrimination parameter can be estimated, as the inverse of the latent *SD* (Bürkner & Vuorre, 2019). As the models are estimated via Markov chain Monte Carlo (MCMC) sampling, translating the discrimination parameter of each posterior sample to the *SD*, before calculating summary statistics, will allow the calculation of random effects for the variance of the latent distribution. CLMMs can therefore provide researchers with analogues to the traditionally reported statistics of means and *SD*s, but with both estimates disentangled from participants' response patterns.

I argue that CLMMs provide a valuable framework for norming items via Likert scales, allowing the calculation of items' latent means and *SD*s, analogous to the traditional estimates of means and *SD*s of responses, but disentangled from overall response patterns. In the first half of this article, through a series of simulations, I demonstrate the following: (1) non-linear response patterns can account for the typical patterns of relationships observed between means and *SD*s of ratings, and CLMMs can appropriately model items' values in the latent distribution underlying Likert responses; (2) such models can be expanded to account for other sources of variability, such as participant random effects, with improvements in the accuracy of item estimates; (3) such models can be further expanded to account for differences in a latent distribution's variance as well as its mean; and (4) while CLMMs make assumptions about the underlying latent distribution, they are relatively robust to modelling responses that result from distributions which violate these assumptions, and are still preferable to the traditional approach of calculating raw means of ordinal responses. In the second half of the article, I apply CLMMs to real norms data from existing datasets, on judgements of orthographic character similarities (Simpson et al., 2013) and on semantic dimensions of words (Scott et al., 2019), showing how these methods and results differ from those of traditional analyses.

3.2 Simulations

To demonstrate that CLMMs provide comparable results across different response patterns, I performed simulations as follows. On each iteration, a single dataset was simulated which had differences between observations, items, and (from Simulation 2 onwards) participants, described in terms of the mean and *SD* of a normally distributed latent distribution. This normal distribution represents latent values before they are distorted by an overall pattern in the Likert responses. Differences between items and participants are similarly drawn from normal distributions - I model these differences via the random-effect structure of the CLMM. The values from this single dataset are then mapped onto one of five possible response patterns. This is done to show (a) how identical effects in latent space can result in divergent estimates and patterns when using traditional means and *SD*s, and (b) how CLMMs provide estimates which are far less biased by overall response patterns.

Throughout the simulations, I use five example response patterns, as follows: equidistant, left-biased, right-biased, edge-biased, and centre-biased. These are similar to the qualitative categories of response styles identified in the item response theory literature (Baumgartner & Steenkamp, 2001). The only difference between the response patterns I simulate is in the locations of the thresholds demarcating the borders between regions of the latent distribution which map onto respective ordinal observations. The differences between the five response patterns are illustrated in Figure 3.4, which shows how the probabilities of ordinal Likert responses differ among the response patterns, even when the change in the latent distribution is identical; differences in the probability of each response are accounted for entirely by changes in locations of thresholds.

In each simulation, I mapped simulated latent values onto a corresponding Likert response according to each response mapping. For example, a latent value of 2.5 on one trial would be recoded to responses *4*, *3*, *5*, *5*, and *4* for the equidistant, left-biased, right-biased, edge-biased, and centre-biased response patterns, respectively. In this way, the results across the different response patterns are directly comparable. The only exception to this is Simulation 4, where I manipulated the distribution of latent variables and random effects but kept the response mapping constant. In every simulation, I recovered the item random effect values after fitting a separate CLMM to ratings simulated for each response pattern, using a probit-link function to reflect the normal distribution of the simulated latent distribution. Here, the retrieved item random effects encode the difference between each item and the overall distribution in a parameter describing the latent distribution (usually the latent mean, but in Simulation 3, also the latent *SD*). For instance, a random effect of -1.2 for a single item's latent mean would indicate a shift of the full latent distribution of -1.2 away from the grand mean (which, for CLMs, is always 0). Each CLMM was fit with either the *ordinal* (Christensen, 2020) or *brms* (Bürkner, 2018) package for R.

The code used in all simulations is available at the OSF project associated with this work, at https://osf.io/ntvmf/.

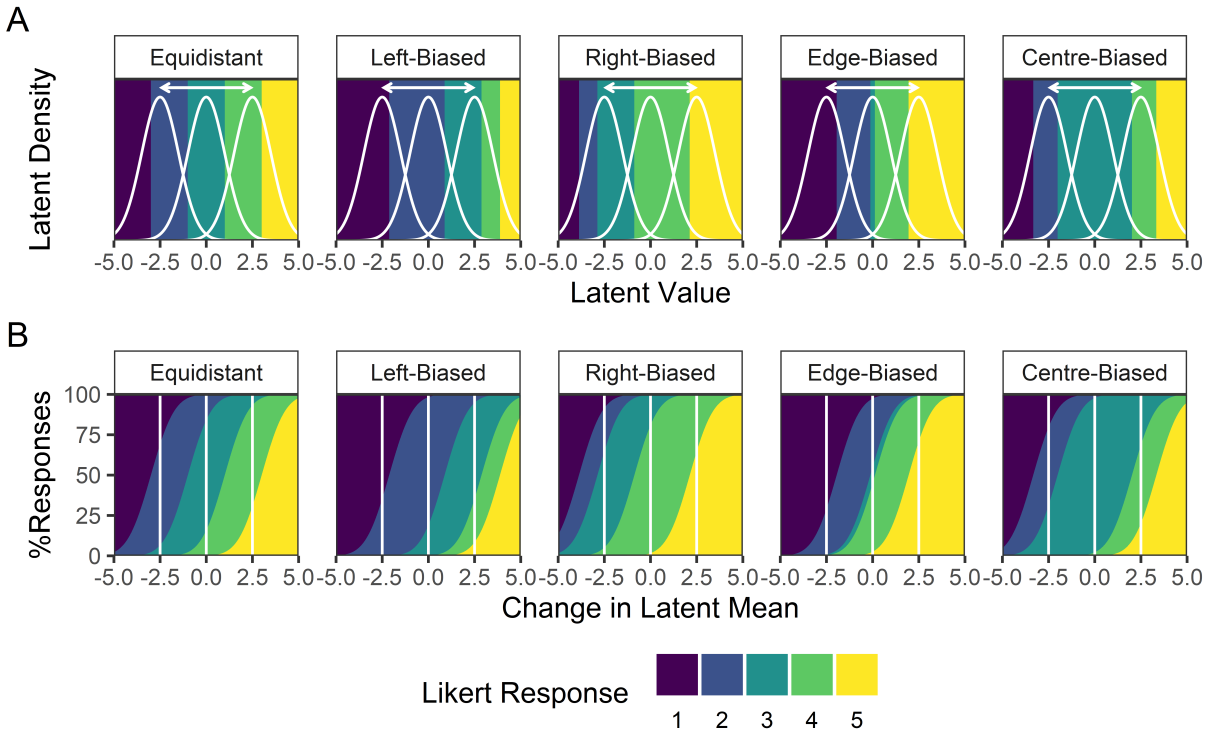


Figure 3.4: Illustration of how response patterns affect Likert responses. Given the same latent distribution, the five example response patterns I used in the simulations alter the probability of different Likert responses. The top half of the figure, (A), shows the locations of the thresholds for each response pattern, highlighting how changes in the latent mean alter the proportion of the latent distribution which maps onto each Likert response. Importantly, this effect differs among response patterns. To illustrate this point, for each response pattern, the densities of three example distributions (white curves) are shown, with means of -2.5, 0, and 2.5, and an identical *SD* of 1. An observation sampled from one of these distributions would fall into one coloured region and would be mapped onto the corresponding Likert response. The bottom half of the figure, (B) shows how the change in the mean of the latent distribution (on the x-axis) alters the probability (cumulative percentage; y-axis) of observing any Likert response differentially in each of the five response patterns for an identical latent distribution of mean 0 and *SD* 1. The same example three distributions as in panel A are highlighted with white vertical lines.

3.2.1 Simulation 1: CLMMs with Item Random Effects

In this first simulation, I demonstrate that cumulative links appropriately account for nonlinear response patterns, and that random effects can be used to accurately calculate differences in a Likert scale's underlying latent distribution for separate items. In each iteration, I simulated 100 individual items' positions on a latent distribution with a mean of 0 and *SD* of 1. I also simulated residual variance in the latent distribution with a normal distribution of mean 0, and *SD* 1. As such, latent distribution values for item i, L_i, were simulated as follows, where μ_i refers to item random effects, and e_i refers to the residuals.

$$L_i = \mu_i + e_i$$

$$\mu_i \sim N(0,1)$$

$$e_i \sim N(0,1)$$

In each iteration, I generated 25 latent means for each item, given that item's random effect μ_i, with these values then recoded to ordinal responses on a 5-item Likert scale, as described above. To recover (via *ranef()*) the item random effect values, I used the *ordinal* package (Christensen, 2020) to fit a CLMM to ratings simulated for each response pattern, with a probit-link function. In the package's syntax, the model was specified as follows:

```
rating ~ 1 + (1 | item_id)
```

Figure 3.5 depicts the results of Simulation 1. This simulation demonstrates that distinct patterns in the relationship between ratings' means and *SD*s can arise from the response patterns alone, even when the underlying latent distribution is identical. The results further show that while the means of ratings are heavily influenced by nonlinearities in response patterns, estimates of item random effects from the CLMMs are more robust to differences between response patterns. Notably, however, the distortions that result from using the raw mean may be less problematic if researchers are only interested in rank order (see Appendix A.1). Nevertheless, whenever researchers are interested in the relative distances between items, CLMMs provide estimates which are far less distorted by overall response patterns.

3.2.2 Simulation 2: CLMMs with Item and Participant Random Effects

In the typical design of a rating norming study, the 25 observations simulated in the previous simulation would have come from different participants, who are likely to systematically differ in how they judge any given word. Each participant would additionally rate a subset of the total set of items in the study, introducing a crossed random effects structure. Variability between participants can be calculated in the same model as item variability with the incorporation of an additional random effect. As a demonstration of this, I re-ran the previous simulation

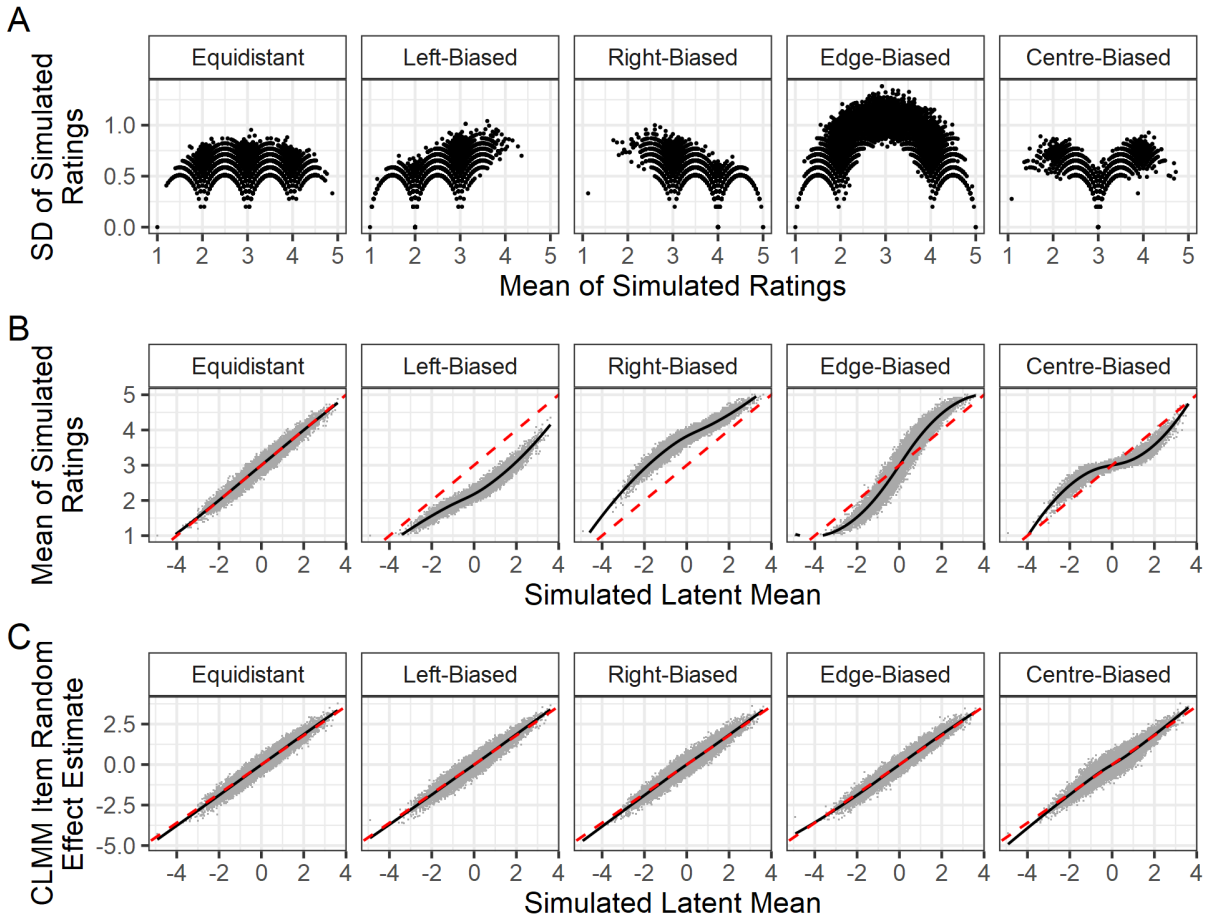


Figure 3.5: Results of Simulation 1: CLMMs recover items' latent distribution random effects from the five example response patterns. The panels show (A) the relationship between means and *SD*s of ratings, (B) the relationship between items' simulated latent means and their mean ratings, and (C) the relationship between items' simulated latent means and estimated random effects from the CLMM. The relationships in panels B and C shown with the black lines were estimated via locally estimated scatterplot smoothing (LOESS), with a span parameter of .75. The dashed red lines show an expected linear relationship for reference, identical across all response patterns. In all panels, results from all simulation iterations are concatenated. The results show that the relationship between items' means and *SD*s of ratings differs markedly between simulated response patterns, even though the simulated values in the latent distribution were identical. While averaging over ordinal responses works well when the responses are generated from equidistant thresholds, any other response pattern leads to nonlinear inaccuracies in the values. CLMMs, meanwhile, account for any pattern of thresholds and more accurately recover the items' distributions in the latent variable.

with the additional inclusion of participant random effects. As in the previous simulation, 25 observations were generated for each item, but each observation for an item was generated by 25 different simulated participants, where each participant rated 25 items in total. Participants were allocated to items pseudo-randomly, such that they rated each item only once. This meant there were a total of 100 participants in each iteration. The latent distribution values were thus simulated from a normal distribution with mean 0 and *SD* 1, with both item and participant random effects also drawn from normal distributions with mean 0 and *SD* 1. As a result, latent distribution value L_{ij} for the ith item and jth participant, was simulated as:

$$L_{ij} = \mu_i + \mu_j + e_i$$

$$\mu_i \sim N(0,1)$$

$$\mu_j \sim N(0,1)$$

$$e_{ij} \sim N(0,1)$$

As before, ratings were simulated by recoding regions of the latent distribution using the five response patterns shown in Figure 3.4. The CLMMs were again fit using the ordinal package with a probit-link function, specified to either omit or include participant random effects from the formula, written in the package's syntax as, respectively:

```
rating ~ 1 + (1 | item_id)
rating ~ 1 + (1 | item_id) + (1 | participant_id)
```

Figure 3.6 shows the Simulation 2 results, demonstrating that the estimation of participant random effects allows the CLMMs to recover the simulated item random effects more accurately.

In examining whether the estimation of participant random effects improves the accuracy of item random effect estimates, I first calculated the difference between each item's simulated item random effect, and that estimated by the two models. As panel D of Figure 3.6 shows, including both participant and item random effects in the fitted model improved the accuracy of the estimates compared to considering only item random effects. However, panel C shows that the magnitude of item random effects was underestimated when participant random effects were not calculated, which I considered could be the cause of the difference in accuracy of estimates between the two models. If the improvement in accuracy is due to a difference in magnitude alone, the improvement may not be meaningful or useful for rating norming studies. This is because it is the ordinal relationships and relative sizes of differences between items which are most informative, while underlying latent values are functionally arbitrary if relative distances are preserved. To examine whether the improvement in accuracy was solely the result of this difference in the magnitude of estimated item random effects, I calculated the error in item random effect estimates when model estimates are normalised by their respective *SD*s, thereby standardising the magnitude of the random effect estimates from each model. These

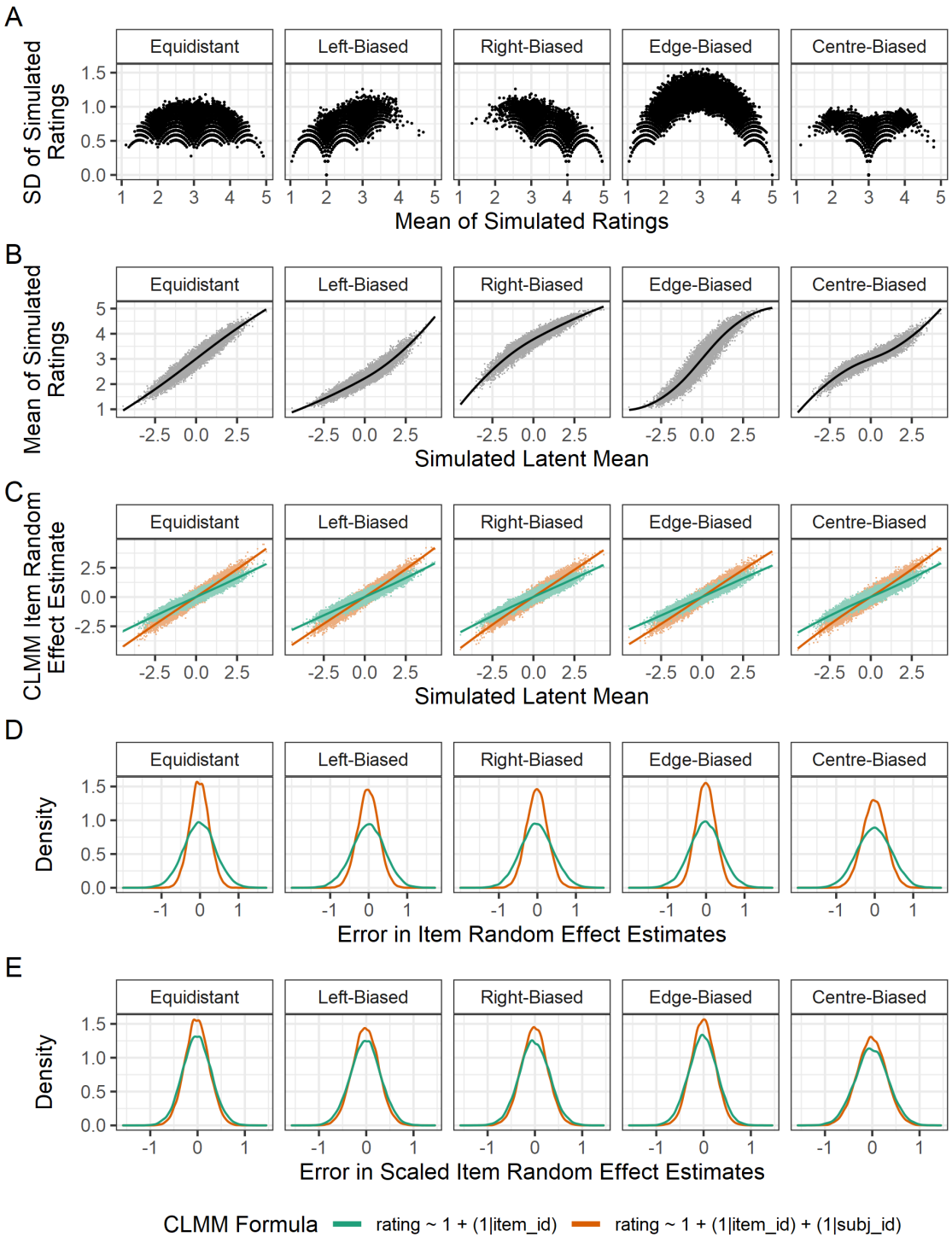


Figure 3.6: Results of Simulation 2: CLMMs recover items' latent distribution random effects when per-participant random intercepts are also simulated. Panels A and B show the same information as the respective panels in Figure 3.5. Panel C shows the relationship between simulated latent distribution values and item random effect estimates, from the CLMM estimating item random intercepts only (green), and from the CLMM estimating both item and participant random intercepts (orange). As in Figure 3.5, the lines represent LOESS estimates (span parameter of .75). Panel D shows densities of the error in the items' random effect estimates (i.e., error=simulated value-estimated value) from both types of CLMM. Panel E shows the same as panel D, but with random effect estimates scaled by standard deviation to account for the differences in estimated magnitude shown in panel C.

results are presented in panel E of Figure 6, showing that while most of the improvement in accuracy with participant random effects can be attributed to differences in the magnitude of effects, there may be some gain in accuracy when participant random effects are additionally accounted for.

The degree to which item random effect estimates increase in accuracy when participant random effects estimates are included will depend on features of the data. One important consideration is the magnitude of variances of the random effect distributions relative to one another, and to the latent distribution. This is because greater variance in the participant random effects distribution will increase the degree to which estimates are distorted by the biases of individual participants. To demonstrate this, I ran additional iterations in Simulation 2b (see Figure 3.7), varying the *SD*s of the participant and item random effect distributions from which the random effects are simulated. For simplicity, and because there would be similar results for each response pattern, I simulated Likert responses from the edge-biased response pattern only. *SD*s of item and participant random effect distributions were varied between .25 and 5, in steps of .25. All other features of the data were simulated as specified above. I ran 50 iterations for each combination of item and participant random effects and calculated the *SD*s of the error in scaled item random effect estimates (Figure 3.7A). I could then calculate the difference between these estimates to estimate the effect of including participant random effects on the accuracy of item random effects (Figure 3.7B). This analysis revealed that estimating participant random effects in the CLMM random effect structure can reduce error in the estimates of item random effects. Specifically, the results suggest that when participants are more variable than items, estimating participant random effects increases the accuracy of item random effects estimates. When items are more variable than participants, the results suggest that while there is no gain in accuracy, there is also no loss of accuracy.

An alternative approach to accounting for participant variability when calculating item norms may be to first *z*-score responses within participants, before calculating per-item averages. I consider such an approach in Appendix A.2. To summarise this evaluation, such an approach is well-considered as a simple approach which will account for per-participant differences in central tendency, but in itself it fails to account for the ordinal nature of a Likert scale, and will accordingly result in a similar distortion of estimates to that observed for raw averages. I further argue that CLMMs provide additional advantages, such as estimating item and participant variability concurrently, rather than accounting for these sources in separate steps, as the *z*-scoring approach does.

In sum, this simulation shows that estimating both item and participant random effects can improve the accuracy of item random effect estimates from a CLMM applied to data with a design comparable to that of a typical rating norming study. Fitting CLMMs which estimate item as well as participant random effects is unlikely to *reduce* accuracy of estimates and will provide a gain in accuracy which is dependent on the relative variability of items and participants. As a result, I argue that modelling both sources of variability simultaneously is both useful and appropriate when the goal is to accurately norm items on the basis of Likert ratings.

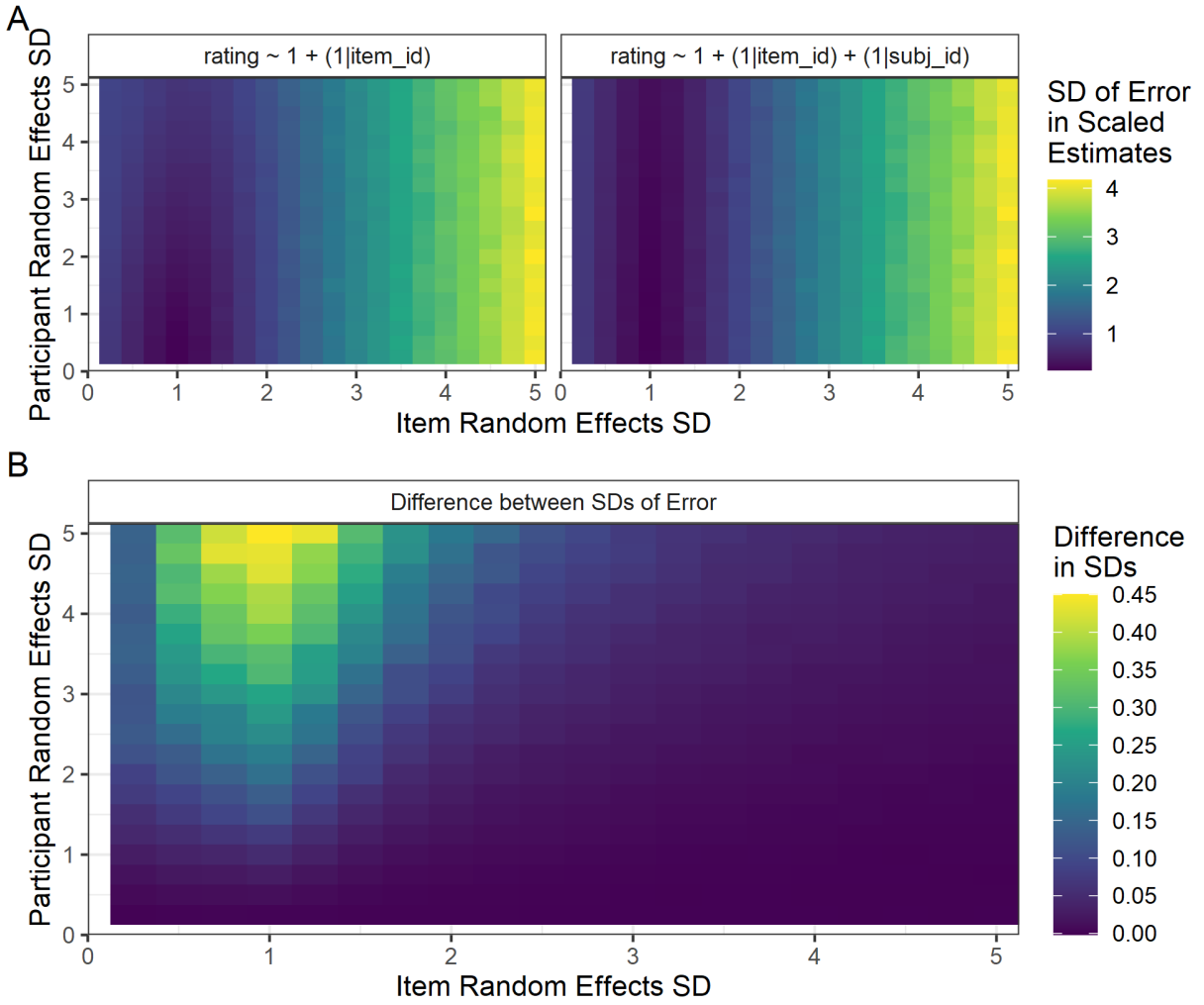


Figure 3.7: Results of Simulation 2b: effect of varying the magnitude of item (x-axis) and participant (y-axis) random effects on estimation accuracy of item random effects. Panel A shows the estimated SDs of errors of item random effects, where estimates are scaled (as in Figure 3.6E) to account for differences in magnitude. Estimates in panel A are shown separately for a model estimating only item random effects (left), and a model estimating both item and participant random effects (right). Panel B shows the difference between the estimates from the two models, calculated as item estimates from the less complex model (estimating only item random effects) minus those of the more complex model (estimating both item and participant random effects). Values in panel B therefore index the reduction in error that results from accounting for participant random effects (e.g., a value of .3 reflects a reduction of .3 *SD*s in the magnitude of errors).

3.2.3 Simulation 3: CLMMs Estimating Latent Variance

All CLMMs shown thus far estimate changes in the mean of a latent normal distribution, assuming homogeneity of variance. The latent distribution's spread may also differ meaningfully between items, however. As an example, consider polysemous words (words with multiple senses): for example, the word *lie* may be used in the sense of a bodily position, or in the sense of spreading falsehoods. As a result, one may expect ratings on semantic dimensions to show greater variance for such ambiguous words. However, if the words are presented with a disambiguating context (e.g., *lie (position)* and *lie (untruth)*), one may expect not only an associated shift in the average of Likert ratings (Scott et al., 2019), but also in the variance of ratings. Like means, however, *SD*s of Likert ratings incorrectly assume continuity in an ordinal scale, and this accordingly causes response patterns to distort estimates (see panel A of Figure 3.5 and Figure 3.6). As with the mean, an estimate of the *SD* of the latent distribution may therefore be used to disentangle such meaningful differences from the ordinal response pattern.

Although packages like *ordinal* generally assume homogeneity of variance, differences in multiple parameters of a distribution function can be modelled concurrently with the *brms* package (Bürkner, 2018) for R. When estimating changes in all parameters which specify a distribution, such an approach can be considered an example of distributional modelling. Extending this method to CLMs and CLMMs can allow researchers to estimate differences in a latent distribution's mean and variance concurrently. Here, the latent distribution's mean is estimated on an identity scale via one linear formula, while a second linear formula allows differences in the latent distribution's variance to be estimated as changes in a discrimination parameter (Bürkner, 2020). This parameter is specified as the inverse of the latent distribution's *SD* (i.e., 1/*SD*), and is modelled on a log scale by default (Bürkner & Vuorre, 2019). In the previous simulations, I have shown that random effect estimates of an item's mean in a latent distribution can be used as a measure of its central tendency in the rated dimension, akin to means of Likert ratings but disentangled from overall response biases. In a similar manner, I argue that random effect estimates of the latent distribution's variance can be used as a measure of an item's spread in the rated dimension, akin to the *SD* of Likert ratings, but again, disentangled from response patterns.

To demonstrate that a distributional CLMM can accurately estimate items' latent means and *SD*s across different response patterns, I simulated data with participant and item random effect estimates for both the latent distribution's mean and *SD*. The numbers of participants and items, and the numbers of observations per participant or item were identical to those used in Simulation 2. The latent distribution value associated with each trial, however, was simulated as follows:

$$L_{ij} \sim N(\mu_{ij}, \sigma_{ij})$$

$$\mu_{ij} = \mu_i + \mu_j$$

$$\mu_i \sim N(0, 1)$$

$$\mu_j \sim N(0,1)$$

$$\sigma_{ij} = \frac{1}{e^{disc_i + disc_j}}$$

$$disc_i \sim N(0,.5)$$

$$disc_j \sim N(0,.5)$$

Here, latent values (L_{ij}) are drawn from a normal distribution with mean μ_{ij} and *SD* σ_{ij}. Latent means (μ_{ij}) are calculated as the sum of item (μ_i) and participant (μ_j) random effects, which are both drawn from normal distributions of mean 0 and *SD* 1. Latent *SD*s ($\sigma_i j$) are calculated as the inverse of the exponent of the sum of item ($disc_i$) and participant ($disc_j$) random effects for a discrimination parameter, which are drawn from normal distributions of mean 0 and *SD* .5.

In total, 100 datasets were simulated, and, as in the previous simulations, the latent distribution was recoded to values on the Likert scale using the five different response patterns. For each of the five response patterns in each of the 100 iterations, a probit-link Bayesian distributional CLMM was fit with *brms*, with 3 Markov chains consisting of 6,000 iterations each (split equally between warmup and sampling). For all CLMMs, the *adapt_delta* parameter was set to .8, and the *max_treedepth* parameter was set to 10. In *brms* syntax, the model formula was specified as follows:

```
brmsformula(
  rating ~ 1 + (1 | item_id) + (1 | participant_id),
  disc ~ 0 + (1 | item_id) + (1 | participant_id)
)
```

The results of Simulation 3 are presented in Figure 3.8. As in the previous simulations, averages of simulated Likert responses were distorted by nonlinearities in response patterns (panel B), whereas CLMM random effects scaled linearly, across all response patterns, as a function of simulated differences in the latent variable (panel C). Similarly, *SD*s of simulated Likert responses less accurately represented the simulated latent variable variance (panel D) than *SD*s calculated from random effects for the *disc* parameter (panel E). Notably, as the simulated latent variable *SD*s increase, the degree to which this is underestimated by *SD*s of Likert ratings increases, to the extent that items with a simulated *SD* of 8 are only estimated as having an *SD* of between 1.5 and 2. This is a consequence of the Likert scale's finite bounds.

The results of Simulation 3 show that unequal variances in the latent distribution can be retrieved by a distributional CLMM.

However, a model assuming *equal* variances across observations can still accurately retrieve differences in the central tendency of the latent distribution, when variances differ systematically between participants and items. To demonstrate this, for each of the 100 datasets in this simulation, I fit an additional model assuming equal variance across

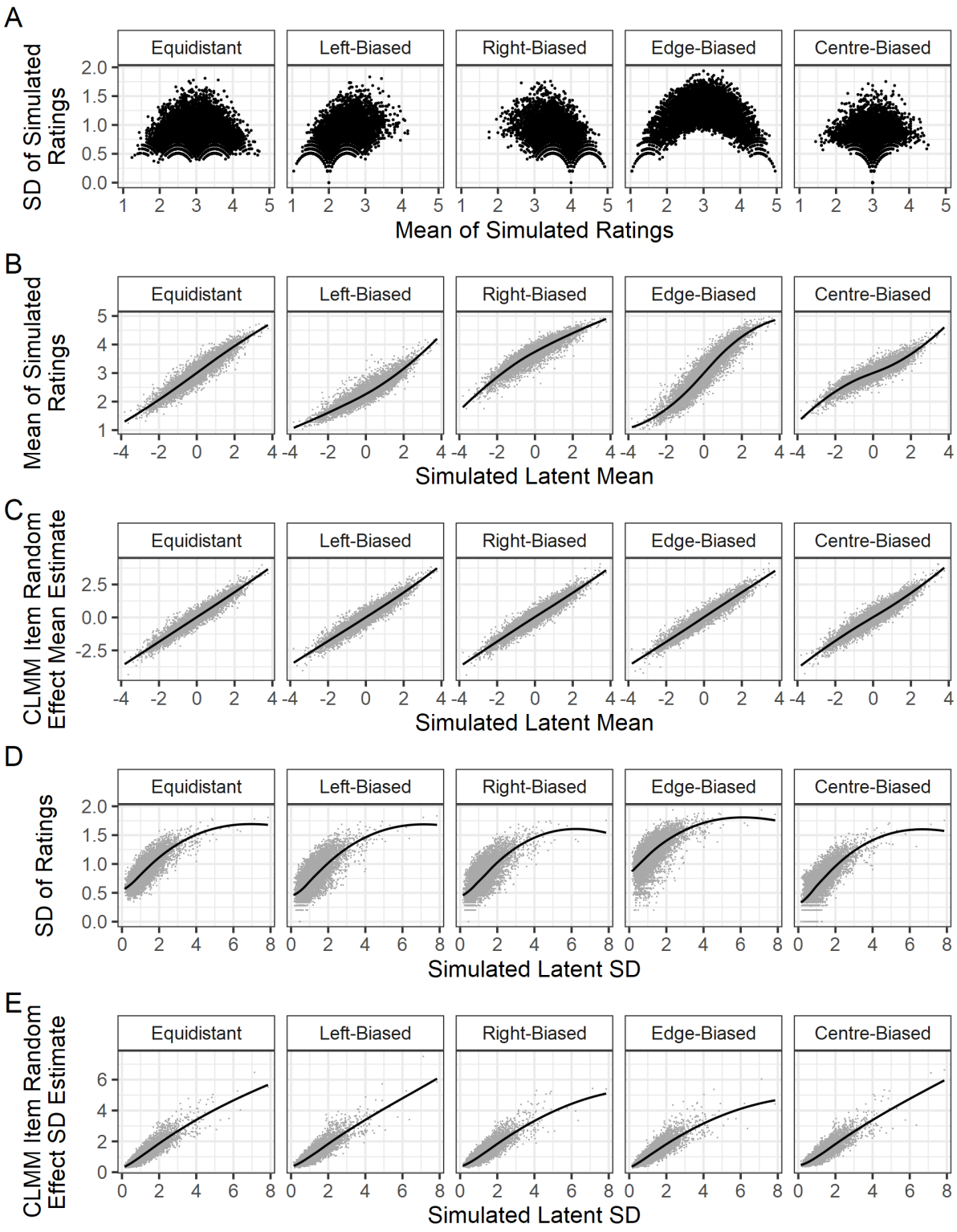


Figure 3.8: Results of Simulation 3: efficacy of distributional unequal variance CLMMs for calculating the *SD* of latent variables' variance from Likert response data. Panels A, B, and C show that the findings from the previous simulation also apply to models which calculate differences in both the mean and variance of the latent distribution. Panel D shows the relationship between items' simulated latent variable *SD*, and the *SD* of Likert ratings, while panel E shows that differences in the latent distribution's *SD* are more accurately retrieved by random effects in the *disc* parameter. Lines tracking relationships in panels B-E are estimated via LOESS (span parameter = .75).

observations. For comparability, this model was specified and fit in a manner identical to that of the distributional CLMM (i.e., using *brms* with identical sampling settings). The sole difference between the model specifications was that the formula for the equal-variance model only estimated differences in the latent distribution mean, assuming homogenous latent variance. In *brms* syntax, this was simply written as follows:

```
rating ~ 1 + (1 | item_id) + (1 | participant_id)
```

I could then compare the accuracy of estimation of each item's latent distribution mean, from each iteration of the simulation. The results of this analysis are summarised in Figure 3.9, showing that accuracy in the estimation of item random effects is very similar for both equal and unequal variance models. As a result, assuming equal variance, when variances across items and participants are in fact unequal, may not be overly problematic, provided the researcher is not interested in the differences in variance of the latent distribution. However, the extent to which accounting for unequal variance may improve accuracy of estimates may be related to the magnitude of differences in latent variance, relative to the magnitude of differences in latent mean. Consequently, researchers should carefully consider whether they expect meaningful differences in the variance of the latent distribution.

3.2.4 Simulation 4: Robustness of the Normal Assumption

The previous simulations all considered a normally distributed latent variable. This choice was motivated by normality of the latent distribution being a central assumption of CL(M)Ms fit with a probit-link function. Other link functions similarly assume other distributions; for instance, the logit-link function assumes the latent variable takes a logistic distribution. Relatedly, CLMMs assume that item and participant random effects are drawn from normal distributions centred on zero. One can imagine scenarios, however, in which a model's distributional assumptions, for either the latent variable or random effects, are inconsistent with reality. For example, item random effects may be bimodally distributed if there are two distinct categories in the data, such as has been argued to be the case for judgements of concreteness (Pollock, 2018). To demonstrate that the use of CLMMs for norming items is relatively robust to violations of the models' distributional assumptions, I ran two simulations fitting probit-link CLMMs via the *ordinal* R package (assuming equal variances) to data where, respectively, the latent variable (Simulation 4a) or the item random effects (Simulation 4b) are drawn from non-normal distributions. In all simulations, for simplicity of the results, I simulated only the edge-biased response pattern.

In these final two simulations, the data were generated equivalently to those described in Simulation 2, except that either the latent distribution, or the item random effects, were drawn from one of five possible distributions (Figure 3.10A). The first of the five distributions was a normal distribution with mean of 0 and an *SD* of 1 (i.e., identical to the $N(0, 1)$ used in Simulation

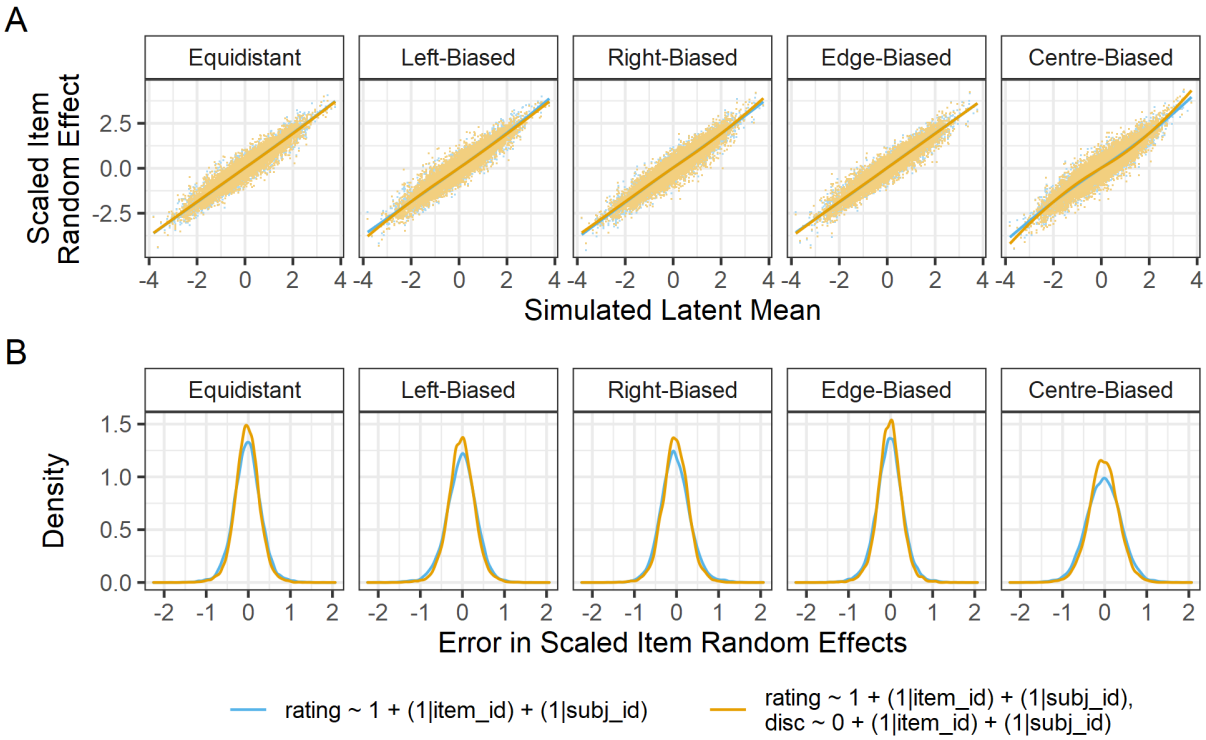


Figure 3.9: Comparison between a CLMM assuming equal variance across observations (blue), and a CLMM estimating differences in the variance of the latent distribution (yellow). Panel A shows that a similar pattern exists for both models between simulated latent means, and those estimated by the models' random effects (scaled for comparability between the models). The estimates are so similar that results for the equal variances model are largely overlaid by results from the distributional model. Panel B shows the density of the differences between simulated latent mean values and scaled estimates from CLMM random effects. While estimating items' differences in latent variance may provide a small gain in accuracy for the estimated means, reflected by the slightly heavier tails in the density plots, this improvement is minimal for data simulated here.

2). This was included such that results for all other distributions could be directly compared to data which conformed to the model's assumptions. The four non-normal distributions were as follows: a logistic distribution ($\mu = 0$, $s = 1$); a uniform distribution ($min = -2$, $max = 2$); a bimodal distribution composed of two normal distributions (respectively, $\mu = -1.5$, $\sigma = .75$, and $\mu = 1.5$, $\sigma = .75$), and a half-normal distribution ($\mu = 0$, $\sigma = 1$). Each of the two simulations was run for 250 iterations. In both simulations, the distributions of either the latent variable or item random effects was altered while all other parameters and results were held constant.

In the first of the two simulations (Simulation 4a), I varied the distribution that the latent variable is drawn from. Item and participant random effects distributions, meanwhile, were drawn from the normal distributions specified in Simulation 2. Figure 3.10 presents the results of this simulation: using the same (probit-) link function to model data, where the latent variable values are in fact drawn from the five distributions described above, may affect the accuracy of estimation of item random effects, but in all cases provides item norms which are more accurate, and scale more linearly, than calculation of mean Likert responses.

In the second of the two simulations (Simulation 4b), I varied the distribution that the simulated item random effects were drawn from. The latent distribution itself, and the distribution of participant random effects, were simulated as the normal distributions described in Simulation 2. Figure 3.11 presents the results of this simulation, showing that, even for very non-normal random effects distributions, the CLMMs still estimate the item random effects more accurately than would a traditional average of Likert ratings.

3.3 Application to Real Data

To demonstrate the viability of CLMMs for norming items, I applied the approach to two real datasets collected from norming studies: orthographic character similarity judgements collected by Simpson et al. (2013), which show a similar mean-*SD* relationship to that identified by Pollock (2018), and norms on semantic dimensions collected in the Glasgow Norms (Scott et al., 2019). I demonstrate that CLMMs exhibit patterns of results like those in the simulations reported in this chapter, and that, unlike traditional summary statistics of Likert responses, they can disentangle meaningful differences among items from participants' overall response patterns. I also show that CLMM-derived norms show greater reliability than traditional summaries across most of the normed dimension, but greater discriminatory power at the normed dimension's extremes. Together, these benefits confer improvements in predictive power of relationships with variables that correlate with the normed variables.

3.3.1 Simpson et al. Analysis

Simpson et al. (2013) collected character similarity judgements for 2,704 pairs of characters, from 677 participants, on a seven-point Likert scale ranging from not at all similar (*1*) to very similar (*7*). Each pair of characters comprised either two lower-case or two upper-case characters. The trial-level data, shared via personal correspondence, consisted of 81,199 total trials, with between 108 and 120 trials per participant, and between 29 and 31 ratings per item.

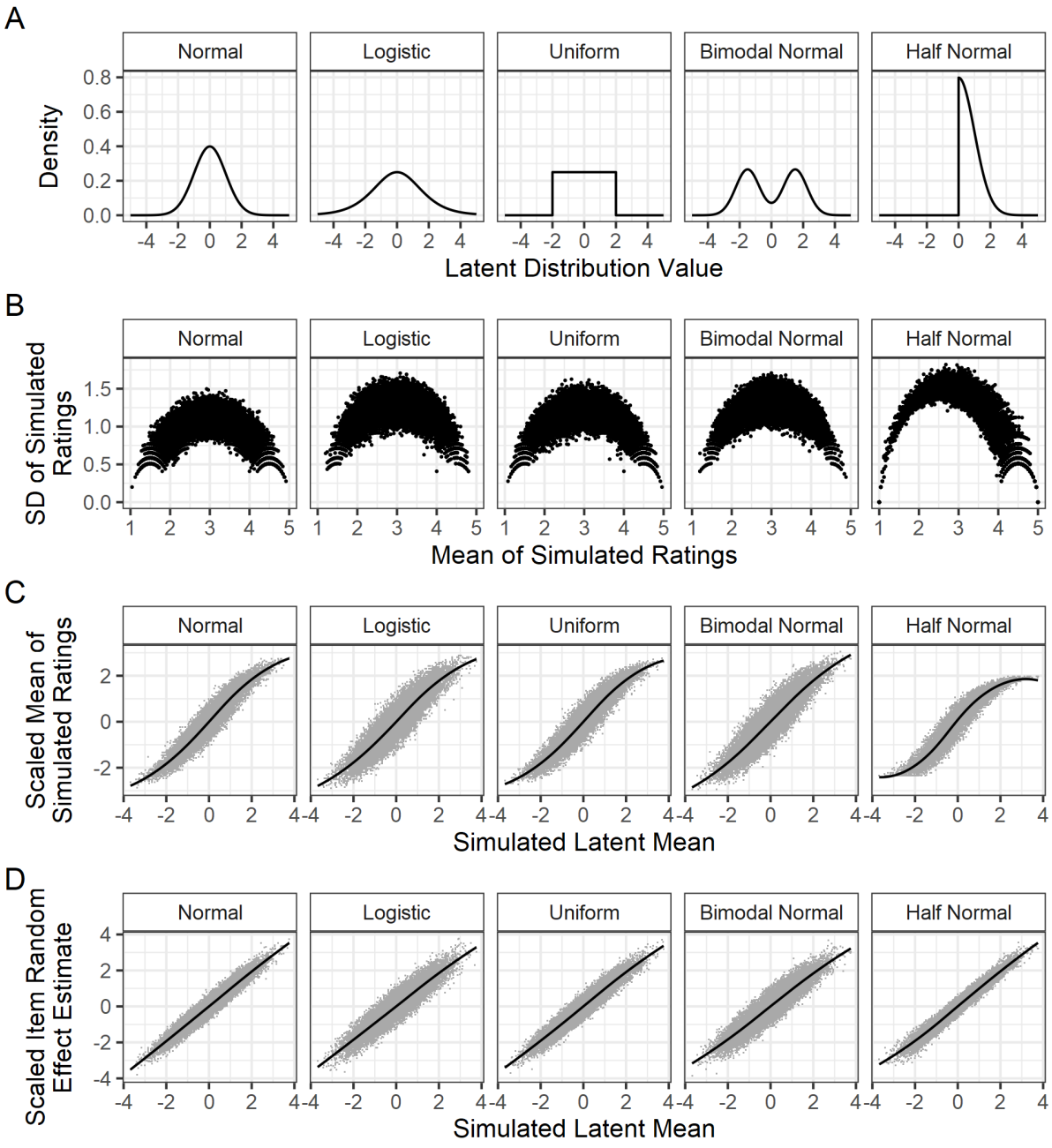


Figure 3.10: Results of Simulation 4a: varying the shape of the latent distribution. Across all non-normal distributions (panel A), a similar heterogenous pattern in the mean-*SD* relationship in Likert ratings was observed, though it is notably asymmetrical in the case of the half-normal distribution (panel B). For all distributions, estimates of items' simulated latent variable values were distorted by the edge-biased response pattern when estimated via the mean of simulated Likert ratings (panel C). In contrast, random effects estimates from the CLMM more accurately retrieved the simulated latent variable values, with similar accuracy across the distribution (panel D).

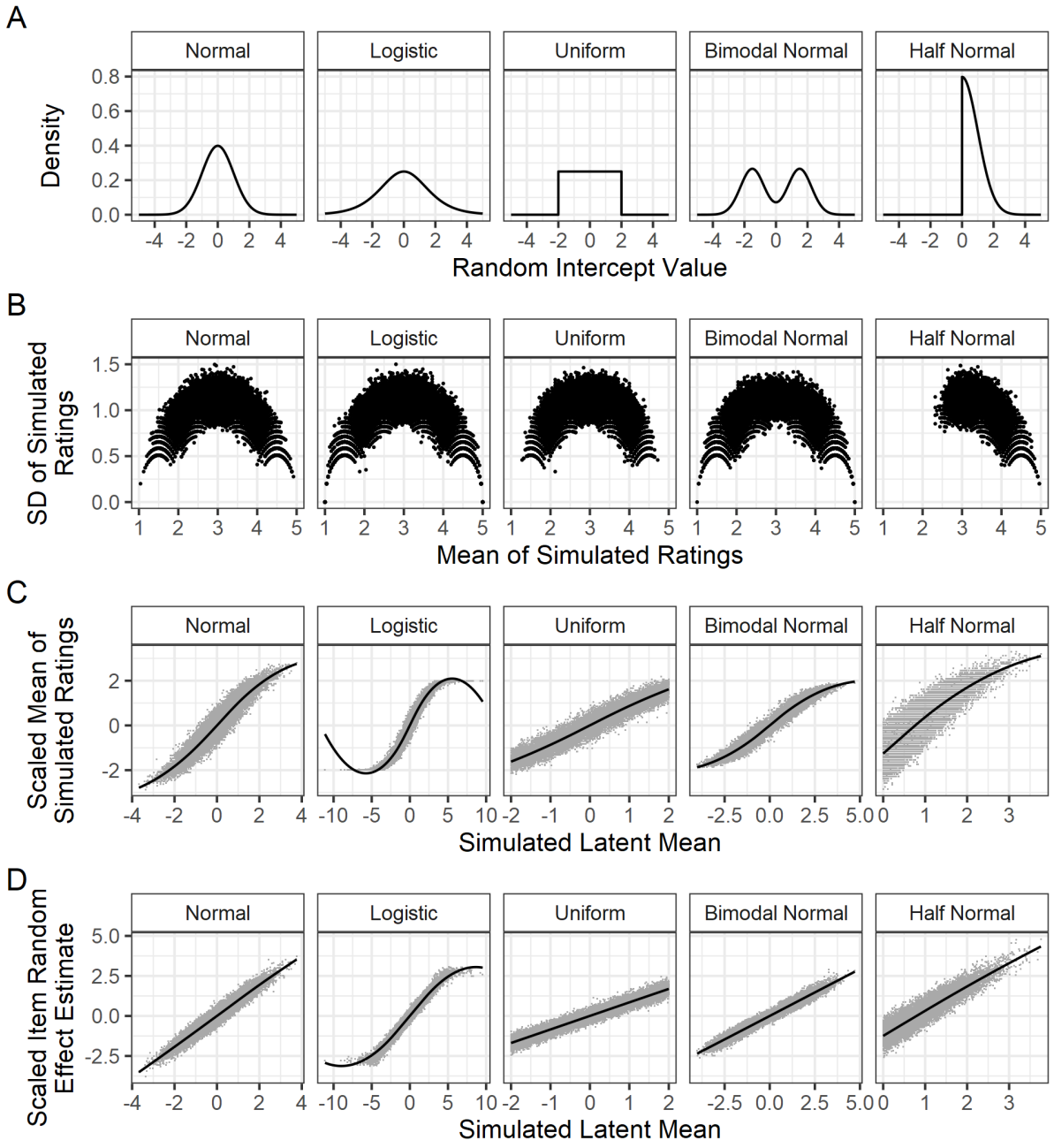


Figure 3.11: Results of Simulation 4b: varying the distribution of the item random effects. Across all non-normal distributions (panel A), the mean-*SD* relationship for Likert ratings (panel B) showed greatest inconsistency at the midpoints of the Likert scale, although this reflects any asymmetries in the random effect distribution, as is shown for the half-normal distribution. For all non-normal random effect distributions simulated, estimates from averages of Likert ratings (panel C) are distorted by the response pattern, and are less accurate than random effect estimates from CLMMs (panel D), which scale more linearly with simulated values. The unusual pattern in panel C for random effects drawn from a logistic distribution does not reflect the pattern of observations well, but is an artefact of the LOESS method of estimation.

Unlike the original analysis, I did not exclude responses based on a ±2 *SD* cutoff, but only excluded missing (i.e., blank) or meaningless (e.g., less than *1* or more than *7*) responses.

Model Results

I fit a Bayesian distributional CLMM to the trial-level data, estimating item and participant random effects. The model was fit with *brms*, using a cumulative probit-link function, and with 6 Markov chains of 6,000 iterations each (split equally into warmup and sampling). The *adapt_delta* parameter was set to .95, and the *max_treedepth* was set to 10. To reduce the size of the model for feasibility of storage, the thin argument was set to 2, meaning that only one half of the posterior samples (i.e., 1,500 per chain) were saved. In *brms* syntax, the model formula was as follows:

```
brmsformula(
  rating ~ 1 + (1 | item_id) + (1 | participant_id),
  disc ~ 0 + (1 | item_id) + (1 | participant_id)
)
```

Table 3.1 presents the estimates and credible intervals for parameters estimated by the model. A summary of the per-item results is presented in Figure 3.12, showing that the problems I identified with reporting means and *SD*s for ordinal Likert responses, namely the distortion of estimates by response patterns, did affect the data. Notably, the patterns of mean-*SD* relationships (Figure 3.12A), and the estimated locations of the thresholds (Figure 3.12E), are similar to those I observed in simulations of edge-biased responses (Figure 3.5). Similarly, given that latent means provide a measure less biased by response patterns, the relationship between means of Likert ratings and latent means (Figure 3.12C) suggests that the use of averaged ordinal responses has distorted the results of the norming study. Furthermore, this nonlinearity in the study's response pattern is accounted for by the CLMM, with the distinct inverted U pattern (Figure 3.12A) disappearing for means and *SD*s in the latent distribution (Figure 3.12B). However, some differences between items' *SD*s of Likert responses is preserved in latent *SD*s (e.g., between items *o-b* and *b-h*), suggesting that *SD*s of ratings are influenced by both response patterns and variability of responses for different items. This is reflected in the scatter plot showing the relationship between *SD*s of ratings and latent *SD*s (Figure 3.12D), which suggests only a noisy relationship, due to the influence of response patterns on *SD*s of Likert ratings. In summary, by estimating differences in the mean and *SD* of the latent distribution, I was able to estimate analogues to the traditional mean and *SD* of Likert responses, but which are disentangled from the raters' response biases. These estimates of latent means and *SD*s can be used to calculate a latent distribution for any presented item (Figure 3.12E). When combined with the threshold estimates, it is possible to probabilistically predict Likert responses for any item. For instance, the character pair *u-ù* would be expected to elicit a Likert response of *7* around 50% of the time.

Table 3.1: Summary of parameters for of character similarity ratings from Simpson et al. (2013): Estimates (medians of posterior distributions) and 89% highest density intervals (HDIs) for the key parameters estimated by the Bayesian distributional CLMM. The first six parameters reflect the estimated locations of the six thresholds in the latent distribution (e.g., *Threshold 3|4* reflects the latent location of the threshold between Likert responses for *3* and *4*). The last four parameters reflect the standard deviation of the random effects distributions for the cumulative link's μ, and $disc$ parameters, for items (i) and participants (j). The symbols and i and j subscripts are used for consistency with the simulations. These estimates revealed that magnitude of differences is larger in the μ parameter than in the $disc$ parameter, and interestingly, that the variability in μ is larger for items than for participants, while the variability in $disc$ is larger for participants than for items.

Parameter	Estimate	89% HDI
Threshold 1|2	.02	[-.04, .08]
Threshold 2|3	.84	[.77, .91]
Threshold 3|4	1.53	[1.46, 1.60]
Threshold 4|5	2.08	[2.00, 2.15]
Threshold 5|6	2.87	[2.79, 2.96]
Threshold 6|7	4.33	[4.22, 4.44]
SD of μ_i	1.47	[1.43, 1.51]
SD of μ_j	.71	[.67, .75]
SD of $disc_i$	.16	[.15, .17]
SD of $disc_j$	.27	[.26, .29]

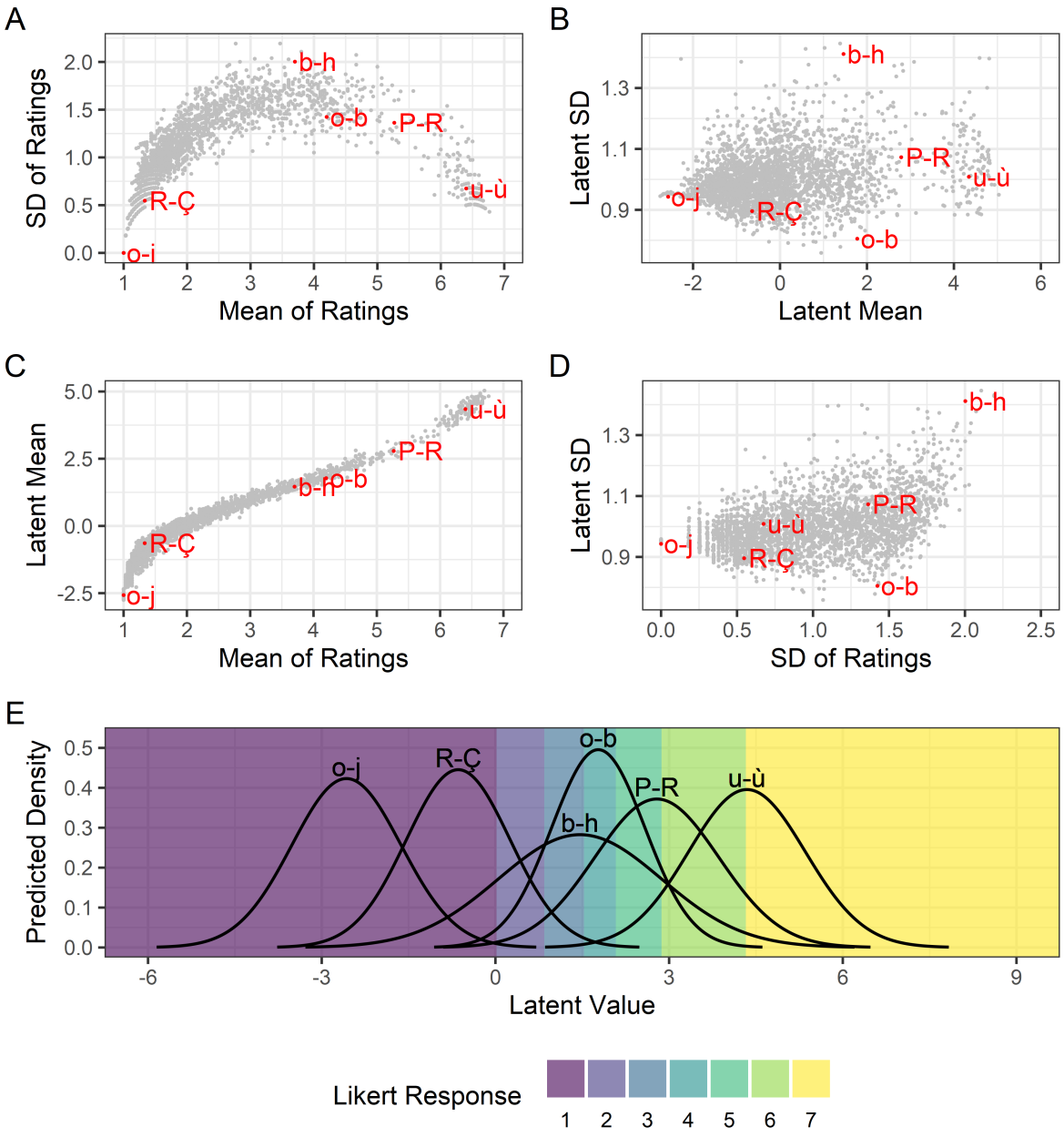


Figure 3.12: Summary of analyis of character similarity ratings from Simpson et al. (2013): Item random effect results of the Bayesian distributional CLMM. Six pairs of Arial characters are highlighted as examples, ranging from *o-j* at a low level of similarity, to *u-ù* at a high level of similarity. Panels depict: (A) The mean-*SD* relationship in items' ratings; (B) The mean-*SD* relationship in the latent distribution; (C) The relationship between each item's mean rating and its latent distribution mean as estimated by the CLMM; (D) The relationship between ratings' *SD*s and the estimated latent *SD*; and (E) The predicted densities of the latent distributions for the six example items, calculated from their random effect estimates for the μ and $disc$ parameters. Coloured regions indicate the mapping from latent values to Likert responses, where the boundaries between coloured regions reflect the estimated locations of the latent thresholds.

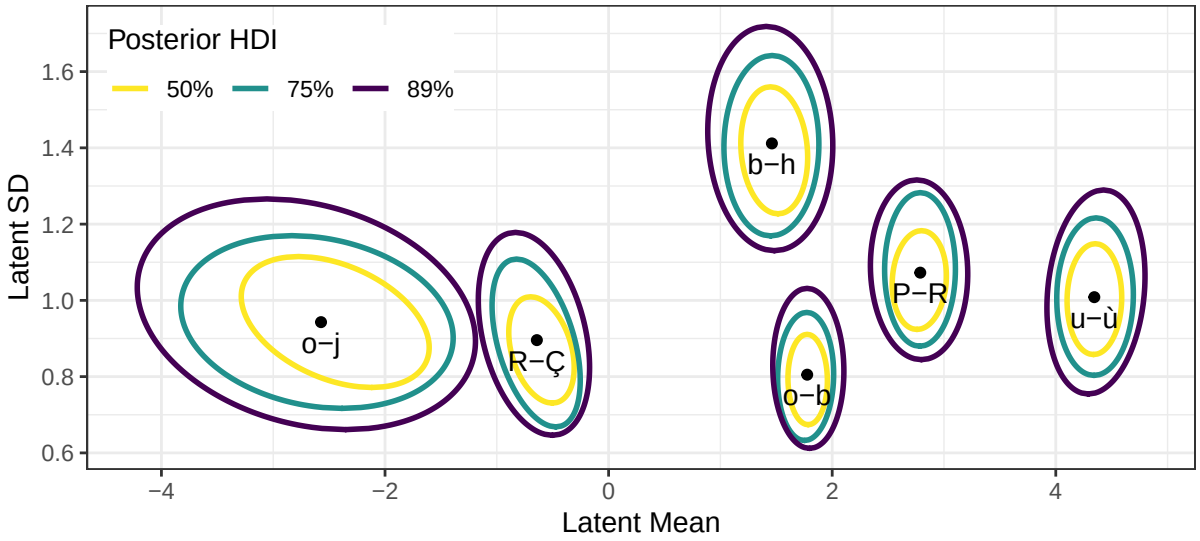


Figure 3.13: Uncertainty in the estimates of example pairs of characters' means and standard deviations in the latent distribution. Ellipsoids present the 50, 75, and 89% HDIs in the posterior samples, while points show the median estimates for each pair of characters.

The use of Bayesian estimation in the distributional model can also allow researchers to examine uncertainty in the random effects of each item. To demonstrate this, I calculated a series of two-dimensional highest density intervals for the latent distribution's mean and standard deviations, for each of the example items highlighted in Figure 3.12. These highest density intervals are presented in Figure 3.13 and demonstrate the degree of uncertainty in the posterior estimates of the Bayesian distributional model. For instance, Figure 3.13 shows that of the six example items, the latent mean and *SD* values for the *o-b* pair are most certain. In contrast, there is high uncertainty about the latent mean and *SD* associated with the *o-j* pair. This is in part likely to reflect that all responses were *1* for this pair, such that a floor effect makes it difficult to estimate the latent distribution (i.e., there are many normal distributions that could plausibly result in the observed number of participants consistently responding with the lowest value in the Likert scale).

The Bayesian distributional CLMM presented above modelled differences in both latent means and latent variances. Following the results of Simulation 3, I was interested in examining how estimates would change if the CLMM assumed equal variance across observations. To additionally show the similarity in the results between Bayesian MCMC and frequentist maximum likelihood models, I fit an equal-variance model using the *ordinal* package. The model was fit using the same link function as the Bayesian distributional CLMM (probit), with a random effects structure specified as follows:

```
rating ~ 1 + (1 | item_id) + (1 | participant_id)
```

I could then compare the per-item latent mean estimates of the two models. In this way,

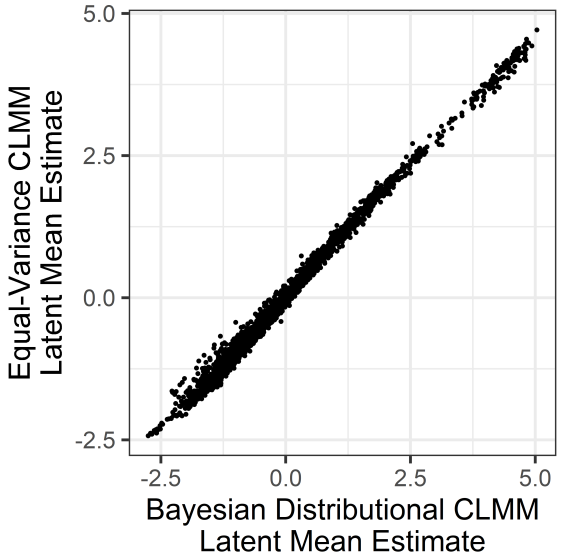


Figure 3.14: Correlation between random effect estimates of items' latent means from CLMMs assuming unequal and equal variance. The latent mean estimates from the Bayesian distributional CLMM, fit with the *brms* package, correlate very highly with estimates from the model assuming equal variance, fit with the ordinal package.

I could examine to what extent assuming equal variance (and fitting via maximum likelihood rather than Bayesian estimation) would affect estimates if the researcher were only interested in latent means. This revealed a Pearson's correlation of r=.997 between the models' estimates Figure 3.14. However, it is worth noting that there were considerable differences in the time it took to fit each of the two models - the Bayesian distributional model took several hours to fit on a typical modern computer, while the equal-variance model fit with maximum likelihood took only a couple of minutes. Nevertheless, the simpler, equal-variance model provided extremely similar estimates of per-item latent means. Were the researcher only interested in latent means, the simpler, equal-variance model would have been arguably sufficient for norming the similarity judgements, though that would mean forfeiting the rich posterior distributions afforded by the Bayesian approach.

Reliability and Discriminatory Power

An important consideration is the variability of norms derived via the CLMM approach between separate samples. To examine this, I used a method of cross-validation whereby the full Simpson et al. dataset was split randomly into two samples, with roughly equal numbers of ratings for each item, and roughly equal numbers of trials for all items per sample. Both samples could then be normed independently, so that consistency of estimates across samples could be examined (Figure 3.16A). I compared the results from CLMM-derived estimates to both raw means, and estimates derived from the random effects structure of a linear mixed effects model (LMM). This latter comparison was employed to delineate the effects of shrinkage and of accounting for participant variability which result from the random effects structure (LMM versus raw means) from the effect of treating the scale as ordinal rather than continuous

(CLMM versus LMM). The LMMs were fit via the *lme4* package for R (Bates et al., 2015), with a Gaussian identity link for comparability to the raw means. Both CLMMs and LMMs were fit with item and participant random intercepts. The process of splitting the data in two, and estimating norms for both samples using each of the three methods, was repeated 100 times. I could then examine the distribution of differences between the two samples of all iterations (Figure 3.16B). I expected this distribution of differences to show reduced spread when results are more consistent, and to have greater spread when results are less consistent. I found that across most of the Likert scale, the CLMM-derived norms were more consistent between separate samples. The exception was at the lower end of the scale, where items were overwhelmingly responded to with a rating of *1* (i.e., "not at all similar"). Here, I found that raw means and estimates derived from LMMs were in fact more consistent than estimates derived from CLMMs. However, this consistency is in fact illusory, since raw means and LMMs fail to account for the finite bounds of the Likert scale, resulting in very small variance estimates due to floor and ceiling effects. In contrast, the CLMM approach provides greater discriminatory power, by estimating differences in a latent distribution which does not suffer from the same bounds. The CLMM-derived estimates are therefore necessarily more variable in the extremes of the scale, where non-ordinal approaches would typically suffer from floor or ceiling effects. In such cases, the estimates provided by CLMMs are more informative than estimates from approaches like raw means and LMMs. Indeed, in the Bayesian distributional model I found that for items affected by floor effects, uncertainty in the latent mean increases markedly with distance from the location of the lowest threshold (lower section of Figure 3.16B). In such cases, the responses are too consistent to provide much statistical certainty, and estimates increasingly rely on other, less directly informative features in the data. For example, items' latent means will be adjusted based on the random effects of participants who provided the items' ratings (as in Simulation 3). This suggests another advantage of Bayesian modelling, in that it can allow researchers to describe the certainty of their estimated norms. However, it is important to note that comparable measures could be calculated for non-Bayesian models via a Monte-Carlo re-analysis of the results. Moreover, the increased uncertainty observed at extreme values highlights the importance of considering the design of rating scales and wording of instructions provided to participants. By carefully wording the task's instructions or anchoring responses with labels (Hollis & Westbury, 2018), researchers may be able to systematically shift participants' responses away from a floor or ceiling effect, such that differences in the probabilities of the possible ratings allow items to be normed with greater certainty, consistency, and precision. Nevertheless, this analysis demonstrates that a CLMM approach to norming items is more robust to floor and ceiling effects than non-ordinal alternatives.

All code, and the fitted Bayesian model, for this re-analysis of the data from Simpson et al. (2013) is available in the OSF project associated with this work, at https://osf.io/ntvmf/. This project additionally contains RMarkdown documents showing and explaining minimal examples of how to use CLMMs to norm items. For researchers who are unfamiliar with R but wish to compare raw to CLMM-derived estimates, I designed and made freely available a web app that supports norming items via CLMMs: https://github.com/JackEdTaylor/shinynorms. This app provides functionality to fit equal-variance CLMMs via the *ordinal* package, with the item and

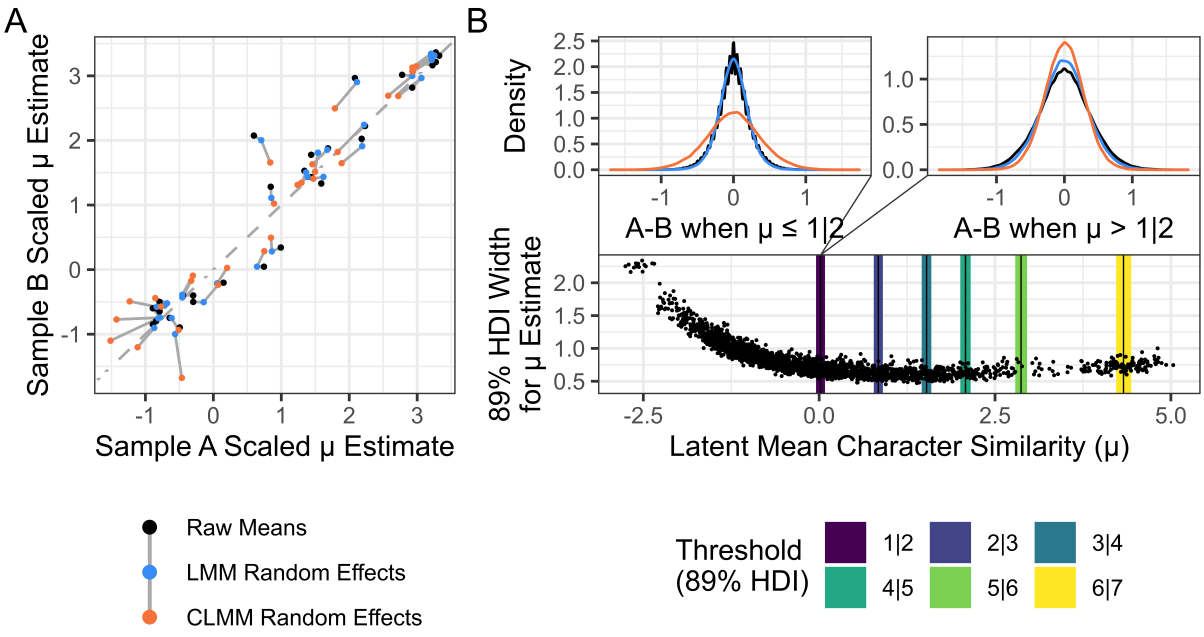


Figure 3.15: Results of the analysis examining consistency in norms estimates for three approaches: raw means, LMM random effects, and CLMM random effects. Panel A depicts the test-retest consistency for 30 example items, from one test-retest iteration. Estimates for the same items, from the three different approaches, are joined by grey lines, while the dashed diagonal line depicts perfect consistency between the two samples for reference. The lower region of panel B shows how uncertainty in latent mean estimates (width of the 89% HDI) from the Bayesian distributional model varies across the latent scale. The coloured bands depict the 89% HDIs for the estimated threshold locations (e.g., 1|2 is the threshold between ratings of *1* and *2*). The upper region of panel B shows the distribution of differences between samples A and B, with observations combined from 100 iterations of the test-retest procedure. The differences are depicted separately for items which were either below (left) or above (right) the lowest threshold. The jagged appearance of the density plot of differences for raw means reflects that there is only a finite number of possible values the average rating can take without the additional discriminatory power provided by random effects.

participant random effects downloadable alongside traditional summary statistics of means and *SD*s.

3.3.2 Glasgow Norms Analysis

Scott et al. (2019) collected ratings for 5,553 words on nine semantic dimensions from 829 participants. Six variables (Age of Acquisition, Concreteness, Familiarity, Gender, Imageability, and Size) were rated on 1-7 Likert scales, while the remaining three variables (Arousal, Dominance, Valence) were rated on 1-9 scales. I fit CLMMs to trial-level data from the Glasgow Norms, for all nine dimensions. I identified and excluded some duplicate entries in the trial-level dataset - 45 trials were excluded because they were duplicate data entries (i.e., identical in all fields, including timestamp, participant ID, item ID, and response), and an additional 153 observations were excluded because, in error, the experiment platform presented the same item to participants more than once (in such cases, only the first presentation of each item was included in the analysis). I also excluded trials where participants, instead of rating a word on the given dimension, indicated that they did not know the word, which was an option for all dimensions except Familiarity (as an unknown word should receive a rating of *1* in Familiarity). Following these exclusions, I then applied the response time filter as used in the original Glasgow Norms analysis, excluding trials with response times less than 600 ms. After these exclusions, there were a total of 1,679,206 observations across all nine dimensions. As in the analysis of reliability and discriminatory power reported for the Simpson et al. data, I estimated norms for each of the nine variables using three different methods: raw means, LMM random effects, and CLMM random effects. Each semantic variable was normed separately. As in the Simpson et al. analysis, distributions of estimates (Figure 3.16) were very similar for raw means and LMM random effects, while CLMM random effects distributions showed larger changes - especially at scales' extremes where ceiling and floor effects were more likely. LMMs were fit with *lme4*, and equal-variance CLMMs were fit with the *ordinal* package, while all model formulae were specified as in the Simpson et al. analysis:

```
rating ~ 1 + (1 | item_id) + (1 | participant_id)
```

In this analysis, I was interested in comparing the predictive power of norms calculated traditionally to norms derived from LMMs and from CLMMs. Scott et al. (2019) identified three strong correlations between the normed variables: Age of Acquisition - Familiarity (r=-.67), Concreteness - Imageability (r=.91), and Valence - Dominance (r=.68). I compared how well the norms calculated from raw means, LMMs, and CLMMs could account for these relationships. I additionally examined the relationship between subjective familiarity norms and objective word frequency estimates (log10 word frequency per million from SUBTLEX-UK; van Heuven et al., 2014), as research has shown these variables to be highly correlated (Brown & Watson, 1987; Scott et al., 2019; Tanaka-Ishii & Terada, 2011). To compare the predictive power of the

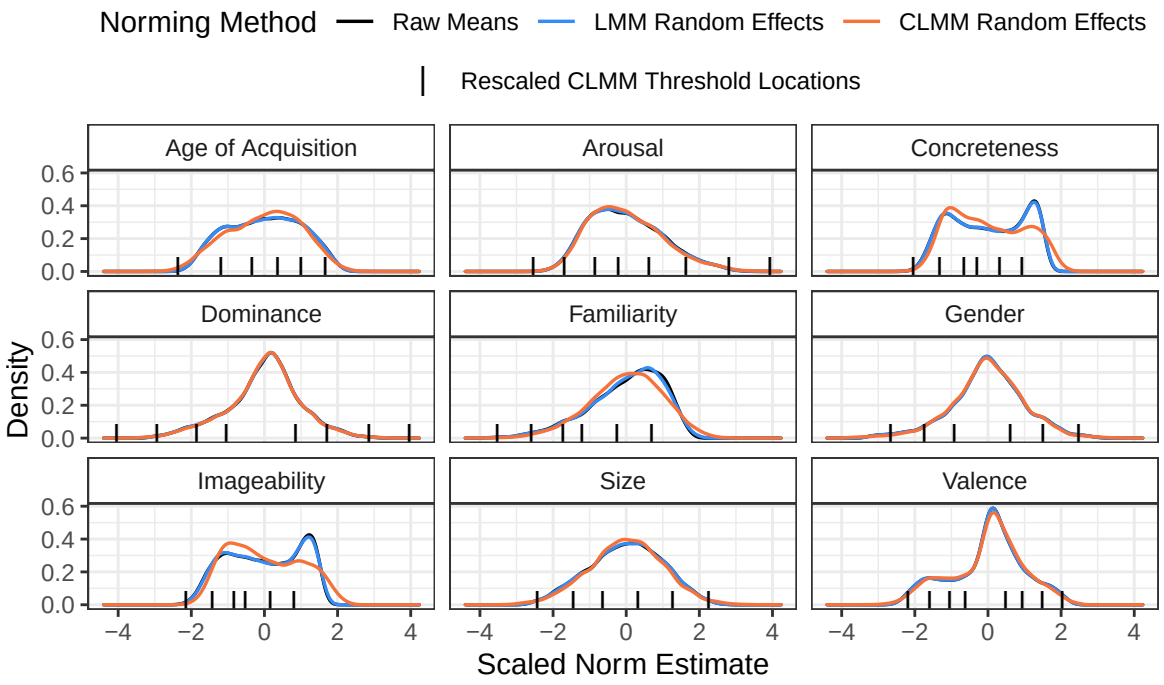


Figure 3.16: Distributions of words' rating norms from the Glasgow Norms dataset, calculated from raw means, or from the random effects estimates of LMMs and CLMMs. Densities of raw means and LMM random effects are usually so similar as to overlay one another. Densities are scaled by *SD*s for comparability between variables and methods of estimation. Vertical black lines depict the estimated locations of CLMM thresholds.

relationships, when norms were estimated using the three different methods, I estimated the R^2 values for variables' linear relationships. Results (Figure 3.17) revealed that estimating norms from LMM random effects sometimes resulted in small improvements in the proportion of variance explained by relationships, and that CLMM-derived norms usually yielded the largest proportion of variance explained. The main exception was the correlation between Valence and Dominance norms, for which similar proportions of variance were explained when norms were calculated from any of the three methods, reflecting that for these two variables, norms calculated from the three methods were very similar. In sum, this reanalysis of the Glasgow Norms showed that the greater reliability and discriminatory power provided by CLMMs confer greater predictive power, and suggest that CLMM-derived norms better index the dimensions that rating scales aim to capture.

3.4 Discussion

Norming studies which use ordinal scales constitute a vital resource for research and are applied to a wide range of scientific applications. It is therefore important that the reported norms accurately reflect the inter-item relations on the dimension of interest. Informed by the simulations reported here, I argue that to this end, norming studies should report estimates which appropriately account for nonlinearities in the ordinal norming scale, rather than inaccurately assuming that the ordinal scale is linear. Specifically, this chapter has shown that,

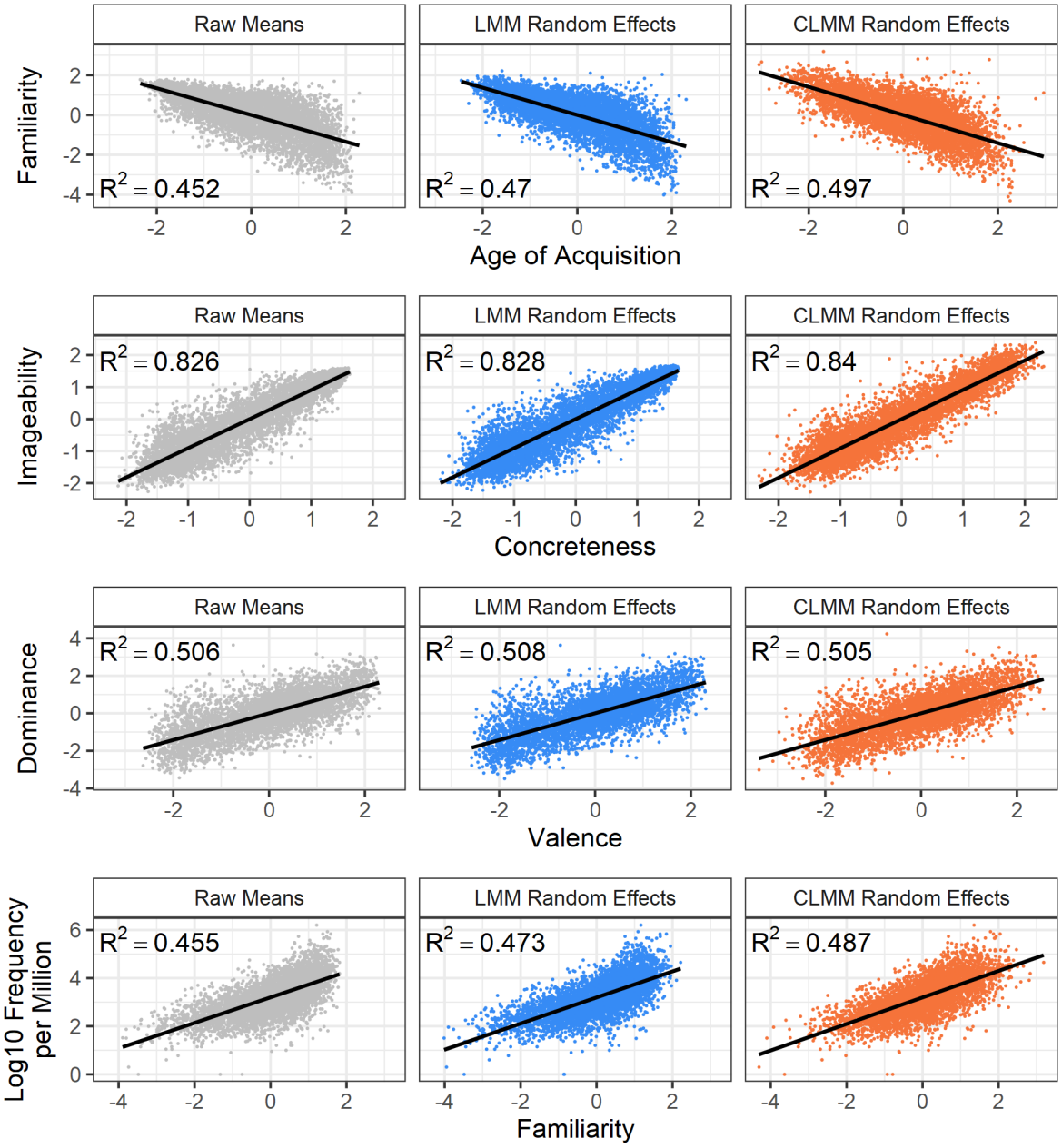


Figure 3.17: Proportion of variance explained in linear relationships when ratings are normed using means, LMMs, or CLMMs. Points depict estimates for individual words, while black lines depict the estimated linear relationships. For comparability, raw means, LMM estimates, and CLMM estimates are scaled consistently, by their *SD*s.

when using an ordinal Likert scale to norm items, traditional methods such as raw means and *SD*s can lead to systematically distorted item comparisons. On the other hand, properly accounting for the ordinal nature of the judgements via CLMMs provides estimates which are far less affected by participants' response mappings. While this problem is generally well-understood in studies that aim to estimate fixed effects of experimental manipulations (Liddell & Kruschke, 2018), the problem has been less widely discussed in relation to norming studies.

From this chapter, the chief recommendation is that items are normed via the random effects structure of hierarchical ordinal models like CLMMs. In this way, researchers will be able to estimate norms which are appropriately disentangled from the nonlinearities in response patterns. However, while random effects estimates extracted from CLMMs will provide more accurate estimates for norming studies, such studies should report them in addition to more traditional measures like Likert means and *SD*s, rather than instead of them. This will be important for ensuring that results are still comparable with existing datasets which only report means and *SD*s, and for users of the data to examine the differences between estimates from ordinal models versus traditional summary statistics.

In addition to more accurately norming items' central tendencies, this chapter has demonstrated how CLMMs can be used to explicitly model and norm latent variances, which requires the application of a Bayesian distributional modelling approach. While researchers are typically most interested in items' central tendencies, ambiguity in judgements can also be of great theoretical relevance (e.g. Brainerd et al., 2021). Explicitly modelling differences in latent *SD*s, rather than assuming equality of variance, could offer a more meaningful analogue to the traditional Likert *SD* reported in norming studies. This recommendation is informed by the finding that *SD*s of Likert ratings reflect both meaningful differences in latent variance, and artefacts of nonlinearities in response patterns. I note that when distributional CLMMs are used to disentangle meaningful and artefactual contributions to items' *SD*s, the striking mean-*SD* relationships identified as problematic by Pollock (2018) are no longer observed. As a result, I argue that this statistical concern raised by Pollock is not inherently problematic. Rather, it reflects consequences of treating ordinal scales as continuous. This conclusion is likely to have important consequences in research concerned with whether concepts like concreteness are categorical or continuous in nature (e.g., compare Pollock, 2018, to Gentner and Asmuth, 2019). While CLMMs necessarily *assume* that a continuous distribution underlies ordinal responses, the results of this chapter demonstrate that conversely, the raw frequencies of different ordinal categories cannot be interpreted as evidence for a concept existing categorically or continuously.

It is notable that the methods required to estimate Bayesian distributional CLMMs tend to be more computationally complex, and correspondingly, take substantially more time to fit, than more standard equal-variance CLMMs. The difficulty of fitting such models may even become unfeasible for especially large datasets. Researchers may therefore wish to ignore differences in latent variance, to focus only on the perhaps more theoretically relevant estimates of latent means. In Simulation 3, I showed that assuming equality of variance in this way does not seem to reduce the accuracy of latent mean estimates to any great extent. In addition, in

the analysis of the character similarity dataset (Simpson et al., 2013), I showed that fitting a maximum likelihood model assuming homogeneity of latent variance can provide highly similar estimates of latent means to those from a Bayesian distributional model. As a result, I argue that if researchers are not interested in reporting latent variances, then simpler, equal-variance models can generally be used to estimate latent means without any great loss in accuracy. However, given that other researchers may be interested in estimates of latent SDs, making the trial-level dataset publicly available regardless would allow other researchers to model such differences if they wish.

All the simulations and analyses presented in this chapter fit models which assume a single response pattern across all observations (but which can take different shapes such as equidistant, left-biased, right-biased, centre-biased, edge-biased, etc.). This assumption is likely appropriate for many normed variables, as reflected in how the artefacts of response patterns are clearly observed when data is collapsed across participants (see Figure 3.2, Figure 3.3, and Figure 3.12). However, researchers may observe that different participants, or indeed items, show distinct response patterns. In this case, a considerable degree of accuracy will be lost by failing to account for such participant- or item-related dependencies. A solution could be, rather than assuming that all participants (or items) display the same overall response pattern, to model response patterns per-participant or per-item, or both (Bolt & Johnson, 2009; Jonas & Markon, 2019). Indeed, CLMMs can be specified to model such variability with *brms*, by using the *thres()* term to provide participant or item IDs as a grouping variable for which thresholds should be calculated separately: e.g. *response | thres(4, gr=participant_id) 1 + (1|item_id) + (1|participant_id)*). Such models can be very computationally intensive, adding a large number (number of participants * number of thresholds) of parameters that need to be estimated. This is especially likely to make results from large-scale norming studies difficult to estimate (e.g., N=4,237 participants in Brysbaert et al. (2014)). However, modelling data in such cases may be made more tractable with statistical approaches like that outlined by Selker et al. (2019), which allows an arbitrary number of thresholds in a latent distribution to be estimated via just two parameters per participant. As a result, researchers may consider calculating thresholds separately for individual participants, although I do not evaluate the performance of such models here. An alternative could be to use a different grouping variable with fewer levels, but which accounts for differences in response patterns relatively well. As an example, it may be that differences in, say, reading skill (high, medium, low) could account for variability in participant-related response patterns such that the skill groups show distinct response patterns. In this case, calculating thresholds separately for each skill group will allow the norms to be better disentangled from response patterns, while only requiring a few more parameters to be estimated. Importantly, whether splitting estimates of threshold locations by grouping variables is appropriate, and which grouping variables it would be most appropriate to split by, will differ between norming studies and participant samples. I also note that such considerations may benefit from further investigation in future research.

Throughout the simulations and reanalysis, I have used CLMMs with probit-link functions to model Likert responses. While the probit-link is convenient for estimating latent parameters more directly comparable to traditional means and *SD*s (as it assumes the latent variable is

normally distributed), other link functions can be equally appropriate, given that the true shape of the latent distribution is usually unknown. Altering the link functions for CL(M)Ms typically results in only small changes in model parameters (McCullagh, 1980). In Simulation 4, I showed that CLMMs fit with a single link function can estimate item random effects similarly well, regardless of different violations in the assumption of the latent variable and random effect distributions. Indeed, no single link function is likely to be superior for modelling rating data in all cases. If researchers wish to check that the link function they use is appropriate, they may want to fit several models to the data, using different link functions but identical formulae. Researchers could then select the model which best accounts for the data, assessed via measures of model fit such as log-likelihood. Regardless, I recommend that researchers always report the link function they used to model responses.

Similarly, all the CLMMs presented in this chapter were fit using flexible thresholds. This means that no constraints were imposed on the possible positions of the thresholds which demarcate the ordered regions of the latent distribution. An alternative would be to specify necessary features of the threshold locations, such as symmetry (around the mode of the latent distribution), or equidistance between thresholds. Estimating flexible thresholds is likely to be the most informative and most generalisable option when fitting a CLMM. There may be cases when specifying constraints on threshold locations is desirable for norming items, but in such cases researchers should clearly explain and justify the use of non-flexible thresholds.

In contrast to the models examined in this paper, which focus on random effects, norming studies have frequently separated results by demographic features of participants, like gender and age (e.g., Engelthaler & Hills, 2018; Grühn & Scheibe, 2008; Kanske & Kotz, 2010; Warriner et al., 2013), or features of experimental design, such as counterbalanced order of task (e.g., Salmon et al., 2010). Similarly, researchers frequently report correlations with features of items, such as other normed or corpus-derived variables (e.g., Pexman et al., 2017; Pexman et al., 2019; Scott et al., 2019; Stadthagen-Gonzalez & Davis, 2006; Warriner et al., 2013). While such effects could be estimated by examining correlations between relevant variables and random effect correlations, such variables could alternatively be incorporated into the CLMM, thereby accounting for the hierarchical variability of such effects in the random effects structure. For instance, a model estimating item norms, while also estimating the effects of age and gender of participants on ratings of individual items, could be estimated with random slopes as follows:

```
rating ~ 1 + (1 + age + gender | item_id) + (1 | participant_id)
```

A key advantage of using CLMMs to more accurately norm items is a reduction in measurement error. As an example, consider studies examining effects of normed features of words like concreteness and imageability on behavioural or neural correlates (e.g., Goh et al., 2016; Khanna & Cortese, 2021). Such studies will be able to estimate effects more accurately, and with greater statistical power, if the normed variables more accurately reflect the underlying variable of interest, disentangled from artefacts of response patterns. Similarly, research that

aims to expand the breadth of norming studies by predicting responses for unpresented items, for example via latent semantic analysis (Bestgen & Vincze, 2012), will be able to provide more accurate predictions, without simply reproducing artefacts of response patterns, if the models predict latent means rather than Likert means. The advantages of CLMMs are also likely to provide benefits in stimulus design. For instance, many variables in the LexOPS dataset (chapter 2) are calculated from averages of raw ordinal responses. Consider that the greater precision and discriminatory power provided by norms derived from hierarchical ordinal models will confer improved precision in stimulus matching when the normed variable reflects an experimental confound, and is used as a continuous control variable to minimise the confound's impact. This improved precision in matching will in turn provide more accurate and meaningful insight into the results of the experiment that the designed stimuli are generated for, reducing the impact of this experimental confound.

Not all norming studies employ ordinal scales; the recommendation to use CLMMs applies mostly to studies norming participant ratings, which are inherently ordinal. Some norming studies, meanwhile, norm variables which are clearly not ordinal. For example, participants may provide norms to a binomial decision, such as whether they know a given word (word prevalence; Brysbaert et al., 2019). I argue, however, that such studies can still benefit from using hierarchical modelling to pool observations and norm items more accurately. For example, random effect estimates from a binomial generalised linear mixed effects model could be used to norm word prevalence more accurately, concurrently accounting for item and participant variability, and appropriately adjusting outliers towards more accurate estimates via shrinkage. On the other hand, researchers may use scales which appear more continuous than the 5-point, 7-point, and 9-point scales most commonly used in norming studies. For instance, participants may be asked to rate items on a scale from 0 to 100 (e.g., Ma et al., 2015; Z. Yao & Wang, 2013). In this case, however, the only difference is in granularity: the latent continuous variable is simply separated into more regions. Participants will still show nonlinear response patterns in their judgements, biased towards some region of the scale. For such a large Likert scale there also likely to be additional sources of nonlinearity, such as ratings biased towards numbers which are multiples of 5 or 10.

Finally, it is important to note that there is a rich literature of existing recommendations for the formulation of Likert scales. Although such recommendations often assume the use of traditional Likert means for norming, such recommendations still hold true for norming studies using the methods of analysis that are recommended here. Researchers should still carefully consider the phrasing of their questions and the instructions given to participants so as to maximise their sensitivity to the underlying variable they are interested in (Connell & Lynott, 2012; Hollis & Westbury, 2018). This will allow researchers to avoid undesirable outcomes such as floor and ceiling effects, which necessarily reduce the precision of estimates (as in Figure 3.16B). Similarly, researchers should still consider whether collecting subjective judgements is informative or useful for the variable they are interested in. Regardless of how subjective judgements are analysed, they will still be inherently subjective. As an illustration, imagine a study utilising the Müller-Lyer illusion, where the sizes of lines are perceptually distorted by inward- and outward-pointing arrowheads at each end. Suppose that participants

are asked to provide a Likert scale rating of how similar the two lines are in their lengths. Even if the ordinal nature of the scale is accounted for, estimates on the latent distribution will still be biased by the perceptual illusion, away from the lines' objective lengths. This is to say, the latent variable will be disentangled from response patterns, but will inherently reflect subjective perceptions, which may not necessarily align with objective reality.

To summarise, in this chapter I have shown that CLMMs support more accurate norming of items than traditional statistics of means and *SD*s, which treat the scale is continuous rather than ordinal. Summarising items via estimates of their latent means and *SD*s provides an analogue to traditional analyses, with the advantage of appropriately disentangling variables of interest from artefacts of nonlinearities in participants' response patterns.

Chapter 4

Category-Level Top-Down Modulation of the N1 via Task Manipulation

4.1 Introduction

A key question in research into visual word recognition is whether early orthographic processing is influenced by top-down modulation. While there is support for higher-level contributions to occipitotemporal regions associated with orthographic processing (Bouhali et al., 2014; L. Chen et al., 2019; Vogel et al., 2012), a demonstration that these connections functionally influence early stages of word recognition requires disentangling early, prelexical processing from later processing that occurs after word recognition. The high temporal resolution of electroencephalography (EEG) affords such insight into the timing of cognitive processes in the brain; the extent to which, and latency at which, activity recorded in EEG is sensitive to higher-level information can delineate the timeline of sensitivity of early processes to top-down modulation. The occipitotemporal N1, the first negative-going event-related potential (ERP) component observed in response to individual words, has been widely associated with orthographic processing (Bentin et al., 1999; Brem et al., 2006; Maurer, Brandeis, et al., 2005). If the N1 is sensitive to higher-level task or semantic information, when this can be predicted from preceding context, this is likely to indicate a functionally meaningful influence of top-down modulation on early orthographic processing. A useful step in research on effects of predictability and top-down modulation is to identify the limits of their influence (Luke & Christianson, 2016; Van der Stigchel et al., 2009). In this chapter, I report the results of an experiment investigating whether the N1 is sensitive to predictions at the level of categorical semantic information, setting an upper bound on top-down modulation of early orthographic processing.

In addition to approaches that bias participants' expectations via linguistic (see Nieuwland, 2019, for a review) or non-linguistic (e.g., Dikker et al., 2009; Kherif et al., 2011) cues in stimuli, a common paradigm for investigating top-down modulation uses task manipulations. Here, changes in task cue attention to, or predictions of, different features of stimuli. An advantage of using task manipulations to investigate top-down modulation is that researchers can present identical stimuli between tasks, such that bottom-up processing can be matched exactly. Most commonly, paradigms using task manipulations have compared responses

in tasks that explicitly require different levels of processing. Commonly used tasks in this paradigm include semantic categorisation tasks (SCTs), explicitly requiring word recognition and semantic processing; lexical decision tasks (LDTs), explicitly requiring word recognition but not necessarily semantic processing; and perceptual or sublexical tasks, explicitly requiring only non-linguistic perceptual processing or attention to sublexical elements such as individual letters in a word. Much research on task manipulations has focused on questions of automaticity in word recognition. For example, if readers show sensitivity to word frequency effects during LDTs, but not during perceptual tasks, then it is possible to conclude that word frequency does not automatically influence word recognition processes, but rather depends on task demands. Combining task manipulations with the temporal resolution of EEG provides insight into the timing of such task-stimulus interactions.

One line of research that has used task manipulations to investigate automaticity has focused on the N400, a centroparietal ERP component observed around 400 ms after word presentation that is typically associated with semantic processing (Kutas & Federmeier, 2011). For instance, Chwilla et al. (1995) presented words in two tasks: an LDT, in which participants discriminated between words and orthographically and phonologically plausible pseudowords, and a perceptual task, in which participants discriminated between lower- and upper-case words. Each target word was preceded by a prime, for 200 ms, with a stimulus onset asynchrony (SOA) of 700 ms, which was either semantically related or unrelated to the target. Results showed that prime-target relatedness only influenced N400 amplitude in the LDT. In the perceptual task, in which lexical or semantic processing was unnecessary, no effect of prime-target relatedness on the N400 was observed. Such effects of task on the N400, indicate that processing in the N400 is to some degree strategic, subject to top-down control, rather than fully automatic (though the N400 is not fully strategic either; Kutas & Federmeier, 2011). Task dependence has also been observed for other late ERP components, such as the P600, a positive-going component, peaking centroparietally around 600 ms after stimulus presentation, that is thought to reflect syntactic and late semantic processing (Brouwer et al., 2012; Molinaro et al., 2011). For instance, syntactic and semantic violations elicit a clear P600 when the violations are task-relevant, such as when participants respond to stimuli with sentence correctness judgements (Martín-Loeches et al., 2006), but are harder to detect or absent when such violations are task-irrelevant, such as in a probe verification task (i.e., "was this word in the sentence?"; Schacht et al., 2014). Hahne and Friederici (1999, 2002) applied a similar task manipulation paradigms to examine whether an earlier syntax-sensitive component, the early left anterior negativity (ELAN; peaking 100-300 ms post-stimulus), shows a comparable modulation from task demands. Results showed that whereas sensitivity to semantic and syntactic violations in the N400 and P600 was modulated by task demands, the ELAN was seemingly insensitive to task demands. On this basis, Hahne and Friederici (1999, 2002) argued that higher-level influences only affect later processing, while early syntactic processing is automatic and impervious to top-down modulation (although see Steinhauer & Drury, 2012, for a discussion of whether the ELAN reflects early syntactic processing or is carried over from processing of preceding words).

A similar argument was originally made for task-insensitive automaticity in early

occipitotemporal processing of word forms (Posner et al., 1989). In contrast to the ELAN, however, much research has suggested that the occipitotemporal N1 may be sensitive to task manipulations. These findings are detailed in section 1.5.2 of the general introduction. To summarise the findings here, the N1 shows a sensitivity to task demands similar to that observed for the N400 and P600. N1 amplitudes are more negative-going in tasks explicitly requiring lexical or semantic processing than in non-lexical perceptual tasks (Y. Chen et al., 2013; F. Wang & Maurer, 2017). Furthermore, tasks explicitly requiring lexical or semantic processing cause the N1 to show greater sensitivity to word forms' orthographic legality (Bentin et al., 1999), and word frequency (Y. Chen et al., 2015; Strijkers et al., 2015). In one interesting use of task manipulations, F. Wang and Maurer (2020) presented Chinese speakers with Chinese characters, or stroke-matched Korean characters. Participants were required to categorise characters as either Chinese or Korean, with coloured (blue or green) frames preceding the stimulus cueing the likely category of the upcoming character. Unlike studies that bias participants' attention towards different aspects of the stimulus (perceptual, lexical, semantic, etc.), this task biased expectations for categories of word forms. F. Wang and Maurer (2020) showed that, in the period of the N1 that follows its peak (i.e., its offset), the effect of orthographic familiarity, where unfamiliar Korean characters elicited more negative-going N1 components than familiar Chinese characters, was greater when the task cued participants' expectations towards Chinese characters, and smaller when participants were cued to expect Korean characters.

Wang and Maurer's paradigm differs from much research on top-down modulation of the N1, which has typically either examined sensitivity to features of word forms like frequency (as summarised above), or has biased expectations towards specific, individual word forms (Nieuwland, 2019). The findings of F. Wang and Maurer (2020) suggest that expectancy for entire orthographic *categories*, when these categories differ in orthographic familiarity or legality, may modulate N1 responses. If the N1 is sensitive to expectations of orthographic categories, it may also be sensitive to predictions of categories of word forms like *semantic* categories. A key difference, here, is that both category members and non-members would be orthographically legal and familiar, such that sensitivity to category membership would indicate an influence of targeted predictions on processing during the N1. One previous study that examined whether early ERP components are sensitive to information at the level of semantic categorisation was reported by Segalowitz and Zheng (2009). Here, Segalowitz and Zheng presented participants (N=14) with different sets of stimuli in two very similar LDTs, where stimuli within each block either all belonged to the same category, or else were drawn from multiple categories. It is notable that, unlike most research that uses task manipulation paradigms, the two tasks did not greatly differ in task demands; category membership was a largely implicit feature of words, with the more semantic LDT only probing knowledge of category membership at the end of each block (asking participants to identify which of four words would be a member of the same semantic category as those presented in the preceding block). Segalowitz and Zheng's stimuli comprised 100 words, drawn from five semantic categories, and 100 pseudowords designed to be orthographically and phonologically plausible. Segalowitz and Zheng reported a task-stimulus interaction on the N1, wherein

amplitudes for words were more negative-going when words were drawn from the same semantic category, while N1 amplitudes observed for pseudowords did not differ between tasks. While this may demonstrate an effect of task demands on sensitivity to lexicality, consistent with research cited above, it may also demonstrate a task-dependent sensitivity to semantic category membership. Disentangling these two explanations would require two types of word stimuli: category-relevant and category-irrelevant words. Presenting these stimuli alongside non-lexical stimuli like pseudowords and nonwords would allow the disentanglement of the two possible explanations for the effect reported by Segalowitz and Zheng (2009).

Such a comparison, between category-relevant and category-irrelevant words in an SCT, was examined by Hauk et al. (2012), who presented words, matched on multiple orthographic variables, that refer to either living or non-living objects. In a go-no-go SCT, analysing results across the full topography simultaneously, Hauk et al. reported that differences emerged as early as 166 ms (although the difference peaked later, at 338 ms). Hauk et al. also presented a separate set of items in an LDT, reporting a difference between words and pseudowords, matched for orthographic features, as early as 168 ms, suggesting an early N1 sensitivity to lexicality, in addition to category relevance. If the effect reported by Hauk et al. in the SCT does indeed reflect category relevance, then the difference between category-relevant and category-irrelevant items should be expected to be absent in an LDT (assuming the category-relevant words are not salient enough for participants to notice the shared semantics and infer the underlying categories). Alternatively, if the category difference is present in both an SCT and LDT, then the difference may not reflect a sensitivity to category relevance, but rather to bottom-up features that differ between the stimuli (e.g., Hauk et al.'s category-relevant and -irrelevant words differed slightly in average word frequency).

In this experiment, I examine whether expectation for a semantic category of word forms modulates the N1. I apply a task manipulation paradigm to compare N1 responses to words during an SCT, in which words' membership of semantic categories is task-relevant, to during a LDT, in which semantics is not explicitly relevant. Unlike Segalowitz and Zheng (2009), the task manipulation has been designed to make semantic category membership explicitly relevant to the SCT, and not the LDT, while, unlike Hauk et al. (2012), the same items are presented in both tasks. By including both category-relevant and category-irrelevant word stimuli in the SCT, in addition to non-lexical pseudowords and nonwords, I am able to disentangle any observed effects indicative of top-down modulation, thereby distinguishing between the two interpretations of the the effect that Segalowitz and Zheng report, and ensuring hat any effect of category relevance is not related to bottom-up differences. If the N1 shows an interaction between task and category-relevance of stimuli, where category relevance influences the N1 during the SCT but not during the LDT, this would provide evidence in favour of top-down modulation of the occipitotemporal N1 at the level of the semantic category. I show that while this task-stimulus interaction does emerge, it is observed later than the N1, setting an upper bound on the top-down modulation of early orthographic processing. Further, while I replicate sensitivity to orthographic features in the N1, no sensitivity to lexicality is observed, and no clear interaction between task and either orthographic plausibility or lexicality is observed in the N1. I discuss the findings, arguing that differentiation between whole categories of words on the basis of

orthography is difficult for an alphabetic script, in which a small number of graphemes are reused and will necessarily occur in both category-relevant and -irrelevant word forms.

4.2 Methods

In this experiment, words commonly listed as members of semantic categories (Van Overschelde et al., 2004) were matched with category-irrelevant words. Words were presented in an LDT (with category-relevant words randomised across blocks) or SCT (with category-relevant words randomised *within* category-specific blocks). In addition to category-relevant and -irrelevant words, pseudowords (orthographically and phonologically plausible) and nonwords (consonant strings) were presented as matched non-lexical stimuli for the LDT, additionally providing a validation of the orthographic sensitivity of the N1.

4.2.1 Design

Participants were randomly assigned to one of four groups: A, B, C, or D. All participants completed both the LDT and SCT, but the order in which the tasks were completed, and the items presented in each task, were counterbalanced. A summary of the experimental design is shown in Table 4.1.

Table 4.1: The order of tasks, and the blocks of stimuli presented in each task, for the four participant groups. The final two columns indicate the number of participants assigned to each response mapping scheme (respectively, where the right-handed response is affirmative, and where the left-handed response is negative) - response mappings are explained in section 4.2.4.

Participant Group	Task Order	Blocks in LDT	Blocks in SCT	*N* per Response Mapping
A	LDT, SCT	1,2,3	4,5,6	3,3
B	SCT, LDT	1,2,3	4,5,6	4.2
C	LDT, SCT	4,5,6	1,2,3	4,3
D	SCT, LDT	4,5,6	1,2,3	5,5

4.2.2 Participants

A total of 38 participants were selected by opportunity, though of these, 9 participants were excluded due to problems with the EEG recording. Of the participants excluded, 5 were excluded due to issues with the EEG setup meaning that data were missing or triggers were not recorded, and 4 were excluded due to high offsets (more extreme than ±20 μV) or excessive noise producing unreliable ERPs .

Of the 29 participants included, 6 were in participant group A, 6 in B, 7 in C, and 10 in D. Ages ranged from 16 to 45 years (mean=24.48, *SD*=6.1), and 20 identified as female, while 9 identified as male. All participants reported right-hand dominance. A screening questionnaire

was used to additionally ensure that all participants were monolingual English speakers, and did not report being diagnosed with any condition or disability that impairs language ability or reading. All participants reported having normal vision, or else wore glasses or contact lenses if they usually wear them to read.

Data collection was approved by the University of Glasgow School of Science and Engineering Ethics Committee (application number: 300170093).

4.2.3 Stimuli

Stimuli comprised 496 items, consisting of 124 category-relevant words, category-irrelevant words, pseudowords, and nonwords, respectively. A list of all items is presented in Appendix B.1. Category-relevant words were drawn from a set of semantic category norms reported by Van Overschelde et al. (2004), which comprise words commonly listed as members of semantic categories. I selected 6 semantic categories from the Van Overschelde et al. (2004) norms: Four-Footed Animals, Fruits, Musical Instruments, Parts of the Human Body, Relatives, and Things that Fly. Items longer than one word (e.g., *star fruit*) or ambiguous in meaning (e.g., *orange*) were excluded. As a result, all category-relevant words were category-specific, unambiguous, concrete nouns. Words in the category norms that were plural in the original were singularised (e.g., *teeth* was changed to *tooth*), and words specific to American English were replaced with British English equivalents (e.g., *ladybug* was replaced with *ladybird*). Finally, some items were excluded from categories such that each category block could be matched in length (number of items) with one other category.

Category-irrelevant words were matched to category-relevant words manually using features of an early version of LexOPS (J. E. Taylor et al., 2020). Category-relevant and -irrelevant items are summarised in Figure 4.1A. Category-irrelevant words were selected to match category-relevant words exactly in word length, and closely in word frequency. Where suitably close matches could not be found, items were hand-selected to match stimuli distributionally. Items were additionally selected to produce similar distributions in variables relevant to orthographic processing, and likely relevant to performance in the LDT and SCT: othographic neighbourhood size (OLD20; Yarkoni et al., 2008), character bigram probability (calculated from SUBTLEX-UK; van Heuven et al., 2014), age of acquisition ratings (from Scott et al., 2019), and concreteness ratings (from Scott et al., 2019).

Pseudowords were formed manually, designed to be orthographically and phonologically plausible, while nonwords were formed as strings of randomly selected consonants (excluding the letter *y*), designed to be unpronounceable and orthographically implausible. Pseudowords and nonwords were matched on length exactly, to pairs of category-relevant and -irrelevant words. Distributions of pseudowords and nonwords on orthographic variables, average character bigram probability and OLD20, are presented in Figure 4.1B, showing that pseudowords show similar distributions to category-relevant and -irrelevant words, while nonwords are highly dissimilar from real words in these variables.

Stimuli were split into 6 blocks, for each of the 6 semantic categories of category-relevant words. For the SCT, stimuli were pseudorandomised within blocks such that no more

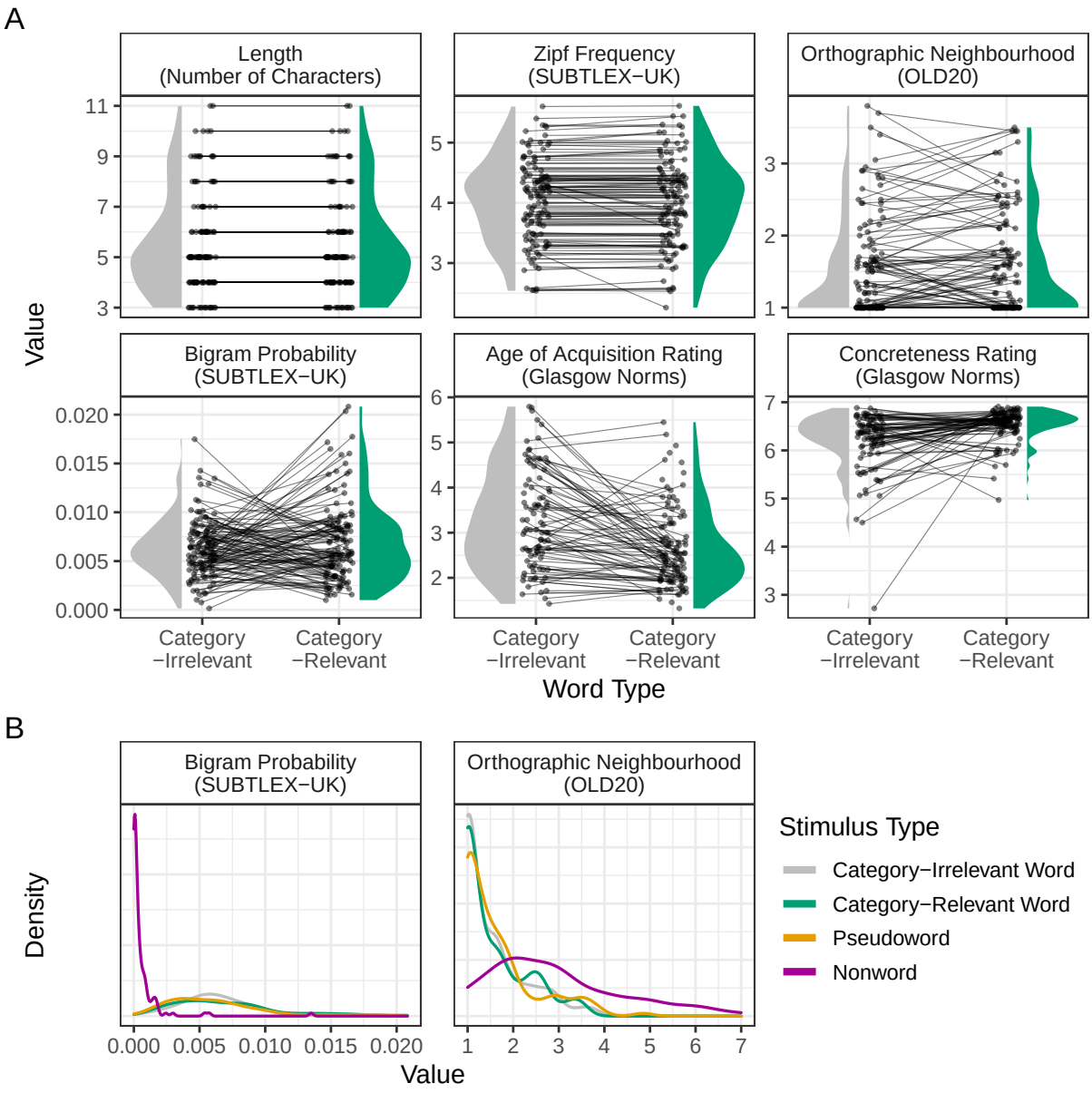


Figure 4.1: Summary of stimuli features. (A) Matched category-relevant and -irrelevant words. Points denote individual items, while lines joining points denote that items were matched. Where possible, pairs were matched item-wise for length and frequency. All other variables were matched distribution-wise. Some points and lines are missing, reflecting words for which values were unavailable. Coloured shapes depict the densities of points. (B) Distributions of all four stimulus types on orthographic variables: character bigram probability and OLD20. Density is not scaled consistently across plots and are only comparable within variables.

than 3 consecutive trials were of the same stimulus type. For LDT trials, stimuli were pseudorandomised across the 3 blocks, rather than within them, to reduce the salience of the shared semantic categories of category-relevant words. Randomisation was performed separately for each participant.

4.2.4 Procedure

Participants completed the experiment resting on a chin rest at a distance of 50 cm from a VPixx Technologies VIEWPixx monitor (resolution = 1920*1080 px, diagonal length = 23") on which the stimuli were displayed. Stimuli were presented using the Psychophysics Toolbox extensions for MATLAB (Psychtoolbox 3; Kleiner et al., 2007). In both tasks, the following events occurred. (1) A blank grey screen was presented for 200 ms (equal to 50% of the maximum intensity in each colour channel). (2) A fixation cross was presented for 200 ms, shown as the "+" symbol (black, 40-point DejaVuSans font, .75° of visual angle high, and 1.03° wide) in the centre of the screen, on a grey background. (3) A blank grey screen was presented for between 300 and 1300 ms (jittered randomly; uniform distribution). (4) The target was presented in the centre of the screen for 400 ms, in black, 40-point DejaVuSans font on a grey background. Target stimuli ranged in height from .7 to 1.22° of visual angle (mean=1.04, *SD*=.16). As the font was not monospaced, width in visual angle was calculated for each stimulus individually, ranging from .47 to .9° (mean=.67, *SD*=.08) per character, and from 1.43 to 8.31° per word (mean=3.69, *SD*=1.36). (5) After the target, a blank grey screen was shown until the participant responded. Participants responded to each trial with either the right or left control (*Ctrl*) key on a standard QWERTY keyboard, where one of these buttons indicated an affirmative response, "yes", while the other indicated a negative response, "no". Mapping of the right and left control keys to affirmative and negative responses was pseudorandomly assigned for each participant before the experimental session, such that there were similar numbers of participants for each type of response mapping. Specifically, 16 participants responded with "yes" mapped to the right control key, and 13 participants with it mapped to the left control key. There were also similar numbers of participants for each response mapping across the participant groups that dictated presentation order (Table 4.1).

For the LDT, participants were instructed to respond "yes" to targets that are real words (i.e., category-relevant and category-irrelevant words), and to respond "no" to targets that are not (i.e., pseudowords and nonwords). For the SCT, participants were instructed to respond "yes" to targets that are members of the current category (i.e., category-relevant words), and "no" to items that are not (i.e., category-irrelevant words, pseudowords, and nonwords). These instructions were presented at the start of every block, which for the SCT included the name of the category that category-relevant words in the upcoming block belonged to.

Each participant completed 3 blocks of the LDT and 3 blocks of the SCT, with the order of tasks dictated by their participant group (see Table 4.1). The first block of each task was preceded by a practice block of 32 trials that were similar to the upcoming task. For the SCT, category-relevant words were names of flowers, while for the LDT, they were just randomly selected words. For the practice trials only, participants were provided with per-trial feedback on

the accuracy of their responses, with text reading "Correct" or "Incorrect" after each response.

4.2.5 Recording

EEG data were recorded from a 128-electrode BioSemi system in an electrically shielded booth, with channel positions conforming to the standard BioSemi 128-electrode arrangement. Four electro-oculography (EOG) electrodes were placed to record eye movements and blinks: 2 were placed to the sides of eyes (on the right and left outer canthi), and 2 below the eyes (on the infraorbital foramen). Data were recorded at 2048 Hz, with an online low-pass filter at the Nyquist frequency. Recordings were downsampled to 512 Hz using the BioSemi Decimator tool. Electrode offset was kept stable and low through the recording, within ±20 mV, as measured by the BioSemi ActiView EEG acquisition tool.

4.2.6 Preprocessing

Stimuli were preprocessed using MNE Python (version 1.0.2; Gramfort et al., 2013). Trials were filtered on the basis of accuracy (excluding 318 trials that were responded to incorrectly), and then on response time (excluding 170 trials with response times <250 or >1500 ms). Following these behavioural exclusions, there were 13,896 trials in total. Recordings were epoched to stimulus onset, with the 200 ms pre-stimulus as a baseline, lasting until a maximum of 1 second after the stimulus. As trials ended after participants' responses, this means that each trial had data until at least 250 ms after the stimulus, but that there were progressively fewer observations for later time points. Signals were bandpass filtered to between .1 and 40 Hz (causal, fourth-order Butterworth filter). To counteract the delay in ERP timing inherent to causal filters (which shift signal phase forwards or backwards in time, depending on the direction they are applied in), the filter was applied in both directions, using the MATLAB function, *filtfilt()*.

Independent component analyses (ICAs) were applied to copies of the data filtered to between .5 and 40 Hz (fourth-order Butterworth filter). The ICA was run using the FastICA algorithm (Hyvärinen, 1999) with a seed for reproducibility, using only data from within blocks (i.e., not during breaks). Here, blocks were defined as beginning one second before the first stimulus of each block, and ending one second after the block's last response. EOG-related ICA components were identified by first determining EOG epochs, defined as one-second windows centred on peaks in the EOG signal at least as extreme as one quarter of the difference between the minimum and maximum EOG amplitude. EOG-related ICA components were then identified as those showing a high correlation with these events, while also showing the typical frontal topography associated with EOG activity. For participants for whom EOG epochs could not be defined automatically, the time-course of all ICA components and EOG channels was compared manually.

Following EOG artefact removal, data were re-referenced with a common average reference. Trials with peak-to-peak amplitudes more than double the 99% confidence interval of the absolute amplitudes or greater were then excluded from analysis (N=620), such that there was a total of 13,276 trials. Some channels were identified as noisy (mean count per

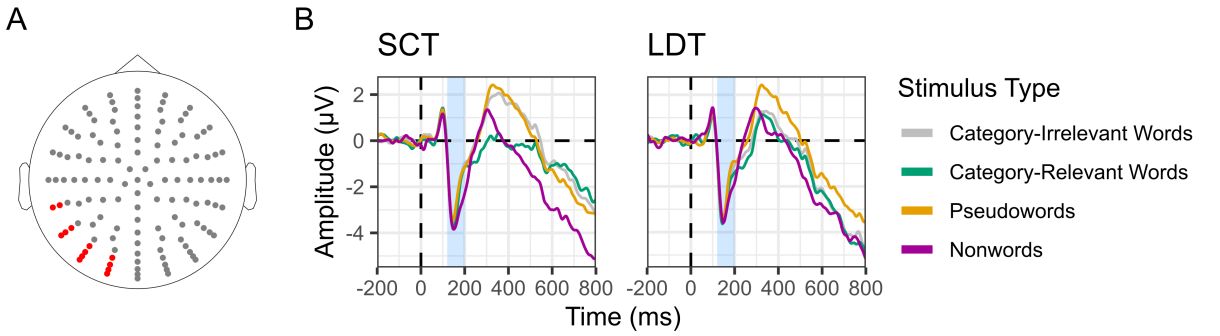


Figure 4.2: Occipitotemporal electrodes and average ERPs from maximal electrodes. (A) The locations of the 13 occipitotemporal electrodes (*red*) from which per-participant maximal electrodes were identified. (B) Average ERPs of the occipitotemporal electrodes for all stimulus types, across the SCT and LDT, with the 120-200 ms window of the N1 highlighted in blue.

participant=1.28, *SD*=2.67), and had their activity interpolated using spherical splines (Perrin et al., 1989).

4.3 Results

4.3.1 Occipitotemporal EEG Activity

To analyse effects of task, stimulus, and task-stimulus interactions on occipitotemporal EEG activity, I identified 13 left-lateralised occipitotemporal electrodes, selected to reflect the typical topography of the N1 (Figure 4.2). Averaging across conditions, the N1 component was observed to peak within this window at 146 ms. I analysed effects of task, stimulus, and task-stimulus interactions on the average amplitude of this occipitotemporal cluster. I first examined effects on trial-level average amplitude during the N1, between 120 and 200 ms. Second, I fit sample-level models, to examine the precise time-course of effects across the full ERP of the occipitotemporal cluster.

Effects on Average Amplitude

Trial-level average amplitude during the N1 was calculated as the mean of occipitotemporal electrodes' amplitudes in the window of the N1 (120-200 ms) for each trial. Distributions of average N1 amplitude for each factorial cell are summarised in Figure 4.3.

Average amplitude was modelled via a linear mixed effects model, fit with R package *lme4* (version 1.1.27.1; Bates et al., 2015). The model's formula estimated the fixed effects of task and stimulus, and interactions between them, while also estimating the maximal random effects structure justified by the design (Barr et al., 2013), which included per-participant, per-item, and per-match-set random intercepts and slopes. Here, match sets refer to the groups of 4 matched items (1 item per stimulus type) described in section 4.2.3. All fixed-effect variables were deviation-coded (i.e., mean-centred on 0 with a distance of 1 between the variable's levels). For the effect of stimulus, category-irrelevant words were used as the null condition, with all other

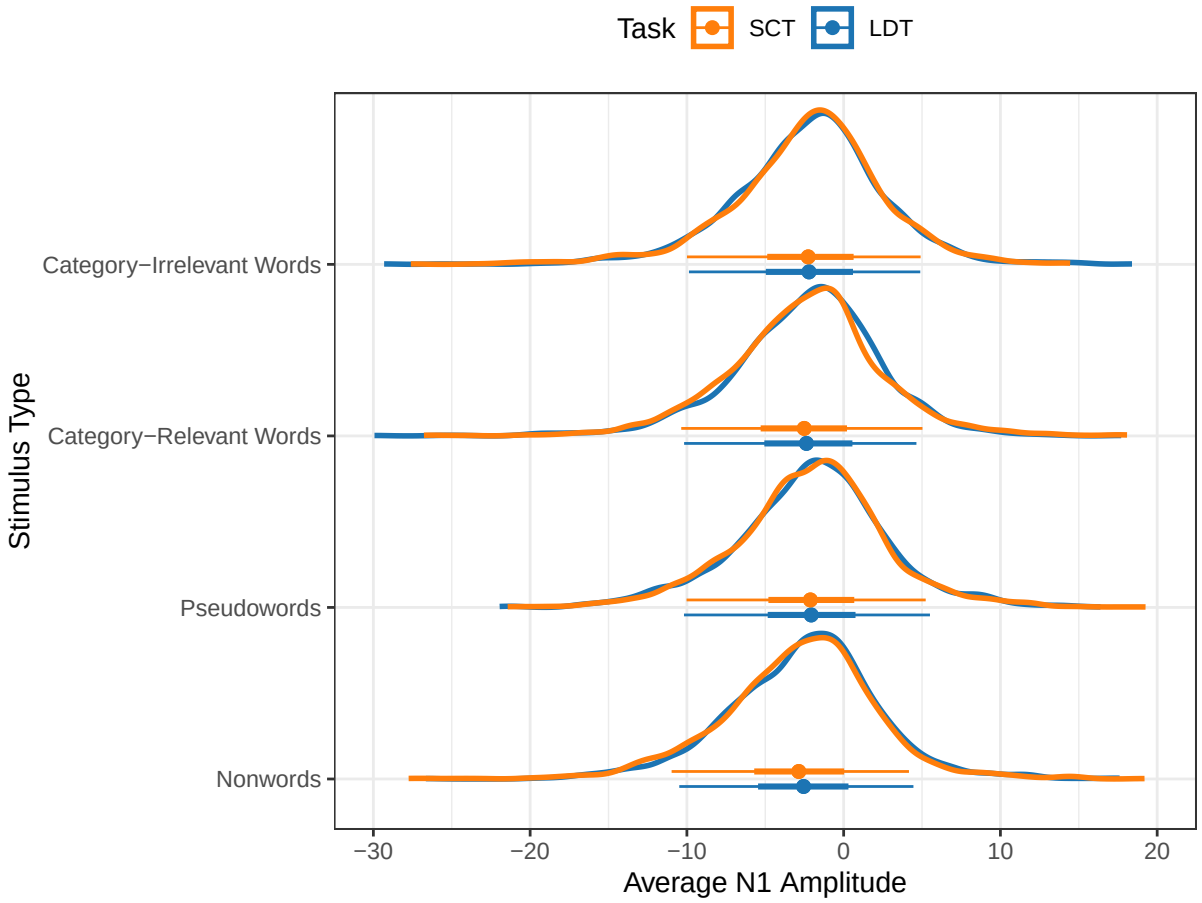


Figure 4.3: Distributions of average N1 amplitude for each factorial cell in the experimental design. Below empirical densities, points depict mean values, thick horizontal lines depict 50% quantile intervals, and thin horizontal lines depict 89% quantile intervals.

stimulus types deviation-coded relative to this condition. For the effect of task, LDT was coded as the null condition. The model formula, in *lme4* syntax, was as follows:

```
amplitude ~ 1 + (category_relevant + pseudoword + nonword) * task +
  (1 + (category_relevant + pseudoword + nonword) * task | participant) +
  (1 + (category_relevant + pseudoword + nonword) * task | match_set) +
  (1 + task | item)
```

The overall average of occipitotemporal amplitude during the N1 (i.e., the model intercept) was estimated to be β=-2.42 μV (*SE*=.4). All reported model estimates are in μV, and the statistical significance of fixed effects was tested via Chi-square model comparisons. The hypothesised effect, the interaction between task and category-relevance, was estimated to be small (β=-.07, *SE*=.26) and was not statistically significant (χ^2(1)=.66, *p*=.798).

Bonferroni corrections were applied to *p* values calculated for all model comparisons that were not related to the hypothesised effect, and these are reported as p_{bonf}, alongside unadjusted *p* values. Where corrected *p* values exceeded 1, this is reported as >.999. Average N1 amplitudes were more negative in the SCT than in the LDT (β=-.18, *SE*=.12, χ^2(1)=7.07, *p*=.008, p_{bonf}=.047).

The average N1 amplitude was more negative for nonwords than for for category-irrelevant words (β=-.57, *SE*=.16, χ^2(1)=11.47, *p*<.001, p_{bonf}=.004). The overall differences between category-irrelevant words and pseudowords (β=.07, *SE*=.16, χ^2(1)=.19, *p*=.661, p_{bonf}>.999), and between category-irrelevant words and category-relevant words (β=-.2, *SE*=.15, χ^2(1)=1.76, *p*=.184, p_{bonf}>.999), were small and not statistically significant.

Task-stimulus interactions for the difference between pseudowords and category-irrelevant words (β=-.15, *SE*=.33, χ^2(1)=.21, *p*=.648, p_{bonf}>.999), and between nonwords and category-irrelevant words (β=-.41, *SE*=.34, χ^2(1)=1.54, *p*=.214, p_{bonf}>.999), were not statistically significant.

Full N1 ERP Analysis

In addition to analysing the full N1 window, I anticipated that effects may be greater during the ERP's onset or offset. Indeed, sensitivity to lower-level differences has been reported to emerge during the early portion of the N1 (Appelbaum et al., 2009; Cohen et al., 2000; F. Wang & Maurer, 2020), while effects indicative of top-down modulation have been reported to be emerge during the later periods of the N1 (e.g., F. Wang & Maurer, 2017, 2020). To analyse the full ERP, I fit a linear mixed effects model to each sample, using a similar model formula to that described for the analyses of average N1 amplitude. A key difference, however, was that amplitudes were averaged across the occipitotemporal region, rather than using maximal electrodes. To deal with model non-convergence at multiple time points but ensure cross-sample comparability, no models estimated random correlations. The sample rate was unchanged for this analysis, kept at 512 Hz. Changes in the model fixed effects are summarised in Figure 4.4A.

The effect of interest, the interaction between task and category relevance and was only observed reliably from around 300 ms onwards (emerging at 225 ms at the earliest), with category-relevant words eliciting more negative amplitudes than category-irrelevant words in the SCT, but this effect being smaller or absent in the LDT. Notably, this effect emerged late into the N1's offset, but the difference was small and not observed reliably prior to 300 ms.

Results also replicated the N1's sensitivity to orthographic features, as revealed by the analysis of average N1 amplitude, captured by the difference between category-irrelevant words and consonant string nonwords, whereby nonwords elicited a more negative-going N1. This difference was greatest around 45 ms after the component had peaked, at 190 ms. Moreover, the direction of the orthography effect reversed after the N1, peaking again at around 260 ms, with nonwords here eliciting more positive ERP amplitudes than words did. The effect reversed again, peaking for a third time at around 350 ms, with nonwords again eliciting more negative ERP amplitudes than words did. This latter difference was sustained up to 600 ms after stimulus presentation, and possibly for longer. Sensitivity to differences between words and pseudowords emerged later, with pseudowords eliciting more positive ERP amplitudes from around 290 ms until 450 ms post-stimulus.

A main effect of category relevance was observed from around 225 ms. As task was deviation-coded in the model, this reflects the average effect of category relevance across tasks. To more clearly decompose the task-stimulus interaction, I fit models dummy-coded to focus on stimulus effects in the SCT and LDT respectively (i.e., simple effects; Figure 4.4B). This revealed that, indeed, the effect of category relevance was clearly observed in the SCT, whereas in the LDT it was much closer to zero. Smaller task-stimulus interactions were also observed for peudowords and nonwords. For both stimulus types, differences between these stimulus types and words were more negative in the SCT than in the LDT, for a sustained period beginning at around 200 ms. The effect of task on occipitotemporal ERPs is more clearly displayed in Figure 4.5, which shows the fixed-effect predictions for each factorial cell in the design.

The main effect of task, that was identified in the analysis of average N1 amplitudes, was less clear in this sample-level analysis. However, a small negative deflection in the effect of task was observed between 150 and 200 ms, suggesting that this small effect was inflated when average amplitude was calculated across the whole N1 window.

For comparison, I additionally analysed effects over right-hemispheric occipitotemporal electrodes (Appendix B.2), with results showing a larger word-nonword difference than was observed over the left hemisphere, but a similar lack of any interaction between task and category relevance in the N1. Notably, post-N1 interactions between task and category relevance were smaller in magnitude than estimates for left-hemispheric electrodes.

4.3.2 Scalp-Wide Analysis of the Time-Course for the Effect of Interest

Analyses of occipitotemporal EEG activity revealed that while the N1 showed robust sensitivity to orthographic features of stimuli, task-relevance interactions were smaller, and emerged later (post-225 ms). Anticipating that top-down modulation may be visible elsewhere in early EEG signals, and that the task-relevance interactions observed in occipitotemporal regions after the

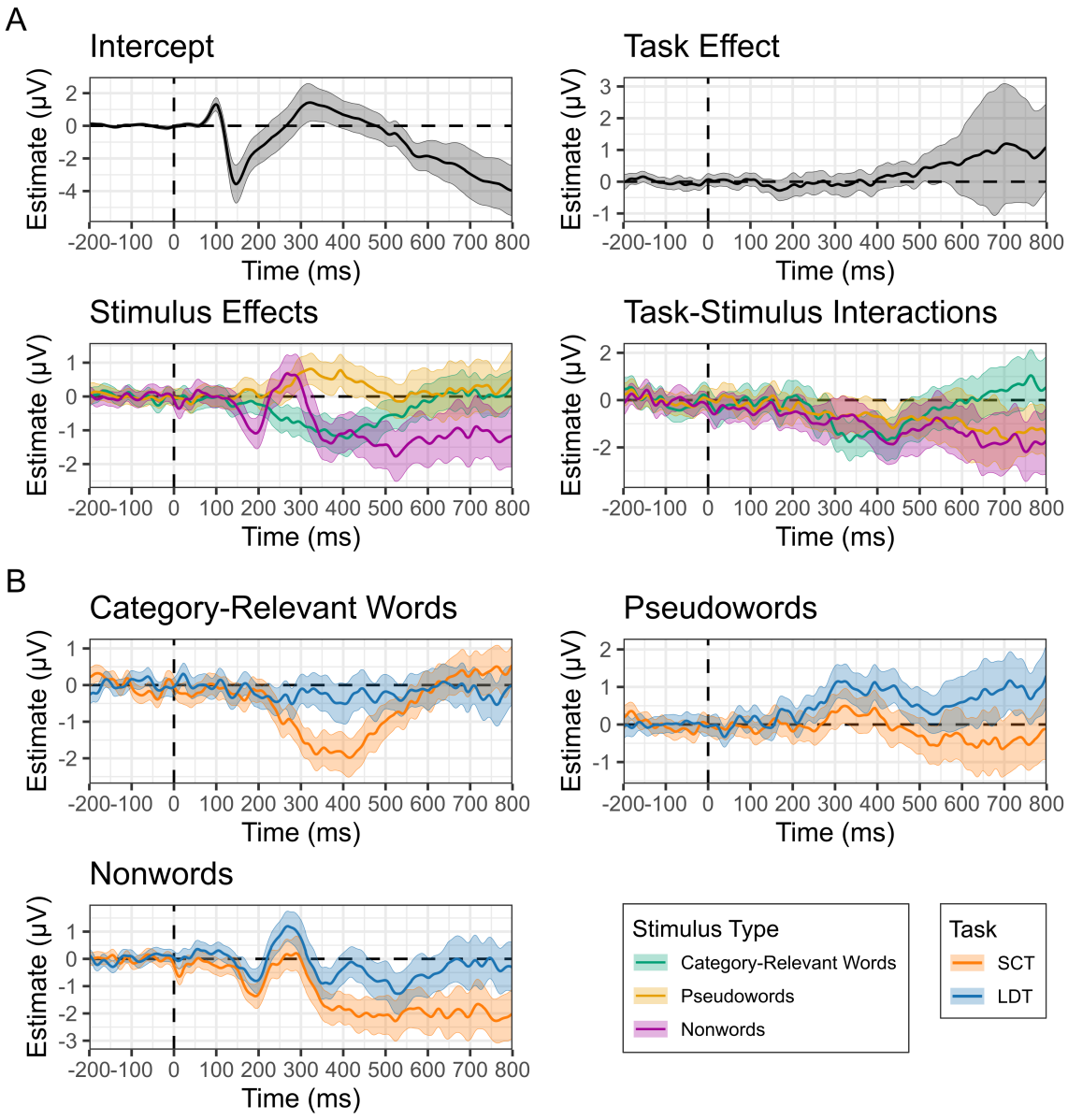


Figure 4.4: Time-course of fixed effects estimates from the per-sample linear mixed effects models of occipitotemporal electrode voltages. (A) Fixed effects estimates from a model with all variables deviation-coded. (B) Simple effects of task for each stimulus type. In both panels, solid lines depict estimates for each sample, while shaded regions depict 95% confidence intervals.

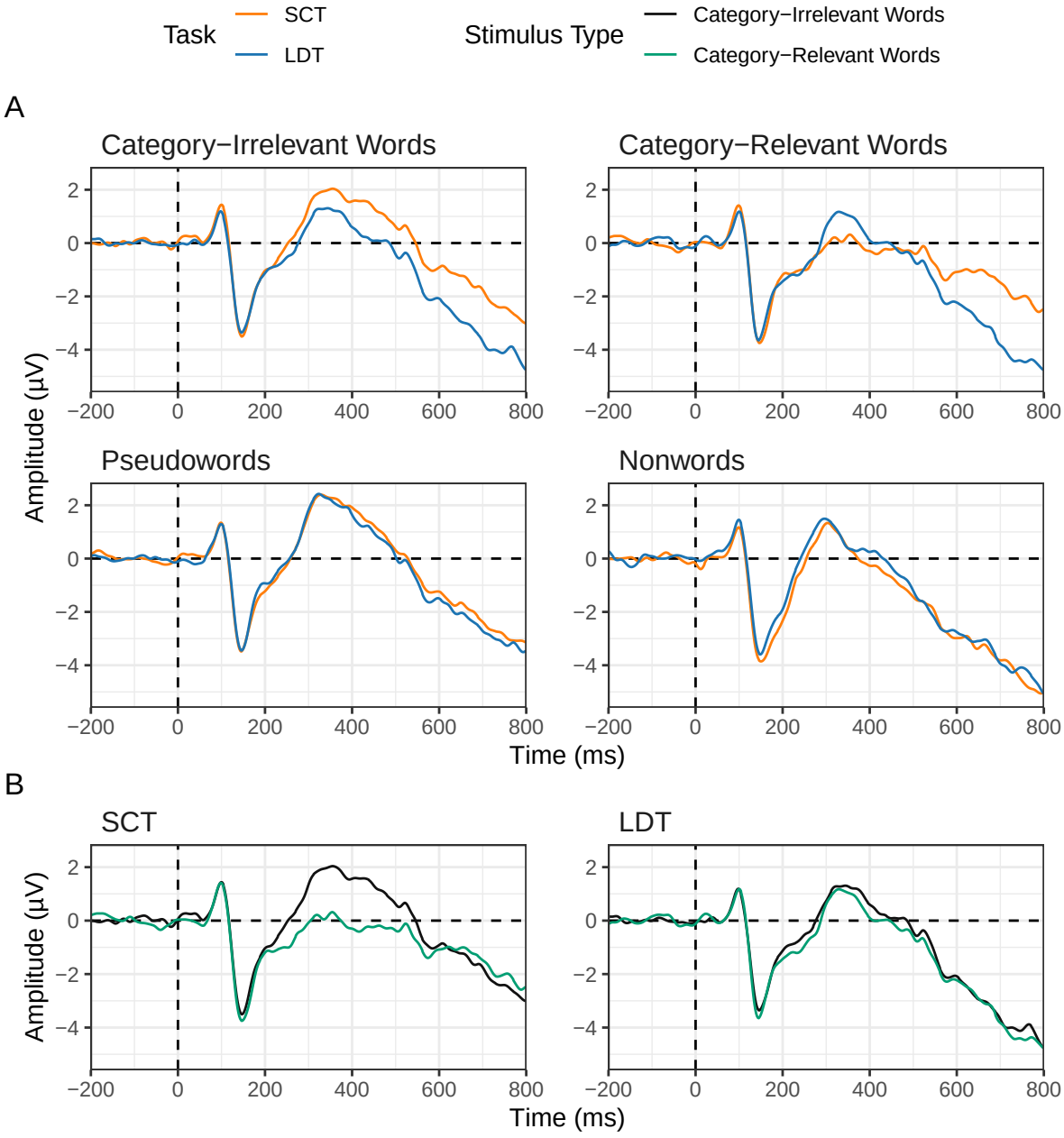


Figure 4.5: Fixed-effect predictions of ERPs for each factorial cell, using estimates depicted in Figure 4.4. These predictions are equivalent to overall average ERPs, but with the influence of random intercepts and slopes removed. Panels focus on (A) the effect of task for each stimulus type, and (B) the effect of category relevance in each task.

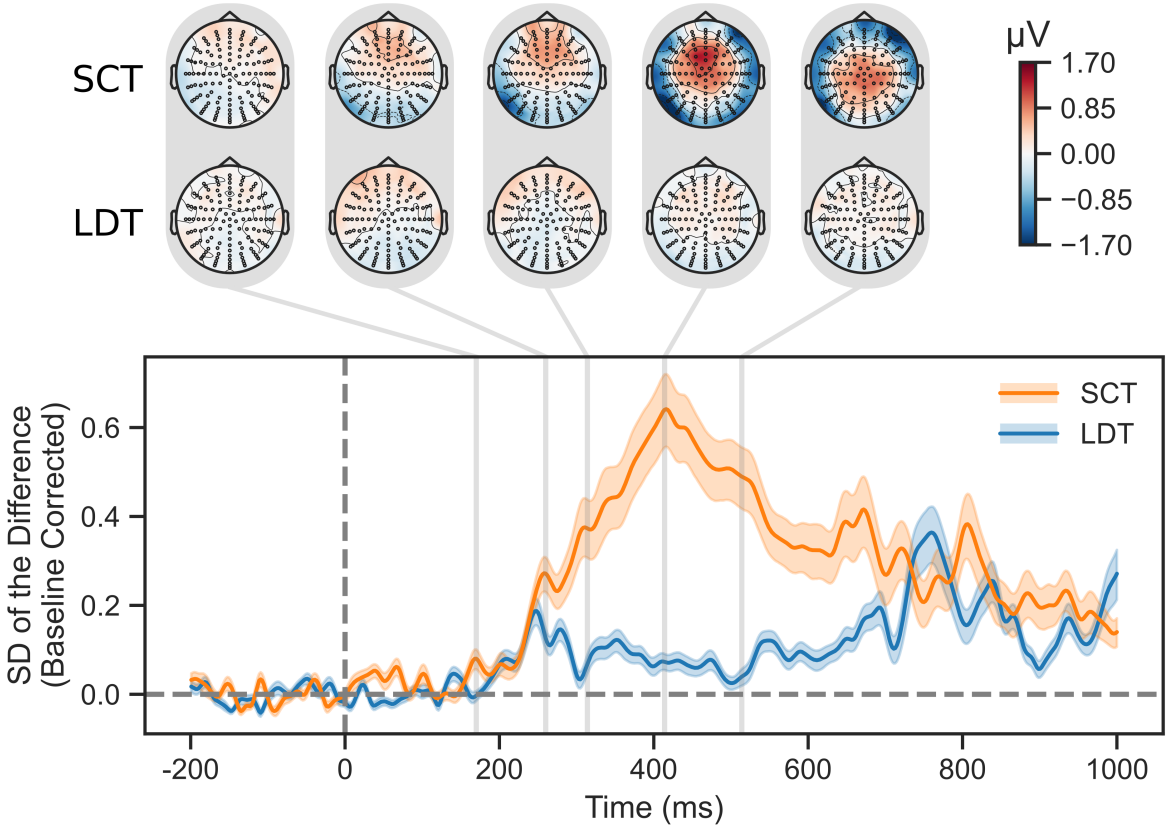


Figure 4.6: Time-course of task-relevance effects in the SCT and LDT. The lower panel depicts baseline-corrected estimates of global field power of the difference between average ERPs for category-relevant and category-irrelevant words, for the SCT and LDT separetly. Shaded intervals indicate 99% bootstrap confidence intervals (10,000 samples). The upper panel depicts the topography of category-relevance effects in the SCT and LDT at 170, 260, 314, 414, and 514 ms. ERP differences were calculated as category-relevant minus category-irrelevant.

N1 may reflect activity indexed by components that originate elsewhere, I analysed the scalp-wide time-course of ERP differences between category-relevant and category-irrelevant stimuli for both the SCT and LDT. Figure 4.6 shows the results of this analysis, which suggests that, consistent with the analysis of occipitotemporal channels, robust interactions between task and category-relevance emerged late into the N1's offset, after 225 ms. The timing (peaking at 414 ms) and centroparietal topography of the differences between conditions suggest sensitivity to this interaction in the N400, with category-irrelevant words eliciting more negative-going N400s than category-relevant words during the SCT but not during the LDT. To examine this centroparietal interaction in more detail, I conducted a sample-level analysis of a centroparietal cluster of electrodes (Appendix B.3). This revealed that, indeed, centroparietal electrodes showed a robust sensitivity to the interaction between task and category relevance, although the observed components did not conform to the typical timing of the N400, peaking instead at around 300 ms.

Topographies of the difference between category-relevant and -irrelevant words also reproduced the finding from the sample-level analysis that showed the difference in occipitotemporal electrodes that was observed after 225 ms (Figure 4.4B) to be larger for left

hemispheric electrodes than it was for right hemispheric electrodes, consistent with differences between left (section 4.3.1) and right-hemispheric (Appendix B.2) sample-level analyses.

4.3.3 Behavioural Results

I also examined effects of task and stimulus on the accuracy and speed of participants' responses. For comparability to the EEG results, I only analysed the behavioural results for participants who were included in the EEG analysis, although no trials were excluded on the basis of noise in the EEG (e.g., based on peak-to-peak amplitude). Similarly, unlike for the EEG results, trials were not excluded on the basis of response times. The only exclusion criterion applied was the removal of trials that were responded to inaccurately, and this was only applied to the analysis of response times.

Effects on Accuracy

As described above, all trials were included when analysing accuracy, for all 29 participants whose EEG results were analysed. This meant that accuracy was analysed in 14,384 trials (i.e., 29 x 496). Overall, empirical accuracy was very high, averaging 97.79%. To examine whether accuracy differed with the experimental manipulations, I fit a logit-link binomial generalised linear mixed effects model (GLMM) to predict accuracy via the R package, *brms* (version 2.16.3; Bürkner, 2018). Flat prior distributions were used for all fixed effects, while the priors for standard deviations (*SD*s) of all random effects were specified as *t* distributions with 3 degrees of freedom, where μ=0 and σ=2.5. The model was fit with 5 chains, and 10,000 iterations per chain (split equally between warm-up and sampling). The *adapt_delta* parameter was set to .8, and the *max_treedepth* parameter was set to 10. The *thin* parameter was set to 2, to reduce the size of the saved model. The same model formula was used as that described in the analysis of average N1 amplitude, with a maximal random effects structure. The model's fixed effect estimates are summarised in Figure 4.7. The model revealed clear effects of stimulus, and task-stimulus interactions. Notably, accuracy was higher for category-irrelevant words in the SCT than it was in the LDT, while the opposite was true for category-relevant words. Further, there was a small difference between category-irrelevant and category-relevant words in the LDT, despite careful matching of these conditions' features and cross-category randomisation of items in the LDT, although there was a high degree of overlap between the posterior distributions.

Effects on Response Time

To analyse response time, I opted to fit a Bayesian mixed-effects model for parameters of the shifted log-normal model. Whereas more traditional methods of modelling response times assume changes in one central tendency parameter of a distribution while other parameters are assumed to be homogenous across observations (e.g., Gaussian models of log-transformed RTs, or Gamma family models), modelling all parameters of multi-parameter distributions, which accurately describe the distribution of observations, allows researchers to assess changes in

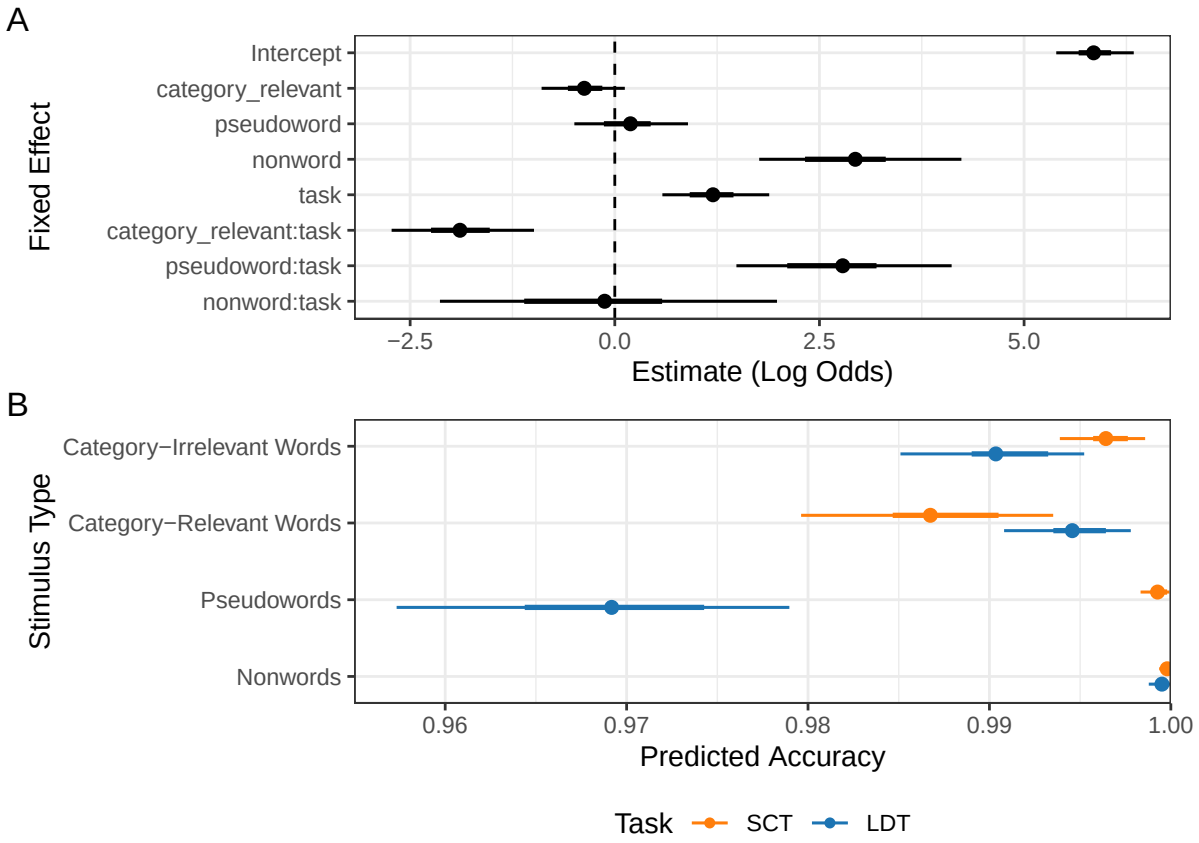


Figure 4.7: Summary of the logit-link binomial model of response accuracy. (A) Fixed effects estimates from the model, where labels on the y axis reflect the names of fixed effect parameters dictated by the model formula. (B) Predicted accuracies for each factorial cell in the experiment. In both panels, round central points depict median posterior estimates, while the thicker and thinner horizontal lines depict the 50% and 89% highest density intervals (HDIs), respectively.

the shape of entire response time distributions (Heathcote et al., 1991; Rouder et al., 2005). The shifted log-normal model was fit via *brms* (version 2.16.3; Bürkner, 2018), estimating fixed and random effects for each parameter of the distribution (i.e., μ, σ, and δ). In modelling the shifted log-normal distribution, μ, which reflects changes in means, was modelled with an identity link function. Meanwhile, σ and δ, respectively reflecting *SD*s (of log-transformed response times) and non-decision time, were modelled with log link functions. For feasibility, random slopes were removed from the model, and moderately informative priors were defined for the model's intercepts. The model formula, in *brms* syntax, was as follows, where *rt* refers to trial-level response times:

```
rt ~ 1 + (category_relevant + pseudoword + nonword) * task +
  (1 | participant) +
  (1 | match_set) +
  (1 | item),
sigma ~ 1 + (category_relevant + pseudoword + nonword) * task +
  (1 | participant) +
  (1 | match_set) +
  (1 | item),
ndt ~ 1 + (category_relevant + pseudoword + nonword) * task +
  (1 | participant) +
  (1 | match_set) +
  (1 | item)
```

Moderately informative priors were constructed for the model's intercepts, based on plausible ranges of shifted log-normal parameters in similar tasks. Specifically, the prior for the μ intercept was specified as $N(5, 2.5)$ (i.e. a normal distribution of mean 5 and *SD* 2.5), the prior for the σ intercept was specified as $N(0, 5)$, and the prior for the δ intercept was specified as $N(0, 7.5)$. All other fixed effects were assigned non-informative prior distributions in the form $N(0, 2.5)$. Priors for all random effects were specified as *t* distributions with 3 degrees of freedom, where μ=0 and σ=2. The model was fit with five chains, each with 10,000 iterations (7,500 warm-up, 2,500 sampling). The *adapt_delta* parameter was set to .99, and the *max_treedepth* was set to 10. To reduce model size, the *thin* parameter was set to 2.

The results of the response time model are summarised in Figure 4.8. Notably, while there were differences in the shifted log-normal parameter estimates for category-relevant and category-irrelevant words (Figure 4.8A), the resultant response time distributions were generally similar in both tasks (Figure 4.8C). However, the increase in response times observed for both types of word stimuli in the SCT, relative to the LDT, was slightly larger for category-irrelevant words than it was for category-relevant words (Figure 4.8B).

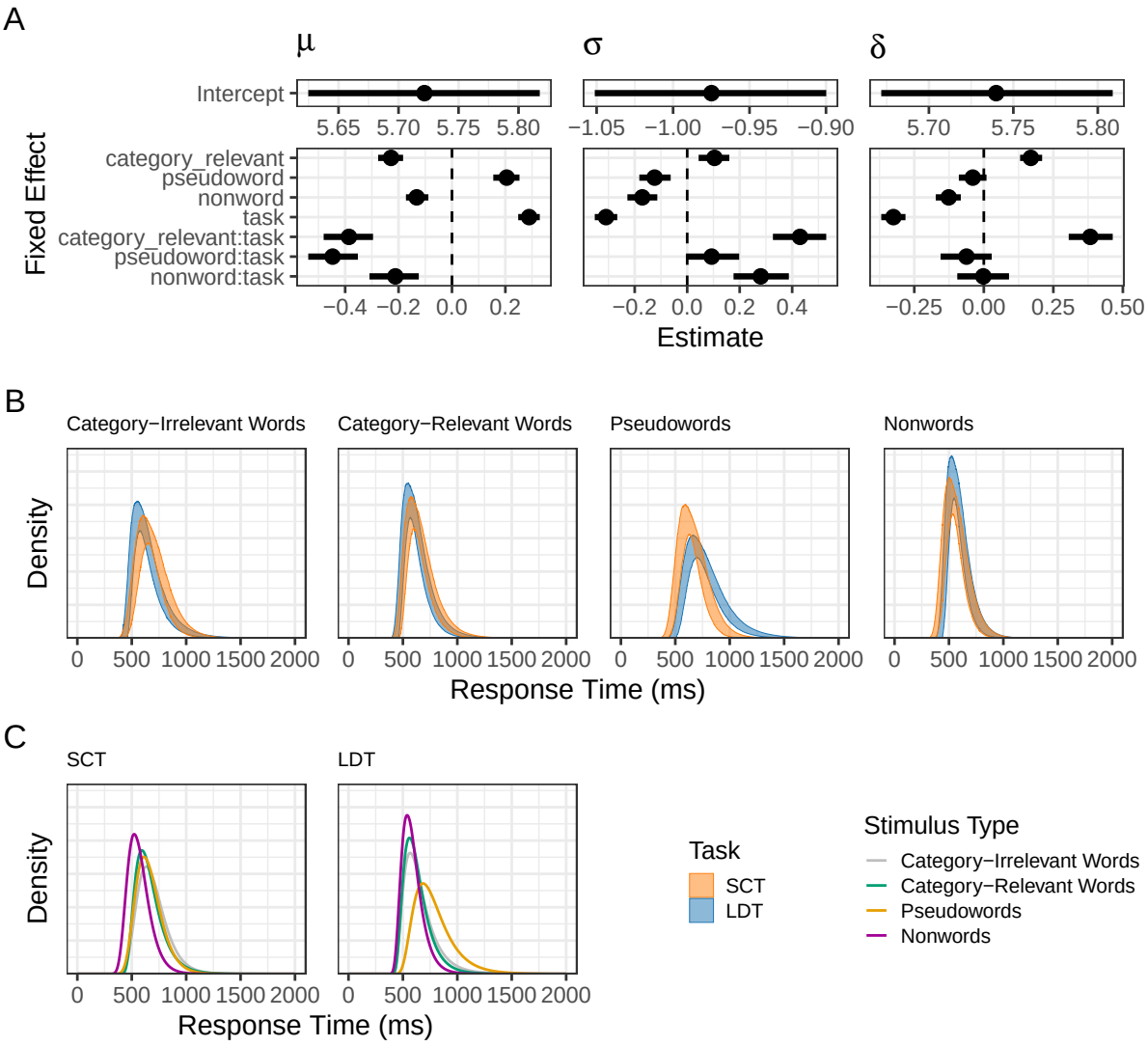


Figure 4.8: Summary of fixed-effect results from the shifted log-normal model of response times. (A) Estimates for each fixed effect, for each parameter of the shifted log-normal distribution, where fixed effect names are dictated by the model formula. Points represent median of posterior distributions, while horizontal lines depict 89% HDIs. (B) Predicted response time distributions for each factorial cell in the design, highlighting the effect of task for each stimulus type. Shaded regions depict 89% HDIs of density estimates for posterior samples. (C) Predicted response times for each factorial cell, highlighting stimulus differences for each task, where lines depict median estimates of densities. Axes for density, in (B) and (C), begin at zero, while densities are only plotted where median estimates for density are greater than zero.

4.4 Discussion

In this experiment, I examined whether the orthography-sensitive posterior N1 ERP component is sensitive to top-down modulation at the level of semantic category. Results showed that robust evidence for sensitivity to category relevance emerged later than the N1, starting at 225 ms at the earliest. From these results, it can be concluded that if early occipitotemporal orthographic processing is indeed sensitive to top-down modulation, the effect is likely small or absent at the semantic category level of word form prediction. Importantly, while evidence for top-down modulation of activity during the N1 was not observed here, the previously reported finding of *bottom-up* sensitivity to orthography was replicated.

4.4.1 Replication of Bottom-Up Sensitivity to Orthography

Previous research has demonstrated a sensitivity to orthographic processing during the N1. This sensitivity has been evidenced by differences between orthographically legal and illegal strings of letters (Bentin et al., 1999; Holcomb et al., 2002; Maurer, Brandeis, et al., 2005) that is dependent on the reader's knowledge of the presented orthography (Brem et al., 2018; Pleisch et al., 2019). Specifically, in both the LDT and SCT, consonant string nonwords elicited N1 components with more negative-going amplitudes. This difference was greatest in the later periods of the N1, peaking at a difference of around 1 μV at 190 ms. Importantly, this clear difference was not observed for pseudowords, which were designed to be orthographically (and phonologically) plausible. While this finding replicates a lack of word-pseudoword sensitivity that has previously been reported (Holcomb et al., 2002; Maurer, Brem, et al., 2005), it stands in contrast to some previous reports of sensitivity to the word-pseudoword difference during or close to the N1 (Eberhard-Moscicka et al., 2016; Hauk et al., 2006; Segalowitz & Zheng, 2009). However, such variability in findings could be partially attributed to differences in features of pseudoword stimuli. While some procedural techniques exist for pseudoword generation (e.g., Duyck et al., 2004; Keuleers & Brysbaert, 2010), evaluation of pseudowords' word-likeness is difficult to formalise, such that researchers must often rely on subjective perceptions of word-likeness. For the stimuli used in this task, I confirmed that the manually generated pseudowords were orthographically word-like by examining the distributions of orthographic variables: character bigram probability, and OLD20. While the selected pseudowords showed similar distributions to those observed for real words on both of these variables, nonwords differed markedly from real words, showing lower bigram probabilities and larger OLD20 values (indicating less orthographic similarity to real words). As a result, I suggest that it is the high orthographic plausibility of the pseudowords used in this experiment that resulted in the high similarity to N1 components observed for real words. In contrast, I suggest that sensitivity to nonwords was observed in the N1 because of the very low orthographic plausibility of consonant strings.

Sensitivity to nonwords-versus-words, and not pseudowords-versus-words, also provides evidence that the N1 is likely sensitive to orthographic information that is more fine-grained (and low-level) than representations of entire word forms. This is because pseudowords,

while orthographically plausible, were unlikely to have ever been observed by the participants prior to the experiment. Were participants processing the entire observed word form as a whole only, the difference between words and pseudowords would be similar to that observed between words and nonwords. The lack of pseudoword sensitivity observed in this experiment is difficult to reconcile with suggestions that some degree of lexical access, or whole-word processing, occurs during or by the N1, as has been claimed on the basis of previously reported word-pseudoword differences (e.g., Eberhard-Moscicka et al., 2016) and sensitivity to word frequency (e.g., Assadollahi & Pulvermüller, 2003; Hauk & Pulvermüller, 2004; Sereno et al., 1998; Simon et al., 2007) in the N1. Indeed, as with the word-pseudoword difference, sensitivity to word frequency in the N1 could result from differences in orthographic features that necessarily covary with word frequency. For instance, both average character bigram probability (Figure 4.9A) and OLD20 (Figure 4.9B) show heterogenous relationships with word frequency, such that high frequency words are more likely to have higher bigram probabilities and larger orthographic neighbourhoods. To consider the relationship between word frequency and words' orthographic features in another way, highly frequent words will, by definition, be perceived more often. Each time a word is observed, its component *N*-grams will also be observed. As a result, the orthographic components of high-frequency words will also be observed more frequently, and sensitivity to word frequency could emerge via sensitivity to the frequency of these orthographic components without recognition of the word form they constitute (i.e., without necessitating lexical access). Of course, character *N*-grams in high-frequency words will also occur in low-frequency words (Figure 4.9A), though the correlation between *N*-gram frequency and word frequency will necessarily become stronger when *N* is higher (e.g., trigrams, quadrigrams, etc.). Furthermore, there is no reason to believe that orthographic processing utilises features like *N*-grams alone. Indeed, it is likely that other featural information informs orthographic processing, and that configural information may also play a role (e.g., characters' or character *N*-grams' relative positions within words; Davis, 2010; Gomez et al., 2008; Grainger & van Heuven, 2004), such that sub-lexical orthographic configurations could occur within words with quite high consistency. As a result, if such orthographic features are not carefully controlled, it is feasible that sensitivity to word frequency could arise from processing that is exclusively orthographic and sub-lexical.

One interesting finding was that, relative to the N1 component's peak, sensitivity to orthography emerged early but peaked late: sensitivity to orthography seemed to emerge close to the N1 peak (146 ms), but peaked later, with sensitivity highest at 190 ms. This finding concords with other research highlighting heterogeneity within the window of the N1, with bottom-up sensitivity to orthography emerging during the component's onset or peak, but often peaking later, during the component's offset (Appelbaum et al., 2009; Cohen et al., 2000; Ling et al., 2019; F. Wang & Maurer, 2020). It has been suggested that top-down effects, meanwhile, begin to influence the N1 during later periods, with effects only emerging in the component's offset (e.g., F. Wang & Maurer, 2017, 2020).

To summarise the interpretation of stimulus effects, this experiment demonstrated a robust bottom-up sensitivity to orthographic information (words-nonwords), but not to lexicality (words-pseudowords), in the later portion of the N1 component.

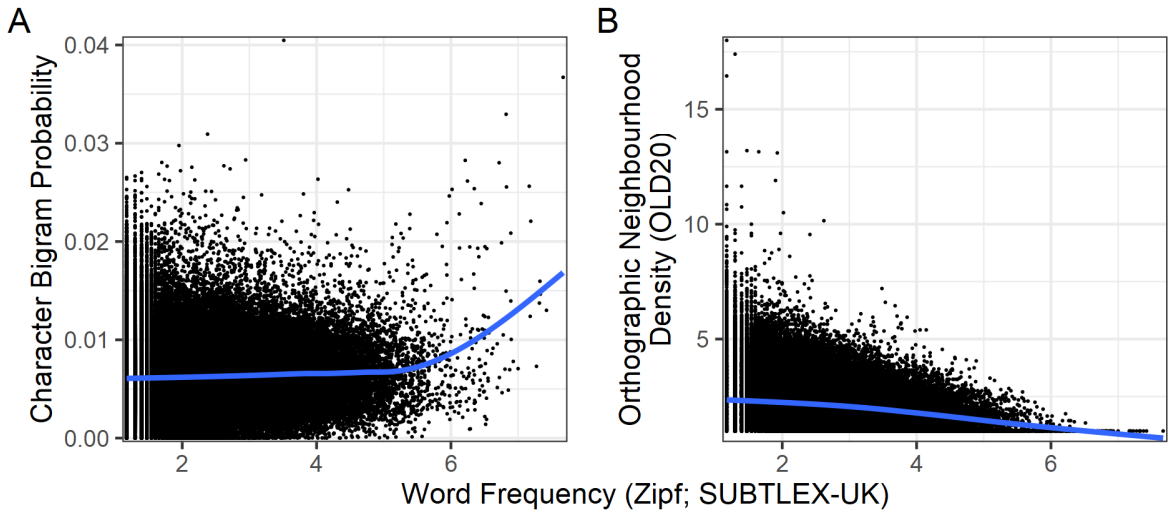


Figure 4.9: The relationships between word frequency (Zipf) in SUBTLEX-UK (van Heuven et al., 2014) and average character bigram probabilities (calculated from SUBTLEX-UK), and OLD20 values (calculated from all words in the LexOPS dataset). The blue line reflects an estimate of the relationship, fit with a cubic spline, to depict changes in central tendency over word frequency.)

4.4.2 Lack of Sensitivity to Category-Level Top-Down Modulation

The hypothesised effect, an interaction between task and category relevance in the N1, was not observed until late into the N1's offset. Indeed, further analyses suggested that this sensitivity to task-dependent category relevance reflected centroparietal effects that peaked during the N400. As a result, it is likely that top-down modulation of the N1, in response to expectations formed at the level of semantic categories, either produces small effects or is entirely absent.

The lack of sensitivity to category relevance, in either LDT or SCT, represents a failure to replicate the finding reported by Hauk et al. (2012). One possible explanation for this discrepancy is the qualitatively different categories employed. The categories of living and non-living objects are semantically broader than the categories employed here (e.g., *Four-Footed Animals* are a subset of living objects). However, as outlined below, if the sensitivity to categories relies on orthographic information, then the more targeted categories employed in this study should be expected to show a *larger* effect, as the mapping from semantics to category should have greater specificity. An alternative explanation for the differences between the results from Hauk et al. and those of this study is that the former study's results do not reflect a *task-dependent* sensitivity to the living-versus-non-living categories of words, but either a bottom-up sensitivity to the categories (e.g., words for living objects may be more familiar) or to other category-irrelevant features that differed somewhat between the categories presented by Hauk et al., such as orthographic neighbourhood density or word frequency, for which effects have sometimes been reported in the N1 (Assadollahi & Pulvermüller, 2003; Hauk & Pulvermüller, 2004; Sereno et al., 1998, see also chapter 6). Finally, it is possible that the high-pass filter of 1 Hz employed by Hauk et al. produced artefactually early effects in the ERP, that in fact originated in later components, by distorting the timing of effects (Tanner et al.,

2015; VanRullen, 2011). In contrast, low high-pass, gentle-slope filters like the fourth-order Butterworth filter employed in the present study do not suffer from the same distortion effects introduced by standard finite-impulse response filters (Rousselet, 2012).

Importantly, however, a lack of top-down influence at the level of broad categories does not discount the possibility of top-down modulation in response to more targeted predictions. Indeed, if activity during the N1 reflects orthographic processing, an influence of top-down modulation from the level of semantics may *require* predictions to be more targeted. This is because recoding information from semantic to orthographic representations is likely to be computationally lossy. Here, lossiness is a concept borrowed from information technology, where transcoding of information between digital formats, or digital compression of data, entail a loss of information. It is argued that transcoding of information between high-level, specific semantic representations, and low-level orthographic representations, results in a similar loss of information, which in turn causes predictions to be less specific. To explain this concept in another manner, semantic and orthographic representations have a many-to-many relationship, where a semantic concept is related to or expressed via many words, which themselves contain many orthographic features. These orthographic features occur in many other words, which similarly each refer to many semantic concepts. Recoding from lexical or semantic representations to sublexical orthographic representations therefore entails a loss of specificity, as these features will also occur in unrelated words. Indeed, in orthographies constructed from alphabetic scripts like English, it is on the basis of strings formed from around just 30 graphemes, which additionally often show high similarity to one another (chapter 6), that readers are able to discriminate between tens of thousands of word forms (Brysbaert et al., 2016). As a result, orthographic representation of predicted word forms, e.g., *brother*, may also lead to facilitation for semantically unrelated words if their word forms share orthographic features, e.g., *bother*, *smother*, *betroth*. Top-down facilitation for predictions of entire semantic categories of orthographically diverse word forms (e.g., Relatives: *brother*, *niece*, *nephew*, *aunt*), where each word would itself confer facilitation for word forms orthographically similar but semantically unrelated to it, could therefore be expected to only produce very small practical effects, if any, on orthographic processing or its neural correlates.

Evidence for predictions leading to such inadvertent facilitation of unpredicted but orthographically similar word forms in the N1 has been reported previously. A. Kim and Lai (2012) showed that sentence stems designed to induce strong predictions of upcoming words (e.g., *She measured the flour so she could bake a-*) led to similar reductions in N1 amplitude (175-205 ms) for the predictable word form (e.g., *cake*) as well as an unpredictable but semantically similar pseudoword (e.g., *ceke*), but not for an orthographically dissimilar pseudoword (e.g., *tont*). This finding suggests that, indeed, facilitation for sublexical orthographic features of word forms can entail a loss of specificity in predictions. Furthermore, the suggestion that this loss of specificity is engendered by the orthographic similarities of items within a script is supported by the previously mentioned finding of sensitivity to categories of orthographic script (F. Wang & Maurer, 2020). Functionally meaningful sensitivity to expectations of the category of script (Chinese vs. Korean) that upcoming characters belong to could be made possible by the intra-script similarity, and inter-script dissimilarity, of

these two orthographic categories (in addition to differences in familiarity). A further prediction of this interpretation of the present study's results would be that if top-down modulation of the N1 does occur in response to semantic predictions, then effects in the N1 of predictions for semantic categories of word forms may be more visible for some non-alphabetic scripts. In particular, logographic scripts, whose word forms are more diverse and numerous, may support predictions at the level of the semantic category. Another factor related to top-down modulation of categories of word forms that requires semantic-to-orthographic recoding may be intra-category orthographic similarity, and this feature is also likely to systematically differ by script. For instance, in contrast to word forms in alphabetic orthographies, Chinese characters for items in a semantic category are often orthographically similar, sharing orthographic components that reflect their shared semantics. For instance, in simplified Chinese, many items in the category of Four-Footed Animals share a variant of the Kangxi radical 犬, specifically 犭, at the left of the characters, for instance: 狗 (*dog*); 猫 (*cat*); 猪 (*pig*); 狮 (*lion*); 狼 (*wolf*)). Indeed, although a direct comparison to non-logographic scripts is lacking, fMRI findings reported by X. Wang et al. (2018) do indicate that for Chinese characters, semantic information (thematic/taxonomic) can be decoded from visual word form area activity in a task-dependent manner (thematic/taxonomic categorisation tasks). Further research could more directly compare such findings to alphabetic scripts to examine whether this finding is indeed script-dependent, and could examine use EEG or MEG to determine whether semantics influence initial occipitotemporal orthographic processing. If script-dependent, research could further delineate the influence of semantics on orthographic processing by examining how categoric-orthographic typicality interacts with the decodability of semantics in vOT or from the N1. For instance, characters whose orthography is less typical of a semantic category, sharing fewer orthographic features with characters they share the category with (e.g., 羊 (*donkey*) is orthographically dissimilar from other characters in the category of Four-Footed Animals) may show a reduced effect of category-level predictions on orthographic processing.

The lossiness of transcoding information from semantic to orthographic representations may explain why much, if not most, published research on early effects of prediction in alphabetic scripts has utilised designs that aim to elicit more targeted predictions (Nieuwland, 2019). As a result, experimental designs that lead participants to form stronger predictions about upcoming word forms (e.g., Dambacher et al., 2012; Dikker & Pylkkanen, 2011; Kretzschmar et al., 2015; Penolazzi et al., 2007; Sereno et al., 2003) may be vital for examining whether the N1 is affected by top-down modulation on the basis of higher-level predictions. Alternatively, considering evidence that the N1's likely neural generator is capable of representing multiple word forms simultaneously (White et al., 2019), neural activity during the N1 may be similarly capable of maintaining *predictions* for specific word forms in parallel. However, even in such a case, this would require the participant to actively predict all members of a category, rather than evaluate their category relevance post-hoc, which may be unlikely for such large categories as Musical Instruments.

Finally, it is notable that interactions between category relevance and task were observed in occipitotemporal amplitudes *after* the N1 (Figure 4.4). The strong left-lateralisation of this effect (Figure 4.6), in contrast to the largely symmetrical topography of effects on the N400

component, may suggest that this reflects meaningful late occipitotemporal activity related to category membership. While possible interpretations include a sustainment of occipitotemporal activation that begins during the N1, or a reactivation of orthographic information driven by feedbackward connections, a specific mechanistic interpretation of this finding is not possible here.

To summarise, the hypothesised category relevance-task interaction was not observed in the N1 until late into its offset. In retrospect, a finding of category-level sensitivity would be a particularly surprising result, given the lossiness inherent to the recoding of information from semantic to orthographic representations. Nevertheless, this study does provide some useful insight into top-down modulation of the N1. Namely, if top-down modulation of the N1 does occur, it probably requires higher-level predictions of words, and hence word forms, that are more specific than those that can be formed at the level of the semantic category.

4.4.3 Main Effect of Task

Previous studies utilising task manipulations like that employed here sometimes report a main effect of task, where N1 amplitudes are more negative going, across different stimulus types, during tasks that more explicitly require lexical or semantic processing, relative to more perceptual tasks (Y. Chen et al., 2013; Segalowitz & Zheng, 2009). In one recent M/EEG investigation, Rahimi et al. (2022) reported general effects of task, between an LDT and SCT, as early as 60 to 65 ms, in the primary visual cortex (V1), with clear effects during a time window that includes the N1 (150-250 ms). The present study also found a main effect of task on the N1, and this was in the same direction, with more negative-going average N1 amplitudes observed in the SCT than in the LDT. Notably, this effect was quite small (.18 μV), relative to the sensitivity to the word-nonword difference, for example, though the full-ERP analysis of occipitotemporal electrodes suggested that if there is a main effect of task, it did indeed coincide with the N1.

4.4.4 Lack of Sensitivity to Top-Down Modulation of Lexical or Orthographic Processing

Previous results have suggested that processes indexed by the N1 that are sensitive to the difference between words and nonwords, or between words and orthographically unfamiliar stimuli, may be influenced by task demands. Specifically, some evidence suggests that sensitivity to this difference is greater during tasks that more explicitly require lexical and semantic processing (Bentin et al., 1999; F. Wang & Maurer, 2017). Indeed, the category-level effect reported by F. Wang and Maurer (2020) is likely to reflect an effect of top-down modulation on general orthographic or lexical processing, given that the experiment compared responses to two orthographic scripts that were respectively familiar (Chinese) or unfamiliar (Korean) to participants, comparable to the word-nonword differences observed for alphabetic scripts. Evidence also suggests that the N1 is more sensitive to word frequency (or orthographic features that covary with word frequency) in tasks requiring more explicit lexical or semantic processing (Y. Chen et al., 2015; Strijkers et al., 2015). I have suggested in this chapter that,

if not a sensitivity to semantic category relevance, the task-dependent effect reported by Segalowitz and Zheng (2009) could result from such effects. Specifically, the greater semantic processing precipitated by the LDT where words were drawn from a common category may have induced deeper semantic processing of words in this condition. Indeed, as previously noted, the absence of category-irrelevant words, as a lexical control, means that it could reflect a task interaction with category relevance, or lexicality, or both.

However, in addition to a lack of sensitivity to lexicality, and the absence of task interactions with category relevance prior to 200 ms, no clear interaction between task and either lexicality (words vs. pseudowords) or orthographic processing (words vs. nonwords) was observed in the present study. As a result, the present study did not replicate the previous findings listed above, although it is notable that none of those studies used the specific tasks employed in the present study.

4.4.5 Possible Limitations

It should be noted that the lossiness of semantic-to-orthographic recoding is not the only possible explanation for this study's lack of evidence for a task interaction with category relevance in the N1. Another possible explanation that could account for this is that participants may not have actively predicted words in the SCT, instead only evaluating category membership after a word has been observed. This is to say, the task may not have sufficiently biased expectations to lead participants to form the predictions necessary for top-down modulation to occur. Indeed, the finding of relatively late sensitivity to category relevance in the SCT, peaking around 400 ms, could reflect largely bottom-up lexical and semantic processing. One analysis that could be used to evaluate this interpretation would examine the extent to which observed sensitivity to category relevance depends on word predictability. For instance, in the category of *Fruit*, category members such as *apple* and *banana* may be more predictable than *cantaloupe* or *papaya*. If the sensitivity to category relevance that emerges in the SCT from 225 ms only emerges for highly predictable items, this would be more suggestive of top-down contributions to such processes, rather than bottom-up evaluation of category membership. This is difficult to evaluate for the data collected from this experiment, however, as although (Van Overschelde et al., 2004) did report the proportion of participants who provided each item as a category member, most items were not highly predictable by this measure. For example, while *dog* and *cat* were highly predictable members of the *Four-Footed Animal* category, respectively provided by 98% and 97% of participants, the average predictability of items in this category was only 23.96%, and over half of items had predictabilities below 20%. Therefore, in addition to designing paradigms which lead to predictions for specific word forms, future investigations could improve on the present study by formalising word predictability, and ensuring that a sufficient number of items are highly predictable to estimate the relationship between predictability and top-down modulation.

Another possible limitation of the present study is that the behavioural analysis suggested that some small differences may have been observed between category-irrelevant and category-relevant words in the LDT. This could be problematic, as it would indicate that either

the stimuli were not suitably well-matched enough, or that participants may have noticed the shared semantic categories of category-relevant stimuli in the LDT, even though items were shuffled across blocks to avoid this. However,in addition to the small magnitude of the difference in accuracy (less than .5% difference), there was considerable overlap between the posterior distributions, such that comparability of responses was still highly plausible. Similarly, although the shifted log-normal parameter estimates for response time distributions differed between category-irrelevant and category-relevant stimuli in the LDT (*category_relevant:task* interaction in Figure 4.8), the resultant response time distributions were very similar. Finally, while there were potentially interesting behavioural results across the two tasks, it should be noted that interpreting data about responses like accuracies and response times is made difficult by the change in responses necessary for category-irrelevant stimuli across the tasks. In the SCT, the correct response for category-irrelevant stimuli was the *negative* response (i.e., "no"), while the correct response for category-relevant stimuli was *affirmative* (i.e., "yes"). In the LDT, however, the correct response for both types of word stimuli was *affirmative*. Indeed, this difference in the frequency of responses, from .5 affirmative to .25 affirmative, is also likely to have affected responses to pseudowords and nonwords, further complicating the interpretation of behavioural data.

4.4.6 Conclusions

In this study, I replicated previous findings of bottom-up sensitivity to orthography, and found evidence suggesting that the N1 is not sensitive to lexicality. Importantly, I showed that if higher-level predictions affect orthographic processing in the N1 via top-down modulation, the effect is likely small or absent at the level of semantic categories when predictions are broad. In this way, this study provides an upper bound for the influence of top-down modulation. If top-down modulation does functionally influence the N1, it is likely to occur when predictions have greater specificity than multiple, orthographically diverse members of a semantic category. Such an influence may also be expected to vary with predictability, affecting processing the most when predictability is highest.

Chapter 5

The Effect of Predictability on Top-Down Modulation of the N1

5.1 Introduction

In chapter 4, I showed that the N1 event-related potential (ERP) component, observed in electroencephalography (EEG) signals around 170 ms after word presentation, is sensitive to orthographic features of word forms in a bottom-up manner. I found no evidence for top-down modulation of the orthographic processing of words at the level of semantic categories. One possible reason for this lack of top-down modulation could be that predictions for orthographic features of entire semantic categories may be too non-specific to effect any meaningful change in orthographic processing. In this chapter, I therefore examine whether top-down modulation of orthographic processing is observed in a task designed to bias predictions towards more specific word forms: a Picture-Word Verification Task. I additionally vary the predictability of the picture-word relationship to examine, if top-down modulation occurs, whether this effect interacts with predictability. In particular, I expected that the effect of picture-word congruency would be largest when predictability is high.

The N1 component shows robust sensitivity to orthographic features of visual word forms (Bentin et al., 1999; Brem et al., 2018; Gagl et al., 2020; Holcomb et al., 2002; Ling et al., 2019; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019, see also chapter 4). However, much less is known about the extent to which processing during the N1 is sensitive to top-down modulation. While research suggests there may be a general effect of task demands on the N1 (Y. Chen et al., 2013), and that task demands may interact with the N1's sensitivity to stimuli's orthographic features (Bentin et al., 1999; F. Wang & Maurer, 2017, 2020) or lexical features that covary with orthography (Y. Chen et al., 2015; Strijkers et al., 2015), less is known about effects of prediction for specific word forms. Although some evidence suggests that targeted predictions for entire categories of word forms may affect the N1 (Hauk et al., 2012; Segalowitz & Zheng, 2009), I did not observe such an effect in chapter 4. Further, as argued in chapter 4, recoding information from semantic to orthographic representations would be an inherently lossy operation - especially for alphabetic scripts that rely on the reuse of a small number of orthographic characters. Indeed, if orthographic processing during the N1 is sensitive to top-down modulation, effects of such influences may only be observed when predictability is high

enough to preserve specificity in the mapping from semantics to orthography.

Rather than investigating the effects of prediction at the level of entire categories, researchers more commonly design paradigms to induce participants to construct more specific predictions, for individual word forms. For example, researchers design sentences with varying cloze probabilities for target words, to examine the effect of such predictions on ERP components including the N1. Published findings using paradigms with such sentential manipulations have broadly suggested an effect of predictability on the N1, with findings generally revealing that less negative-going N1 components are elicited by words that are predicted. Such a finding is consistent with a predictive coding account of orthographic processing (Gagl et al., 2020) that involves top-down modulation (e.g., Price & Devlin, 2011), according to which predicted features of a word form would be "explained away" (A. Clark, 2013; Eisenhauer et al., 2022; Gagl et al., 2020), while unpredicted features would elicit a component of more extreme amplitude, reflective of the relatively higher orthographic prediction error (the difference between the predicted and observed orthographic representations; Gagl et al., 2020). Nevertheless, key features of the effect of predictability on early visual word processing, like its reported topography and timing, differ considerably across studies, as does the oft investigated interaction between predictability and word frequency (see section 1.5.2 for a summary). It is similarly notable that in a review of studies using such sentential paradigms to bias predictions towards specific word forms, examining effects on multiple components including the N1, Nieuwland (2019) concluded that reported effects had thus far been weak, inconsistent, and in need of replication.

While sentential approaches are useful for forming strongly biased expectations with naturalistic stimuli, they can also introduce several issues, such as that ERPs elicited by the target word can become difficult to disentangle from ERPs elicited by preceding words (as has been argued for the early left anterior negativity, i.e., ELAN: Steinhauer & Drury, 2012), especially if the delay is short or unjittered. Further, it has long been recognised that paradigms using linguistic stimuli to bias expectations for word forms must be cautious of direct semantic associates priming target word forms intra-lexically, to differentiate between modular and interactive accounts of word processing (Fodor, 1983; Forster, 1979). This complication also applies to investigations of top-down modulation of orthographic processing. For instance, priming the word *chips* with the words *fish and* may not necessarily require active top-down modulation, with input from regions that process semantic information, if the word forms frequently co-occur. Rather, an orthographic processing module could, upon encountering orthographic features of the words *fish and*, learn to facilitate processing of orthographic features in the word form *chips* via local non-feed-forward connections, without such connections necessitating top-down modulation (Barlow, 1997; Rauss & Pourtois, 2013).

An alternative approach utilises non-linguistic contexts to bias participants' predictions. For example, researchers can alter task demands while presenting identical or highly similar target stimuli. In chapter 4, I applied such a paradigm, finding that top-down modulation of orthographic processing probably doesn't occur at the level of semantic categories. Some evidence suggests that top-down modulation may produce a main effect of task (Y. Chen et al., 2013; Segalowitz & Zheng, 2009), or that sensitivity to orthography or lexicality may interact

with task (Bentin et al., 1999; F. Wang & Maurer, 2017), though these differences were not observed between the tasks I employed, of lexical decision and semantic categorisation.

In addition to varying task, researchers can also bias participants' predictions by providing non-linguistic stimuli with semantic content. For instance, one approach involves preceding word forms with pictures. This method was employed in a picture-word verification paradigm applied by Dikker and Pylkkanen (2011). Here, Dikker and Pylkkanen preceded noun phrases (e.g., *the apple*) with congruent or incongruent pictures of either single objects (e.g., an apple vs. a banana), or multiple related objects (e.g., a bag of groceries vs. a collection of animals). In this way, it was possible to manipulate a specific word form's predictability, independently of its semantic congruency. Results showed a main effect of picture-word congruency on the M100 component observed in magnetoencephalography (MEG) at around 100 ms, wherein less extreme M100 amplitudes were elicited by word forms that matched high-predictability images. It was also reported that this effect was observed only for the picture-word pairs with high predictability. A similar finding was observed during a window from 250 to 400 ms, with a main effect of congruency, but only for the picture-word pairs with high predictability. The early predictability-congruency interaction reported for the M100 is interesting, and is consistent with an account of early effects of top-down modulation of visual or orthographic processing, but as it is (a) less directly relevant to the present thesis, and (b) reported for MEG rather than EEG data, I do not focus on the comparable P1 ERP component observed in EEG, though I discuss the plausibility of, and possible explanations for, this effect in section 7.5. Dikker and Pylkannen did not examine effects in an M170 window with timing comparable to the N1 in EEG, though it is possible that no effects were reported in such a window because none were observed. Nevertheless, it is possible that the study's small sample size (N=7 participants, each presented with 320 trials) may have been insufficient to reliably detect this effect if it exists. Indeed, functional magnetic resonance imaging (fMRI) findings have suggested that predictions in a picture-word verification task may modulate activity in the N1's likely neural generator in the ventral occipitotemporal cortex (vOT; Kherif et al., 2011), though, as previously noted (see section 1.5.1), the coarse temporal resolution of fMRI fails to provide insight into whether top-down contributions influence activity in the early, initial stages of word processing.

As a non-sentential approach, the picture-word verification task has potential to offer insight into the possible top-down modulation of early processes in visual word recognition, while the stimuli, analyses, and sample size used by Dikker and Pylkkanen can be improved and made reproducible. The present study applies a picture-word verification task similar to that used by Dikker and Pylkkanen (2011), to examine whether the N1 component is sensitive to top-down modulation, as operationalised by the difference between picture-congruent and picture-incongruent word conditions. I draw a methodological distinction between congruency, defined as whether a given word matches its preceding picture, and predictability, defined as the likelihood that the image is associated with its congruent word. This stands in contrast to previous studies, which only examined an effect of predictability, but is similar to the design employed by Dikker and Pylkannen, which additionally manipulated congruency. In this way, as Dikker and Pylkkanen did, I was able to examine whether the possible congruency effect is contingent on the predictability of the word given the image that precedes it. Instead of

dichotomising predictability into categories of low and high predictability, however, predictability is here operationalised similarly to Cloze probability, as the proportion of people who identify the given image as the word that follows it. This allowed me to estimate effects more accurately, avoiding an unnecessary loss of information that arises from the dichotomisation or binning of continuous variables (MacCallum et al., 2002; Royston et al., 2006). I predicted that were the N1 sensitive to top-down modulation, there should be an interaction between picture-word congruency and predictability, with no effect of congruency at the lowest level of predictability, and, consistent with findings from sentential studies and the predictive coding account of top-down modulation (see section 1.5), less negative-going N1s for congruent words at higher levels of predictability.

This study was pre-registered at https://osf.io/389ce/, and the reported planned analysis conforms to that specified in the pre-registration.

5.2 Stimuli

Stimuli were designed for two separate tasks in this experiment: the picture-word task, and a localiser task to account for between-participant variability in the N1's timing and location.

5.2.1 Picture-Word Task Stimuli

A total of 400 words (200 per congruency condition; one congruent and one incongruent word per image) were selected with LexOPS (J. E. Taylor et al., 2020), a package in the R programming language (R Core Team, 2021). A list of the full set of stimuli is available in Appendix C.1. The experimental stimuli are summarised in Figure 5.1. First, stimuli were filtered by word prevalence according to Brysbaert et al. (2019), such that at least 90% of participants knew each word. In addition, stimuli were filtered such that all words were nouns according to the dominant part of speech data from SUBTLEX-UK (van Heuven et al., 2014), and had a mean concreteness rating above 4 (on a Likert scale from 1, least concrete, to 5, most concrete) according to Brysbaert et al. (2014). Images were taken from the Bank of Online Standardised Stimuli (BOSS) norms (Brodeur et al., 2014), a large database of images with normed statistics, including percentage of name agreement, which I use here as a measure of predictability. Words were identified as possible picture-congruent words if they were listed as a modal name for any image in the BOSS norms, and were identified as possible picture-incongruent words if they were not.

Picture-congruent and -incongruent words were matched item-wise in terms of several lexical variables as follows: word length (number of characters) exactly; concreteness according to Brysbaert et al. (2014) within ±.25; Zipf frequency (a logarithmic scale of word frequency) according to SUBTLEX-UK within ±.125; character bigram probability (calculated from SUBTLEX-UK) within ±.0025; and OLD20 (the average orthographic Levenshtein distance of the 20 closest neighbours to a given word; Yarkoni et al., 2008) calculated from the LexOPS inbuilt dataset within ±.75. To ensure that picture-incongruent words were not inadvertent possible descriptors for images, the cosine positive pointwise mutual

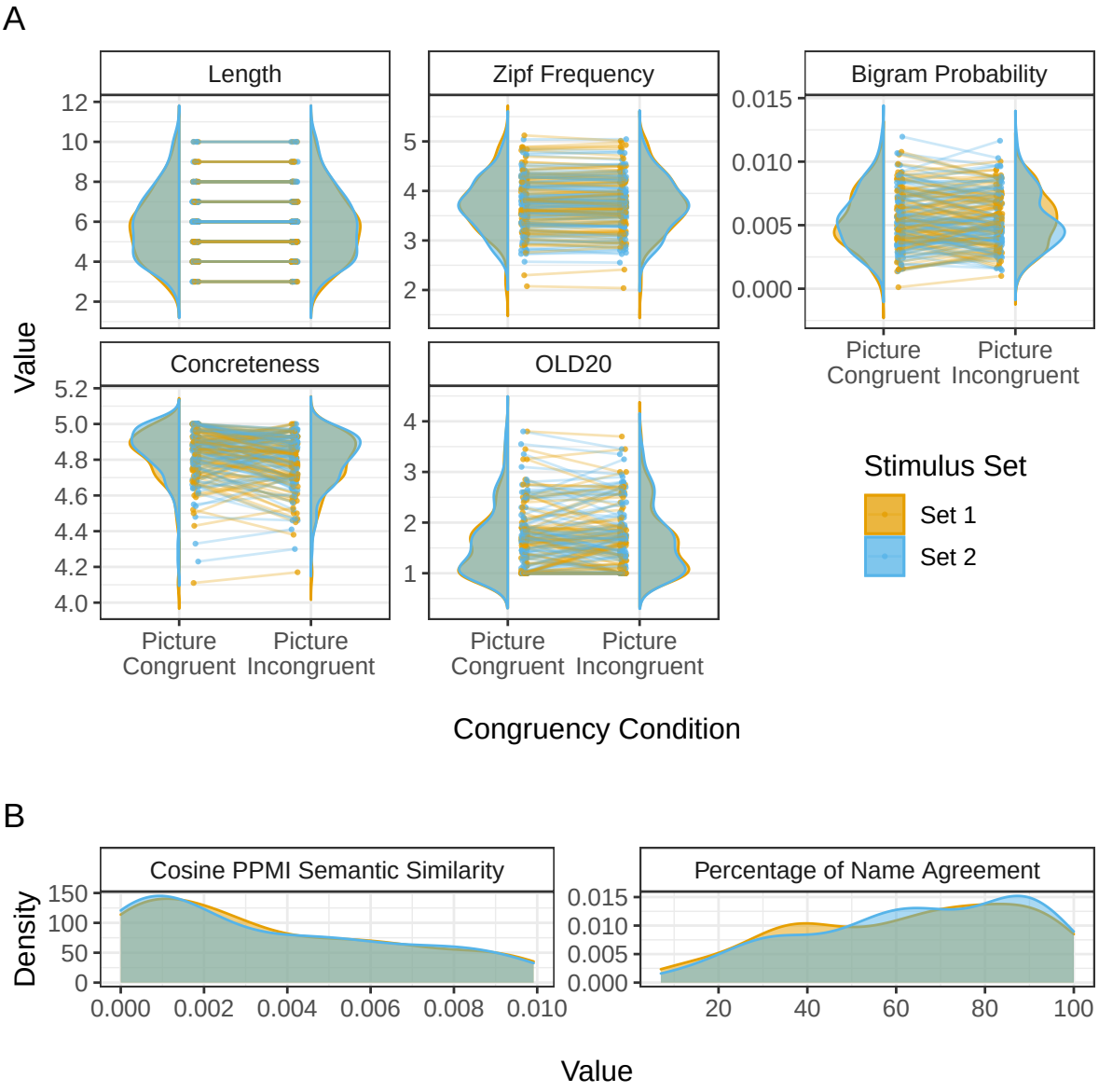


Figure 5.1: Summary of the picture-word stimuli. Each panel depicts how a single variable was controlled. (A) Probability densities for variables which were matched item-wise between picture-congruent and picture-incongruent conditions, and distribution-wise between counterbalanced stimulus Sets 1 and 2. Points representing pairs of words which are matched item-wise are joined by lines. Points' positions are jittered slightly along the x-axis for visibility. (B) Probability densities for variables matched distribution-wise between the counterbalanced stimulus sets for which values are only available for half of the items: Cosine PPMI (Positive Pointwise Mutual Information) Semantic Similarity from SWOW (Small World of Words; De Deyne et al., 2019), and modal name agreement from the BOSS norms. Semantic similarity was actually matched item-wise, but values are only meaningful for half of the stimuli, as items will have a similarity of 1 with themselves, while for percentage of name agreement, values are only available for picture-congruent words.

information (PPMI) measure of associative semantic similarity calculated from the Small World of Words (SWOW) word association norms (De Deyne et al., 2019) was minimised to be $\leq$.01 between each image's matched picture-congruent and picture-incongruent words. To ensure picture-incongruent words did not share orthographic features with their respective picture-congruent words, orthographic Levenshtein distance between matched items was maximised. As items were also matched in word length, this meant all matched pairs of words had a Levenshtein distance equal to their numbers of characters. The variable used to index the predictability of picture-congruent words was percentage of modal name agreement, which was sampled pseudo-randomly (picture-congruent words were not selected if no match could be identified fitting the constraints specified above) from the BOSS norms, and varies continuously in the generated stimuli from 7 to 100%.

As the participants were recruited in the United Kingdom, picture-congruent and -incongruent words were excluded if identified as Americanisms (e.g., *sidewalk*) or if they were modal names for images that the Canadian participants of the BOSS norms are likely to have been more able to name or distinguish (e.g., *buffalo*, *bison*). In addition, words were excluded if identified as shortened versions of proper names (e.g., *limo*, *chimp*) or alternate names for the same object (e.g., *motorbike*, *motorcycle*). Candidate picture-incongruent words were additionally excluded if images the objects would refer to would be markedly dissimilar from the images in the BOSS (e.g., *waiter* or *church*, as there were no images of people or entire buildings in the BOSS), or if they were unimageable despite their high concreteness value (e.g., *item*). Some plural words (e.g., *sticks*) were excluded to ensure that the number of plurals in the incongruent words was similar to that seen in the congruent words. Finally, four images with the modal names *nut*, *trumpet*, *spinach*, and *tuba* were excluded, as I judged the modal names as incorrect descriptions of their images.

To avoid repetition effects from observing the same image twice, each image was presented once, with participants viewing either the picture-congruent or picture-incongruent word. This was counterbalanced by splitting the stimuli pseudo-randomly into two equally sized stimulus sets, referred to as Set 1 and Set 2. Participants were each presented with one of these stimulus sets. Pictures followed by congruent words in Set 1 were followed by incongruent words in Set 2, and vice versa. To minimise any systematic difference between the counterbalanced groups, the stimulus split was selected to maximise the empirical distributional overlap (Pastore & Calcagnì, 2019) between the two stimulus sets in relevant variables, using the distributional matching method detailed in chapter 2. Specifically, the stimulus sets were selected from 50,000 random splits to maximise overlap between the stimulus sets in the distributions of the following variables: percentage of modal name agreement according to the BOSS norms; cosine PPMI semantic similarity according to the SWOW; Zipf word frequency and character bigram probability according to SUBTLEX-UK; word concreteness (Brysbaert et al., 2014); word length; and OLD20. Variables that were also matched item-wise between the conditions were matched distribution-wise separately within each congruency condition. This ensured there were no systematic differences in distributions between conditions or stimulus sets.

To generate stimuli for practice trials, 20 matched pairs of picture-congruent and -incongruent words were generated using the same pipeline as above, except that word

frequency, word concreteness, and character bigram probability were not matched item-wise. The practice trial stimuli were generated from images and words not used in the experimental stimuli. The same practice trials were presented to all participants and were not counterbalanced.

Behavioural Validation

To validate the stimulus generation method for the picture-word stimuli, a behavioural experiment was run using a different (earlier) stimulus set generated from a very similar pipeline. The only differences in the pipeline were that (a) Zipf frequency was controlled within ±.2, (b) Levenshtein distance was not maximised, (c) OLD20 was not controlled for, and (d) the split into stimulus Sets 1 and 2 was optimised from only 20,000 iterations. The stimuli generated for the validation experiment varied in predictability from 12 to 100%. The procedure was also identical to that described in the Procedure section of the present study, except that participants could respond as soon as the word was presented, rather than 1 second after presentation, and the word did not change colour (see Procedure, section 5.4.2, below). Participants comprised 35 monolingual native English speakers (15 female, 19 male, 1 non-binary) who were not diagnosed with any reading disorder. Age varied from 18 to 26 years (M=21.4, *SD*=2.05), and all participants reported being right-handed with normal or corrected-to-normal vision. Trials were excluded if response times (RTs) were less than 250 ms or more than 2000 ms. The logic for the validation experiment was as follows: assuming the stimulus pipeline produces suitably controlled stimuli, increased predictability should facilitate task performance for congruent trials and have either no effect or a minimal effect on performance for incongruent trials.

More traditional methods of modelling response times assume changes in one central tendency parameter of a distribution while other parameters are assumed to be homogenous across observations (e.g., Gaussian models of log-transformed RTs, or Gamma family models). Meanwhile, modelling all parameters of multi-parameter distributions which accurately describe the distribution of observations, like the ex-Gaussian or shifted log-normal distributions for RTs, allows researchers to assess changes in the shape of entire response time distributions (Heathcote et al., 1991; Rouder et al., 2005). The shifted log-normal distribution, for instance, allows researchers to describe changes in the means (μ) and standard deviations (σ) of log-transformed RTs, while also modelling changes in shift (δ). In modelling the validation experiment data, I fit a Bayesian mixed-effects model estimating the same fixed and maximal random effects structure for each parameter (μ, σ, δ) of the shifted log-normal distribution using the *brms* package for R (Bürkner, 2018). More details on this shifted log-normal model are presented in Appendix C.2.

The predictions from the shifted log-normal model showed the expected effects, with predictability leading to faster responses for congruent trials, but having almost no effect on incongruent trials (Figure 5.2). It also demonstrated that when predictability is low, response times show similar central tendency for congruent and incongruent trials though a larger spread in the distribution for congruent trials. When predictability is high, on the other hand,

the difference is mostly due to changes in shift, whereas other features of the distribution are very similar. A more traditional analysis modelling changes in RT as changes in the scale parameter of the Gamma distribution is reported in Appendix C.2, with results corroborating the conclusions drawn from the shifted log-normal model.

5.2.2 Localiser Task Stimuli

The precise timing and location of the N1 varies among studies and participants. Rather than identifying a single electrode and timepoint for all participants before data collection, I designed a localiser task to identify, within an appropriate region and time period of interest, the timepoint and electrode at which each participant's maximal sensitivity to orthography emerges. This data could then be used to extract N1 amplitudes in the picture-word task, while accounting for variability among participants in timing and topography of orthographic processes.

For the localiser task, three categories of stimuli were presented for 100 trials each (Figure 5.3). These consist of matched triplets of words (Courier New font), false-font strings (BACS2serif font), and phase-shuffled words. While the comparison between words and false-font strings is a more traditional measure of N1 sensitivity to orthography, with previous evidence suggesting a more robust difference from words than exists between nonwords and words (Brem et al., 2018; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019), phase-shuffled words have been designed for this study as a more controlled alternative comparison for exploratory analyses, with equal spatial-frequency amplitude and permuted spatial-frequency phase. Similar phase-shuffled word stimuli have shown robust differences to word forms in fMRI investigations of vOT activity (Rauschecker et al., 2012; Rodrigues et al., 2019; White et al., 2019; Yeatman et al., 2013).

To generate these stimuli, a large list of suitable words (N=27,332) was identified by filtering the word prevalence norms of Brysbaert et al. (2019) to only contain words known by at least 90% of participants and which were not selected for the main experiment. A representative sample (N=100) of this list was generated by maximising distributional overlap (Pastore & Calcagnì, 2019), between the sample and the full list of candidates, on 13 variables where observations were available: word prevalence (Brysbaert et al., 2019); length (number of characters); word frequency in Zipf in SUBTLEX-UK (van Heuven et al., 2014); part of speech according to SUBTLEX-UK; character bigram probability calculated from SUBTLEX-UK; OLD20 (Yarkoni et al., 2008) calculated from the LexOPS dataset (J. E. Taylor et al., 2020); concreteness (Brysbaert et al., 2014); age of acquisition (Kuperman et al., 2012); average lexical decision response time (RT) and accuracy according to the British Lexicon Project (Keuleers et al., 2012); and the emotion ratings of valence, arousal, and dominance (Warriner et al., 2013). Similarity in the categorical variable of part of speech was maximised with dummy-coded variables. Distributional similarity was maximised by selecting the sample from 500,000 random samples with the highest total distributional overlap with the full list of possible words. The selected sample of words is summarised in Figure 5.4. The full list of word stimuli for the localiser task is presented in Appendix C.3.

The false-font strings consist of characters from the Brussels Artificial Character Sets

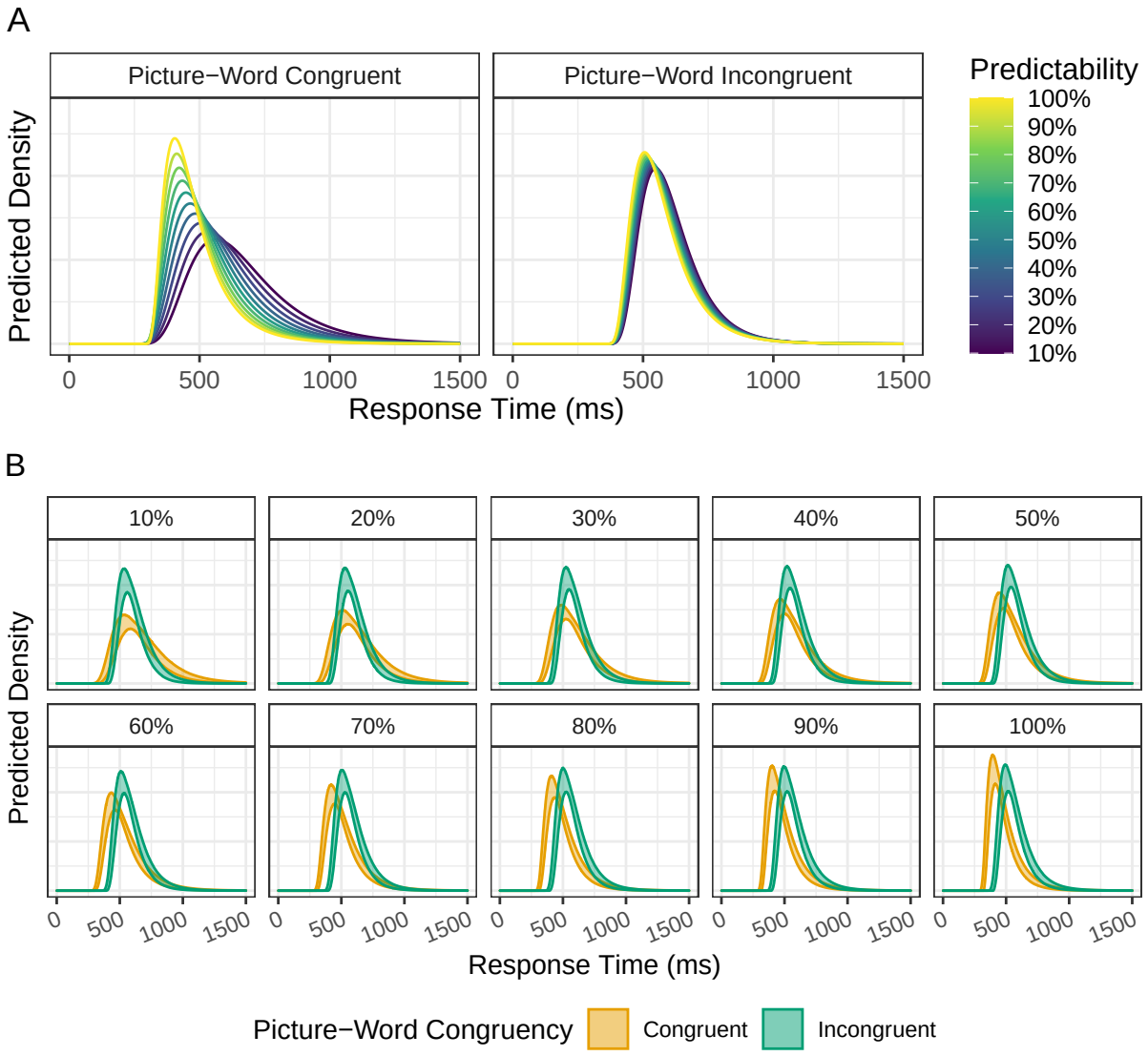


Figure 5.2: Fixed effect predictions of RT distributions in the behavioural validation experiment for the picture-word stimuli. Predictions of RT distributions are shown for congruent and incongruent trials for values of percentage of name agreement, from 10 to 100% in steps of 10. These predictions were estimated from a single Bayesian mixed-effects model, modelling the same fixed and random effects structure for each parameter of a shifted log-normal distribution. The two panels show the same results but highlight (A) the effect of predictability for picture-congruent and picture-incongruent words, and (B) the effect of picture-word congruency at different values of predictability, showing the degree of certainty in the predictions with the 89% highest density intervals (HDIs) of the predictions from all posterior samples. Density is scaled consistently across panels.

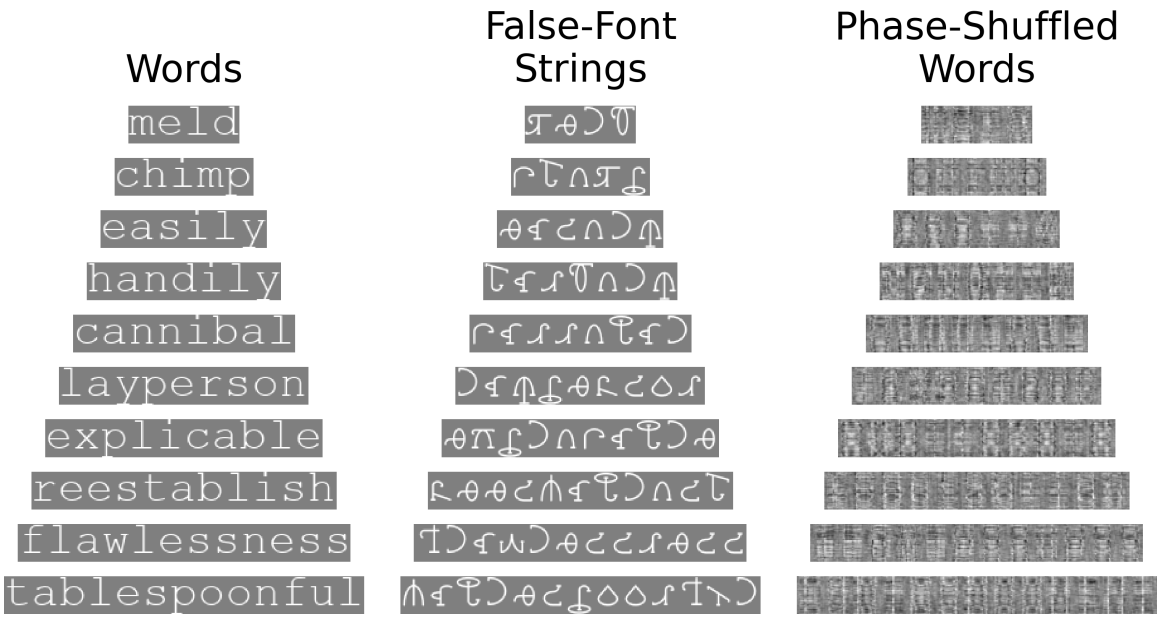


Figure 5.3: Ten example stimuli for each stimulus type in the localiser task. Each row represents a matched triplet of word, false-font string, and phase-shuffled word stimuli. The phase-shuffled word images were generated uniquely for each trial.

(BACS; Vidal et al., 2017) font. Specifically, I used the font BACS2serif, to create an item-wise false-font match to each word, where every Courier New character in the word stimuli is replaced with a BACS character matched in the number of strokes, junctions, terminations, and serifs. The phase-shuffled stimuli were generated by using a Fourier transformation to extract the phase and amplitude from the word images. Phase values were randomly shuffled (i.e., permuted), while amplitude values were preserved. An inverse Fourier transformation was then used to generate a new image with the original amplitude values, but with phase randomly shuffled. To prevent phase shuffling from resulting in noticeably large changes in contrast, the phase shuffling was done on a version of the word image with 50% of the original contrast. After the inverse Fourier transformation, the contrast of the generated phase-shuffled image was readjusted to equal that of the original word image. The phase-shuffling method was chosen because unlike replacing phase with random noise (e.g., uniformly distributed phase between -π and π), permuting the original phase values preserves the original image's overall distribution of phase, preserving coarse spatial frequency information such as spaces between letters. To avoid repeating the same stimuli across participants more than necessary, unique phase-shuffled images were generated for each trial, for each participant.

Versions of the localiser task's stimuli were also created in green, to be displayed when the participant is required to respond. For words and nonwords, this was done by simply changing the font colour to be green. To preserve image intensity, the colour of phase-shuffled images was changed by altering pixels in the following way. For pixels in which the value in the green channel was less than 50% of the maximum intensity (i.e., the intensity of all channels in the grey background), values in red and blue channels were altered to equal the value in the green channel for that pixel. For all other pixels, the values in red and blue channels were set to 50% of the maximum intensity.

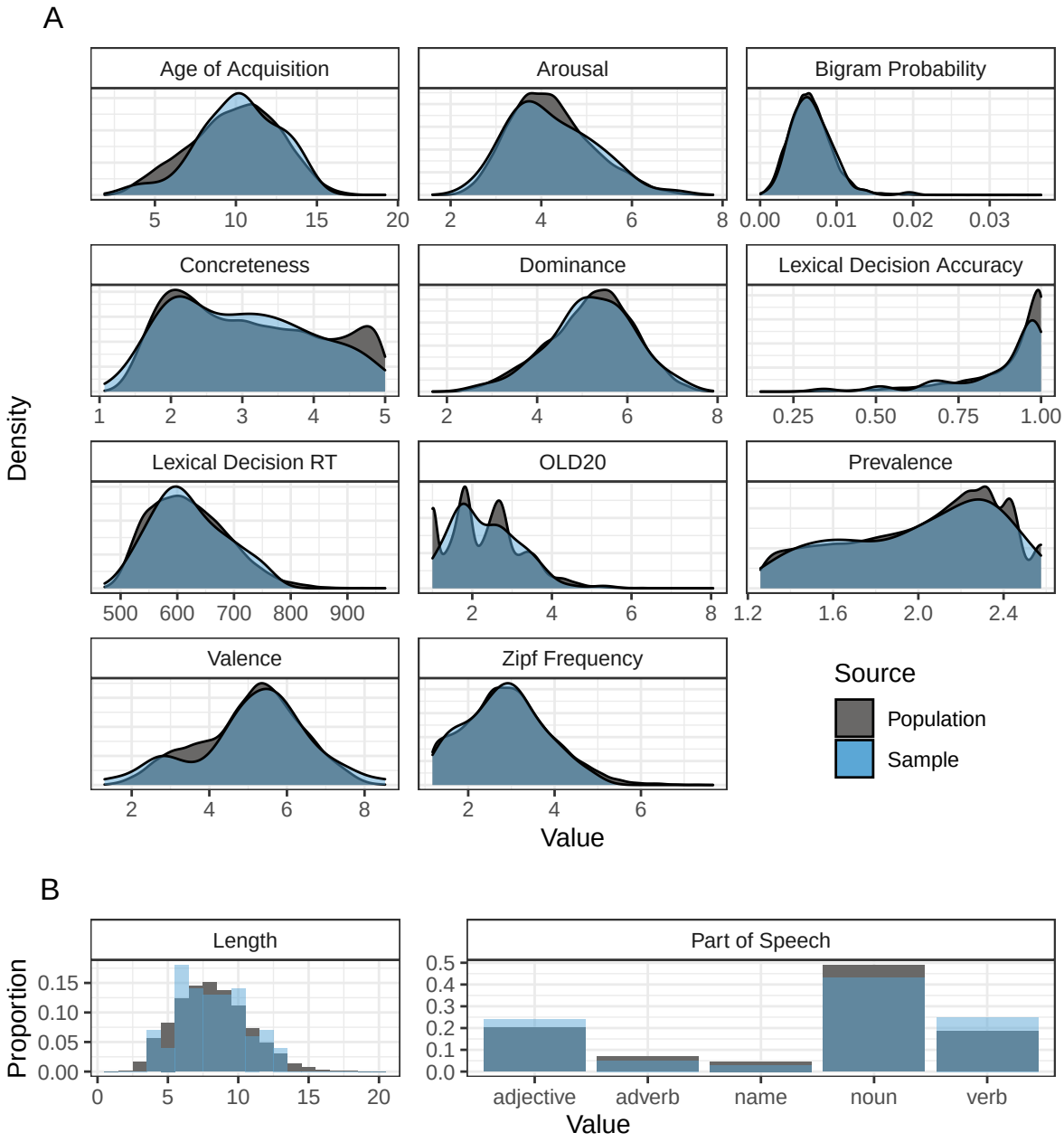


Figure 5.4: Distributions of key variables illustrate the similarity between the selected localiser stimuli words (*sample*) and the list of all words known by at least 90% of participants (*population*). Panel A shows distributional similarity of continuous variables. Panel B shows similarity in length (all integer values) as a histogram showing proportions, and the similarity in the counts of each part of speech category as a bar plot of proportions. Only the part of speech categories which were present in the sample are shown. No members of less common part of speech categories like determiner or number were selected in the sample.

5.3 Power Analysis

I conducted simulations to identify the number of participants required to reach at least 80% power (an arbitrary but commonly used target for statistical power), if I were to carry out the same experiment a large number of times. Since I expected a much larger effect of predictability in the congruent trials (incongruent trials were originally included as a control condition), the planned analysis for this experiment focuses on the congruency-predictability interaction. A fixed effect coefficient for the interaction in the expected direction would be evidence for an effect of predictability on the N1. The expected fixed effects coefficients were calculated assuming an interaction between predictability and image-word congruency consisting of a .75 μV reduction in N1 amplitude for the most predictable congruent trials relative to the least predictable trials, and no difference for incongruent trials. This effect size was based on proportional effects observed in occipitotemporal electrodes' peak N1 amplitude in a single-electrode analysis of chapter 4.

To determine the .75 μV effect size, first I decided to simulate the difference as a proportion of the maximum N1 amplitude, because different EEG systems and setups can result in vastly different voltage measurements. Next, to identify a realistic proportional difference at the maximum level of predictability (100% name agreement) between picture-congruent and picture-incongruent words, I considered the design by A. E. Kim and Gilley (2013), which is as close to this design as I could find. In their study, 53 participants were presented with highly predictable target words which were either prediction-congruent or prediction-incongruent. Kim and Gilley observed left-lateralised occipitotemporal electrodes' N1 peaks that were less negative when the word was prediction-congruent (-2.6 μV) than when the word was prediction-incongruent (-3.9 μV), equal to a proportional difference of .33. A less comparable, though still possibly informative, study from A. Kim and Lai (2012) presented 20 participants with 180 high Cloze probability sentences (with 550 ms SOAs such that overlap of ERPs was minimised). The last word in each sentence was either a highly predictable word, an orthographically similar pseudoword, an orthographically dissimilar pseudoword, or a consonant string nonword. Here, the N1 (170-205 ms) for a left occipitotemporal electrode was shown to be more negative for nonwords and orthographically dissimilar pseudowords (both around -4 μV) than for the predicted word and an orthographically similar pseudoword (both around -3 μV). This is equal to a proportional difference of .25. I decided that other potentially comparable studies, published at the time of the power analysis, were too different in their experimental design, either because they used manipulations other than varying predictability (Y. Chen et al., 2013, 2015; Segalowitz & Zheng, 2009; Strijkers et al., 2015; F. Wang & Maurer, 2017, 2020) or they presented the target items midway through sentences using an SOA of 300 ms or less resulting in ERPs overlapping with those of preceding words (Dambacher et al., 2012; Kretzschmar et al., 2015; Sereno et al., 2019).

As a conservative estimate, given the lack of relevant data, I decided a proportional difference of .15 was a realistic effect size for the difference between picture-congruent and picture-incongruent trials at the maximum level of predictability. In the electrodes that show the greatest N1 peak in the data collected for chapter 4, I observed a mean peak N1 amplitude

of around -5 μV. Assuming a proportional difference of .15, I would therefore expect a .75 μV reduction in N1 amplitudes at the highest level of predictability, relative to the lowest level of predictability, in the picture-congruent condition. The values I predicted for the extremities of each independent variable are presented in Table 5.1.

Table 5.1: The coding method and predicted N1 amplitudes for the extremities of each predictor variable. As congruency is deviation-coded and there are an equal number of congruent and incongruent trials, the values for $Cong_{spw}$ are presented as between -.5 and .5, though the actual values are likely to differ slightly after observations fitting exclusion criteria are removed (in both the simulation and actual analysis). $Pred_{spw}$ values are calculated as proportion of agreement normalized between 0 and 1.

Congruency	$Cong_{spw}$	Percentage of modal name agreement (%)	$Pred_{spw}$	Predicted N1 amplitude (μV)
Incongruent	-.5	7	0	-5.00
Incongruent	-.5	100	1	-5.00
Congruent	.5	7	0	-5.00
Congruent	.5	100	1	-4.25

In each iteration of the simulation, I simulated 200 (100 per congruency condition) trials for each of *N* subjects with subject-, picture-, and word-specific random intercepts and slopes. The predictability values were taken directly from the generated stimuli. The simulation can be understood through reference to the formula that describes the linear mixed effects model:

$$y_{spw} = \beta_0 + S_{0s} + P_{0p} + W_{0w} + (\beta_1 + S_{1s} + P_{1p})Cong_{spw} + (\beta_2 + S_{2s})Pred_{spw}$$
$$+(\beta_{12} + S_{12s})Cong_{spw}Pred_{spw} + e_{spw}$$

Table 5.2 explains each term in this model and presents the values simulated for the power analysis. The simulated values for the fixed effects were calculated based on the predictions and coding scheme, and are also presented in Table 5.2. The simulated values of subject random intercepts were based on mixed effects models for N1 amplitudes in chapter 4, where subject random effects showed much greater variability between subjects than items. The variance for the distribution residuals was also based on estimates from mixed effects models in chapter 4. Due to the coding method of the coefficients, the β terms in the table and equation above can be interpreted as follows:

β_0 reflects the average amplitude at the lowest level of predictability,

β_1 reflects the difference between congruent and incongruent trials at the lowest level of predictability,

β_2 reflects the overall effect of predictability across congruent and incongruent trials, and

β_{12} reflects the difference between congruent and incongruent trials at the highest level of predictability.

Table 5.2: The meaning of each term in the design's linear mixed effects model, and the value simulated for the power analysis. Where simulated variables were drawn from distributions, $N(\mu, \sigma)$ indicates that the respective variable's values were drawn from a normal distribution with mean μ and standard deviation σ.

Term	Meaning	Simulated Value (μV)
y_{spw}	Trial-level N1 amplitudes for subject *s*, picture *p*, and word *w*	
β_0	Grand intercept	$= -5$
S_{0s}	Subject random intercept for subject *s*	$\sim N(0, 2.5)$
P_{0p}	Picture (image) random intercept for picture *p*	$\sim N(0, 2.5)$
W_{0w}	Word random intercept for word *w*	$\sim N(0, 2.5)$
β_1	Fixed effect of congruency	$= 0$
S_{1s}	Subject random slope for congruency for subject *s*	$\sim N(0, .75)$
P_{1p}	Picture (image) random slope for congruency for picture *p*	$\sim N(0, .5)$
$Cong_{spw}$	Trial-level congruency values (deviation-coded)	
β_2	Fixed effect of predictability	$= .375$
S_{2s}	Subject random slope for predictability for subject *s*	$\sim N(0, 1)$
$Pred_{spw}$	Trial-level predictability values	
β_{12}	Fixed effect of congruency-predictability interaction	$= .75$
S_{12s}	Subject random slope for congruency-predictability interaction for subject *s*	$\sim N(0, 1)$
e_{spw}	Residual random noise	$\sim N(0, 3)$

Due to a lack of relevant data from similar designs, variance-covariance matrices for the power analysis were simulated with all random effects correlations set to zero. To check this did not result in heavily biased estimates, the power analysis was also run with all random effects correlations set to values of .2, .4, .6, and .8. Each of these analyses estimated a strikingly similar relationship between the number of participants and statistical power (see Appendix C.4). In each simulation, simulated participants were pseudo-randomly assigned to stimulus sets 1 and 2 in equal number, or with randomly allocated counts of $\frac{N}{2} - 0.5$ and $\frac{N}{2} + 0.5$ if the number of simulated participants were odd. N varied from 10 to 100 in steps of 5, with 500 iterations run at each value. Before models were fit to simulated data of each iteration, data exclusion was simulated as a random 10% loss of trials. This consisted firstly of the 5% of data loss observed in the stimuli validation due to trials being responded to incorrectly or with response times less than 250 ms or greater than 1500 ms. No lower bound for response time exclusions was applied in the EEG experiment, as the word was visible for 1 second before responses are permitted. As a conservative estimate, however, I expected a similar percentage of data loss to that seen in the validation for the picture word stimuli. The remaining 5% of data loss was simulated because, given the participant exclusion criteria, this is the maximum allowable loss of data due to a combination of technical problems with the EEG system. This conservative estimate can be considered a worst-case scenario in terms of EEG data loss. The possibility of participants being excluded was not simulated, as I opted to simply continue collecting data until I reached the desired number of participants, and excluded participants' data would not be analysed. Covarying random effects were simulated using the R package *faux* (DeBruine, 2020). Linear mixed effects models were fit using the same functions, formula, and optimiser as those used for the analysis of the actual data (section 5.5.1). In the case of non-convergence, models were re-fit without random correlations before significance testing, as this is the action I would take when modelling the actual data. Likelihood ratio Chi-square model comparisons were conducted between the full model and a version of the model lacking the interaction term, and the resulting *p* values were recorded from each iteration.

Given that the hypothesis was directional, simulated significance tests were performed using one-tailed comparisons with an alpha level of .05. Running only 500 simulations is likely to give noisy estimates of power when simulating data which can vary in many parameters. Since fitting a much larger number of models would be unfeasible due to the time taken to fit each mixed effects model, the underlying relationship between the number of participants and the design's statistical power was estimated by fitting log-linear binomial generalised linear models (GLMs) to all iterations for one-tailed and two-tailed comparisons. Figure 5.5 depicts the resulting power curves. The power analysis suggested that a sample size of 68 participants (divisible by four, so as to assign an equal number of participants to each combination of counterbalanced response and stimulus groups) would be sufficient to reach at least 80% power for detecting the effect of interest in the predicted direction with a one-tailed comparison. Specifically, the model predicted that at this number of participants, I would have 81.72% power (99% confidence interval = [80.46%, 82.91%]) to detect the predicted effect.

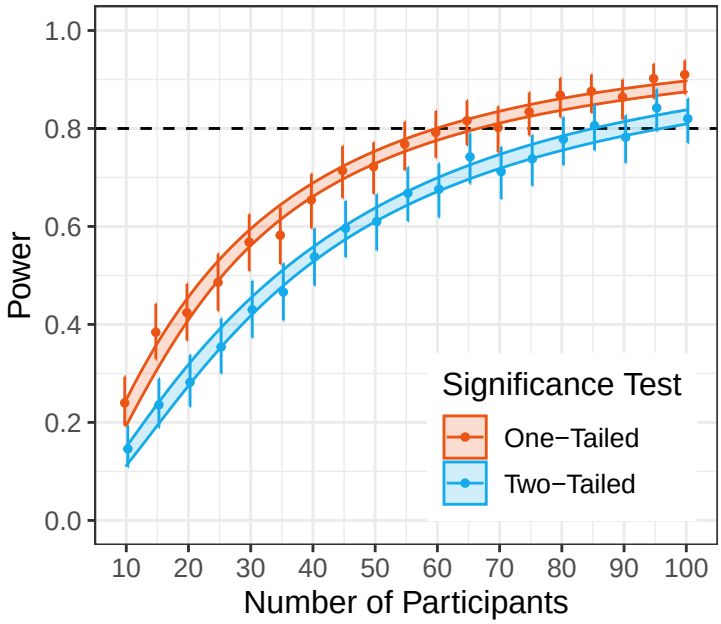


Figure 5.5: Power curves calculated from the simulations. For comparison, both one-tailed and two-tailed power are presented, though the *p* value used in the actual planned analysis is one-tailed. Points (shifted horizontally for visibility) present the observed proportions of simulations which resulted in statistically significant p values. Vertical error bars present 99% binomial confidence intervals of these individual proportions. The coloured lines showing a logarithmic relationship depict the upper and lower bounds of 99% confidence intervals of predicted of probabilities from log-linear binomial GLMs fit to the data. The dashed horizontal line highlights the 80% power target.

5.4 Methods

5.4.1 Participants

68 monolingual native English speakers (40 female, 27 male, 1 non-binary) participated in the study. Participants were randomly allocated into one of the four combinations of stimulus set and response group (i.e., the mapping of the two response buttons to affirmative and negative responses), such that each combination of stimulus set and response group comprised 17 participants. No participants were diagnosed with any reading disorder. Ages varied from 18 to 37 years (M=22.69, *SD*=4.9), and all participants reported having normal or corrected-to-normal vision. Participants' handedness was assessed via the revised short form of the Edinburgh Handedness Inventory (Veale, 2014), with participants only permitted to take part if they scored a laterality quotient of +40 indicating right handedness.

Exclusion criteria for participants were determined prior to data collection as: (1) if 10 or more channels show an offset more extreme than ±25 mV (as measured on the BioSemi acquisition software, ActiView), or (2) if more than 5% of the trials are lost due to technical issues with the EEG system. As no participants satisfied these criteria, no participants were excluded after data collection.

Data collection was approved by the University of Glasgow School of Science and Engineering Ethics Committee (application number: 300200117).

5.4.2 Procedure

Stimuli were presented on a VPixx Technologies VIEWPixx screen (resolution 1920*1080 pixels, diagonal length 23", model VPX-VPX-2004A). Participants completed the experiment on a chin rest positioned 48 cm from the centre of the screen. Stimuli were presented on a grey background equal to 50% of the maximum intensity in each colour channel, roughly 12.3 cd/m^2. The experiment was written using the Python library *PsychoPy* (Peirce, 2007), and all code and materials are available in the GIN repository. All stimuli were presented centrally (horizontally and vertically). All trials in both tasks were presented in a pseudo-randomised order, such that no more than five consecutive trials required the same response from the participant. Trials were randomised across blocks, with the exception of the practice block, for which trials were randomised within the block such that all participants observe the same practice stimuli but in a random order.

Participants started with the localiser task, in the form of a lexical decision task (Figure 5.6A). The localiser task began with 30 practice trials, and was then followed by 300 trials split into 5 blocks of 60 trials. Each trial began with the bullseye fixation target recommended by Thaler et al. (2013) (outer circle diameter: 0.6° of visual angle, inner circle diameter 0.2°), presented for 300 ms. This was followed by a jittered interval of between 300 and 1300 ms, during which the screen was blank. The stimulus (word, false-font string, or phase-shuffled word image) was then presented at a height of 1.5° (width of 1.07° for one character). Words and false-font strings were presented in white (80 cd/m^2), in the respective fonts of non-proportional Courier New and BACS2serif font. The stimulus was visible for 500 ms, after which the font colour changed to

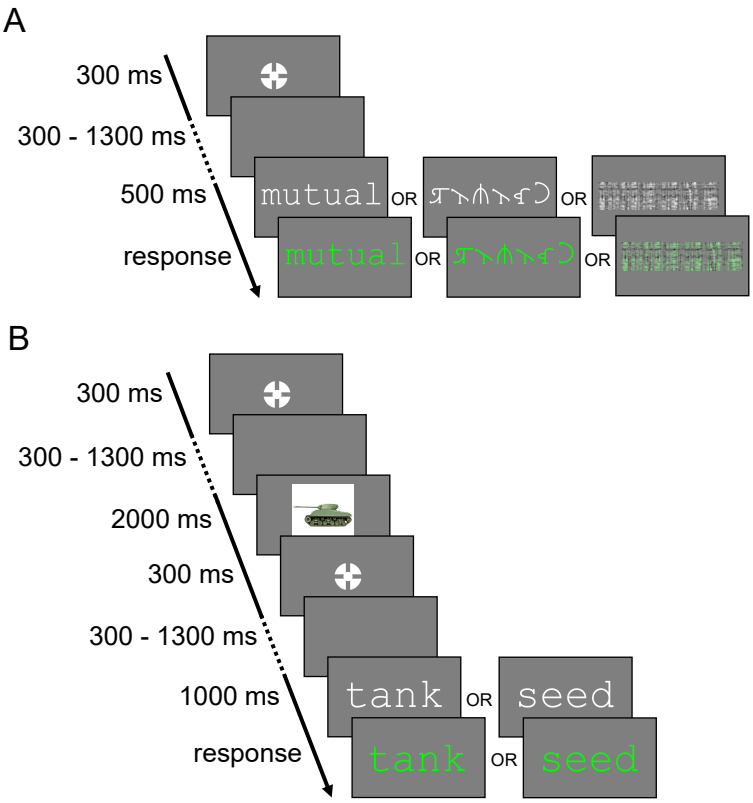


Figure 5.6: Trial structure of the (A) localiser task and (B) picture-word task. This figure is illustrative and the sizes are not to scale; in the experiment, images were in fact presented at a much larger scale than the words.

green and participants could respond. The stimulus changed colour to signal that a button press could be made. Although I did not plan to analyse data between the colour change and the participant's response, I anticipated that this data could be of interest to other researchers. Participants were requested to respond after the stimulus changed colour once, quickly and accurately, to indicate whether the stimulus they saw in each trial was either a word or not a word. Responses were given with the right and left control ('Ctrl') keys of a QWERTY keyboard, with the mapping of affirmative and negative responses counterbalanced across participants. After the participant had responded, the next trial began.

After the localiser task, participants completed the picture-word task (Figure 5.6B), which is composed of an initial practice block of 20 trials, followed by 200 trials split into 5 blocks of 40 trials. As in the localiser task, each trial in the picture-word task began with the bullseye fixation point, presented for 300 ms, after which there was a blank screen for a jittered interval of between 300 and 1300 ms. An image was then presented for 2000 ms, at a size of 10x10°. After the image, the bullseye fixation point was presented again for 300 ms, followed by another interval jittered between 300 and 1300 ms. The word was then presented in white Courier New font, at a height of 1.5° (width 1.07° for one character). After 1000 ms, the word turned green, and the participant could provide their response to indicate whether the word describes the image they saw. As in the localiser task, responses were given with the right and left control ('Ctrl') keys of a QWERTY keyboard, with the mapping of affirmative and negative responses

counterbalanced across participants, but kept consistent within participants across the two tasks. After the participant had responded, the next trial began. There was no deadline for participants to respond. The instructions given to participants for the picture-word task are presented in Appendix subsection C.5.

The first blocks of both tasks consisted of practice trials with 10 exemplars for each stimulus type (word/false-font string/phase-shifted image and congruent/incongruent trials for each task, respectively), during which participants were additionally given immediate feedback on their accuracy for each trial. These practice trials were followed by green text reading "CORRECT!" if the participant responded correctly, or else by red text reading "INCORRECT!", presented in Courier New font with a height of 1.5°, for 1000 ms. Participants had self-paced breaks between blocks for each task. Before the practice trials and at the start of every experimental block, participants were presented with instructions for the task (see Appendix C.5), summarising what would occur in each trial, and specifying that they should respond as quickly and accurately as possible once the stimulus turns green. These instructions also specified which keys participants should press to indicate their decision. After each experimental block, including the practice trials, participants were presented with their average accuracy and median response time. After the practice trials, participants were additionally given the option to run the practice trials again if they wish.

5.4.3 Recording

EEG data were recorded using a 64-channel BioSemi system, sampling at 512 Hz, with an online low-pass filter at the Nyquist frequency. Electrodes were positioned in the standard 10-20 system locations. Four electro-oculography (EOG) electrodes were placed to record eye movements and blinks: 2 were placed to the sides of eyes (on the right and left outer canthi), and 2 below the eyes (on the infraorbital foramen). Electrode offset was kept stable and low through the recording, within ±25 mV, as measured by the BioSemi ActiView EEG acquisition tool. Electrodes whose activity exceeded this threshold were recorded, to be removed in data preprocessing.

5.4.4 Preprocessing

The following section details the procedure applied to EEG data from each individual session, with the same pipeline being applied to both the localisation task and picture-word task unless otherwise specified. EEG preprocessing was achieved using functions from the EEGLAB (Delorme & Makeig, 2004) toolbox for MATLAB (MATLAB, 2020) or OCTAVE (Eaton et al., 2020). For both tasks, trials were excluded if responded to incorrectly (N=368 in localiser task, N=226 in picture-word) or later than 1500 ms after the word (or nonword) changed colour (N=41 in localiser task, N=42 in picture-word).

Channels recorded as having offsets ±25 mV during data acquisition were removed from the data, with their activity to be later interpolated. The EEG data were then be re-referenced to the average activity across all electrodes and filtered with a 4th order causal Butterworth filter between .5 and 40 Hz. To counteract the distortion in signals' timing (phase) that is inherent

to causal filters, the filter was applied in both directions, with the MATLAB function, *filtfilt()*. Segments of data outside of experimental blocks (i.e., in break periods) were identified and removed so they do not impact the independent components analysis (ICA) applied later in the pipeline. Here, blocks were identified as beginning 500 ms before stimulus presentation in the first trial of each block, ending 500 ms after the end of the last trial's epoch. To reduce the impact of occasional non-stationary artefacts with high amplitude (such as infrequent muscle movements), artefact subspace reconstruction (ASR; C. Y. Chang et al., 2020) was used with a standard deviation cutoff of 20 to remove non-stationary artefacts. Following this, an ICA was run on the data to identify more stationary artefacts. The ICA was run using the FastICA algorithm (Hyvärinen & Oja, 1997), with a recorded random seed for reproducibility. The ICA was run on a copy of the data with channel offsets removed to allow for better sensitivity to electro-oculogram (EOG) artefacts (Groppe et al., 2009). The ICLabel classifier (Pion-Tonachini et al., 2019) was used to automatically identify artefacts which were eye-related or muscle-related. Components classified by ICLabel as eye-related or muscle-related with a probability of $\geq$85% were removed from the data. Following eye movement artefact removal, activity from channels which were removed was interpolated via spherical splines, as implemented in EEGLAB. Trials were then epoched and baseline-corrected to the 200 ms preceding stimulus presentation. For the localiser task, stimulus presentation refers to the time point at which words, false-font strings, or phase-shuffled images were presented; in the picture-word task, stimulus presentation refers to the target word.

5.5 Results

The planned analysis (pre-registered at https://osf.io/389ce/) examined the whether the predicted effect of predictability-dependent reduction of N1 amplitudes for picture-congruent words, as outlined in the power analysis (section 5.3), was observed in the electrode which for each participant showed the maximal sensitivity to orthography. I also examine the time-course of the effect of picture-word congruency, and of the congruency-predictability interaction, and behavioural results, in the picture-word task. I then report patterns of results for the time-course of sensitivity to orthographic features in the localiser task, and corresponding effects on participants' behaviour.

5.5.1 Planned Picture-Word Analysis

Electrodes that showed maximal sensitivity to orthographic information in the N1 were identified for each participant using data from the localisation task (Figure 5.7). Specifically, the maximal electrode was identified, from an occipitotemporal region of interest (Figure 5.7A), as that which shows the largest mean amplitude difference, in the expected direction, across all localiser trials between word and false-font string stimuli, in the time window between 120 and 200 ms. Here, the expected direction, based on previous findings (Appelbaum et al., 2009; Bentin et al., 1999; Eberhard-Moscicka et al., 2016; Pleisch et al., 2019; J. Zhao et al., 2014) was a more negative-going N1 for words than for false-font matches. In contrast to some previous studies

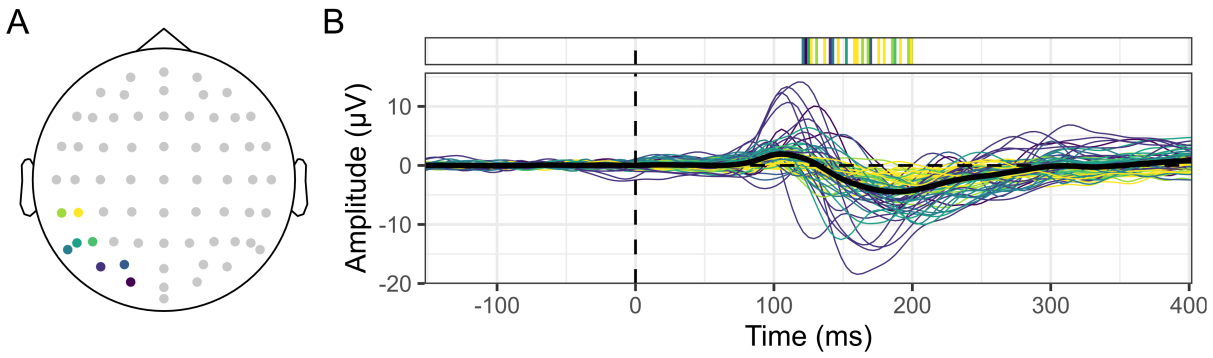


Figure 5.7: The method by which trial-level amplitudes were extracted for the planned analysis. (A) Electrodes in the left-lateralised occipitotemporal region of interest, from which maximal electrodes were identified. The selected electrodes were in the standard 10-20 locations: *O1*, *PO7*, *PO3*, *P9*, *P7*, *P5*, *TP7*, and *CP5*. (B) Maximal timepoints were identified for each participant's maximal electrode (upper panel), from which trial-level amplitudes were extracted from the ERP of each participant's maximal electrode. In the larger, lower panel, coloured lines depict average ERPs for per-participant maximal electrodes, while the thicker black line depicts the overall average.

whose N1 windows extended beyond 200 ms, I decided to set 200 ms as an upper bound for the possible maximal timepoint in the main analysis, to ensure effects were indeed restricted to the N1, and not later components like the N400. The topographic region of interest consists of a cluster of eight left-lateralised occipitotemporal electrodes (Figure 5.7A) which reflect the typical topography of the N1. The timepoint in the window at which the maximal electrode shows the greatest sensitivity to the word-versus-false-font difference in the expected direction, that is, the "maximal timepoint", was also recorded (Figure 5.7B). The identified maximal electrode and maximal timepoint were then used to extract trial-level N1 amplitudes from the picture-word task. To reduce the influence of noise on trial-level data, the trial-level N1 amplitudes in the picture-word task were calculated as the maximal electrode's mean amplitude across 3 time points: the participant's maximal timepoint, one sample preceding the maximal timepoint, and one sample following the maximal timepoint. At the recorded sample rate of 512 Hz, this is equivalent to a window of 3.91 ms centred on the maximal timepoint.

The trial-level N1 amplitudes from the picture-word task were modelled using a linear mixed-effects model fit with the R package *lme4* (Bates et al., 2015), estimating the maximal random effects structure justified by the experiment's design (Barr et al., 2013) as detailed in the section on the power analysis (section 5.3). The model was fit using the *bobyqa* optimiser (Powell, 2009). In *lme4* syntax, the formula for the mixed-effect model was specified as:

```
amplitude ~ 1 + congruency * predictability +
  (1 + congruency * predictability | participant_id) +
  (1 + congruency | image_id) +
  (1 | word_id)
```

In this formula, *amplitude* is the trial-level N1 amplitude in microvolts, while *congruency* is a deviation-coded categorical variable indicating whether a given trial's word was picture-congruent or picture-incongruent, and *predictability* refers to the proportion of name agreement in the BOSS norms, normalised between 0 and 1. A consequence of this coding method is that the model's intercept reflects the predicted amplitude at the lowest level of predictability, averaged across both levels of congruency, while the slopes' coefficients are standardised and directly comparable in their magnitude. The variables of *participant_id*, *image_id*, and *word_id*, in the formula, identify each trial's participant, image, and word, respectively.

The fixed effect relationships predicted by the model are presented in Figure 5.8. The model intercept, reflecting the average N1 amplitude at the lowest level of predictability, was estimated to be β=-3.49 (*SE*=.47). The fixed effect of congruency from this model was estimated as β=.55 (*SE*=.3), which, because predictability was scaled to between 0 and 1, means that at the lowest level of predictability (7%), N1 components for picture-incongruent words were .55 μV more negative-going than those for picture-congruent words. The main effect of predictability was estimated as β=.54 (*SE*=.25), meaning that N1 amplitudes, averaged across congruent and incongruent conditions, were .54 μV less negative-going at the highest level (100%) than at the lowest level of predictability (7%). The effect of interest, the interaction between congruency and predictability, was in the opposite direction from that predicted in the power analysis (β=-.93, *SE*=.43), with a larger, positive effect of predictability for picture-incongruent trials than for picture-congruent trials. A likelihood ratio Chi-square model comparison yielded a two-tailed *p* value of .019 (χ^2(1)=5.5) for this effect, though as the planned analysis was one-tailed for an effect in the opposite direction, this was not interpreted as significant.

To briefly describing the interaction in an exploratory manner, I report two-tailed *p* values and their Bonferroni-corrected counterparts (p_{bonf}). For picture-incongruent words, the effect of predictability was estimated to be β=.99 μV (*SE*=.31, χ^2(1)=9.38, *p*=.002, p_{bonf}=.004), while for picture-congruent words, the effect of predictability was estimated as β=.06 μV (*SE*=.34, χ^2(1)=3.24, *p*=.07, p_{bonf}=.14).

For comparison, I also analysed the data using the maximal electrodes that would be identified from the comparison between words and phase-shuffled words (Appendix C.6). This analysis revealed a similar pattern of effects, with picture-incongruent words eliciting less negative-going N1 components as predictability increases, and with this effect being much closer to zero for picture-congruent words.

5.5.2 Exploratory Picture-Word Analysis

To better understand the pattern of results observed in the planned analysis, I conducted an exploratory analysis examining the time-course of effects in the occipitotemporal region of interest. I also examined the behavioural results, and compared these to the results from the behavioural validation study (section 5.2.1).

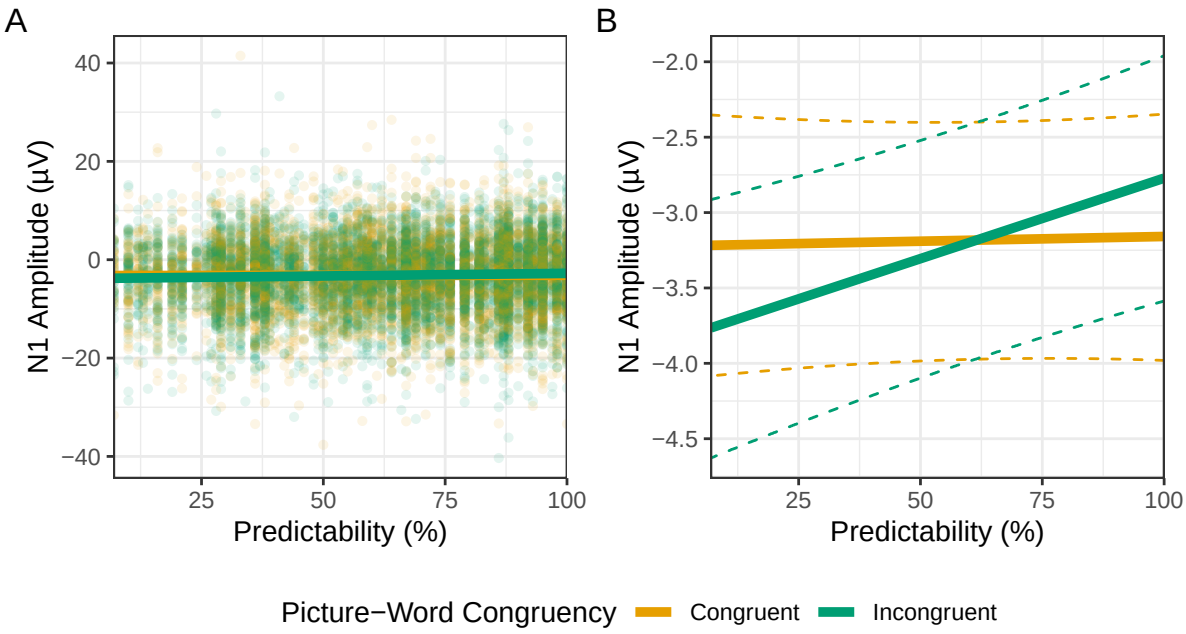


Figure 5.8: Fixed effect predictions from the planned analysis of the picture-word task. (A) Model-derived fixed-effect predictions, visualised over results from all trials (individual points). (B) Fixed-effect predictions visualised alone for visibility, where dashed lines depict the bounds of 95% bootstrapped prediction intervals (estimated from 5,000 iterations), where bootstrapped predictions were generated using the *bootMer()* function of *lme4*. For feasibility, bootstrapped predictions were generated from a version of the model that lacked random slopes.

Time-Course of Effects in the Region of Interest

To examine the time-course of effects, I fit separate linear mixed effects models to sample level data for the left-lateralised occipitotemporal region of interest, with variables coded as described for the planned analysis. For feasibility, data were downsampled to 256 Hz, and the models did not estimate random slopes. To account for variability between electrodes, and for per-participant differences in topography, random intercepts were estimated for each combination of participant and electrode. In *lme4* syntax, the model formula was specified as:

```
amplitude ~ 1 + congruency * predictability +
  (1 | participant_id) +
  (1 | participant_id:electrode_id) +
  (1 | image_id) +
  (1 | word_id)
```

The results (Figure 5.9) reproduced findings from the planned analysis, with the N1 peak becoming less negative-going as predictability increases, but only for picture-incongruent words. Indeed, an effect of congruency at the lowest level of predictability first emerged during the N1's onset and peak, with more negative amplitudes for picture-incongruent words than for picture-congruent words, while the effect of congruency for more predictable words emerging later, after the N1 peak. The analysis also revealed effects during the N1's offset, with higher predictability eliciting more sustained negative amplitudes for picture-congruent words, but more positive amplitudes for picture-incongruent words, while the offsets for picture-congruent and -incongruent words were much more similar at lower levels of predictability. An additional predictability-congruency interaction also emerged after the N1, peaking at around 400 ms (likely resulting from effects in the N400 component) in the opposite direction to that observed for the N1's offset.

To better understand the interaction, I also examined the time-course of the effect of predictability for picture-congruent and -incongruent words separately (i.e., simple effects; Figure 5.10). This showed more clearly that predictability reduced amplitudes in the N1 for picture-incongruent words, but increased amplitudes for picture-incongruent words. This difference peaked around 225 ms, but reversed in direction after 300 ms. It is of note that the timing of the observed effects were later than originally anticipated (the planned analysis was limited to $\leq$200 ms). Nevertheless, the model intercept clearly shows that these effects peaked during the N1's offset, as the component overall peaked later than that observed in chapter 4.

For comparison to previous studies that examined effects of prediction bilaterally, I also analysed effects on right hemispheric electrodes (Appendix C.8), revealing no clear congruency-predictability interaction prior to 300 ms. I similarly examined whether the observed effects interacted with word frequency, finding that while there may be a main effect of word frequency on the N1, no clear interaction was observed between frequency and congruency or predictability (Appendix C.8).

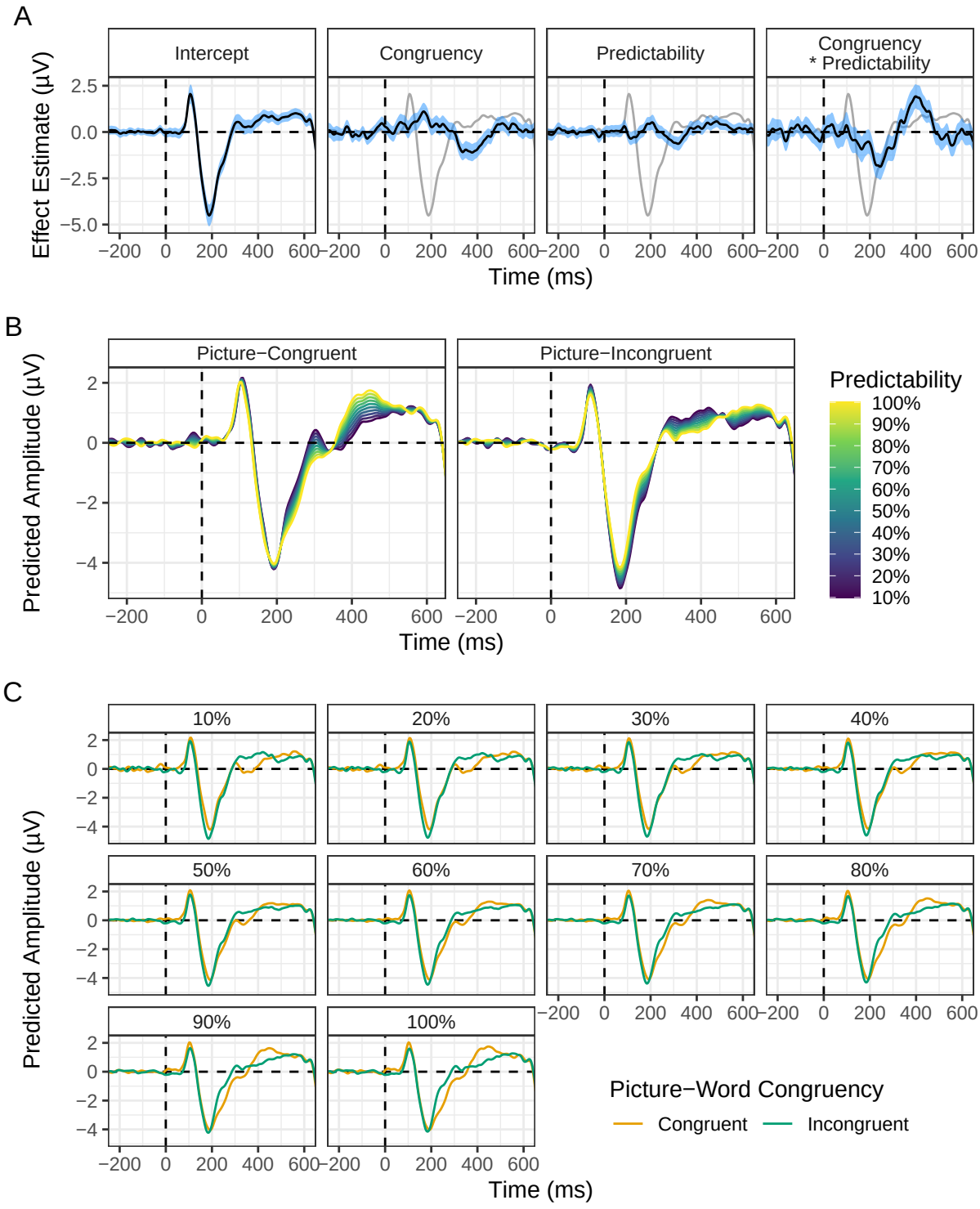


Figure 5.9: Time-course of fixed effects from the sample-level analysis of the left-lateralised occipitotemporal region of interest. (A) Time-course of fixed effects estimates, with shaded regions depicting 95% confidence intervals. The model intercept (reflecting average amplitudes at the lowest level of predictability) is depicted as a grey line on each panel to provide a reference for the timing and magnitude of effects. (B) Fixed-effect predictions for picture-congruent and -incongruent words at levels of predictability from 10 to 100%, in steps of 10%. (C) Same data as (B), but split by predictability rather than congruency.

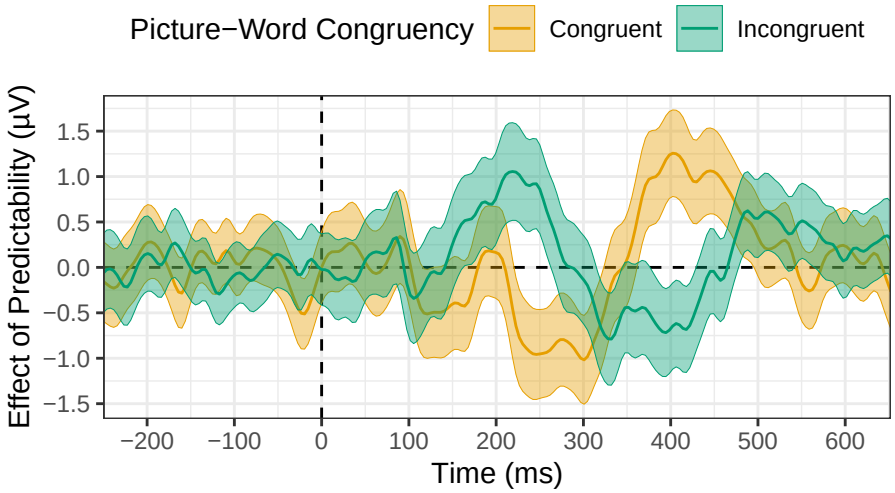


Figure 5.10: Time-course of the effect of predictability (i.e., the difference between the ERPs predicted for words at the maximum and minimum levels of predictability) for picture-congruent and -incongruent words. Central lines depict effect estimates, derived from sample-level models that were coded such that the model intercept lay at the respective levels of picture-word congruency. Shaded areas depict 95% confidence intervals of model estimates.

Behavioural Results

I analysed RTs, to examine whether the pattern of effects was similar to that observed for the behavioural validation experiment (section 5.2.1). I fit a Bayesian distributional shifted log-normal model, estimating the same model formula as that described for the behavioural validation experiment for all shifted log-normal parameters (μ, σ, and δ), using prior distributions based on the posterior distributions from the behavioural validation experiment (full details are presented in Appendix C.7). Priors for the behavioural analysis of the EEG experiment were not exact replicas of the validation experiment's posteriors, but were rather specified with greater uncertainty than that observed in the validation experiment's posteriors. I decided to specify this uncertainty because of key differences in the task demands; participants in the validation experiment could respond to stimuli with no lower limit, whereas responses were only permitted in the EEG experiment 500 ms after stimulus presentation. As a result of the additional time for participants to consider their responses, and because RTs were measured from the time point at which the stimulus changed colour, I reasoned that (1) responses would be faster overall in the EEG experiment (reflected in a reduced prior for the δ parameter intercept), and (2) effects observed in the validation experiment's RT data would be likely smaller in the EEG experiment's RT data. Results revealed that, although the effects were smaller than in the validation experiment, the main finding was replicated, with low predictability eliciting later RTs for picture-congruent words, to a greater extent than it does for picture-incongruent words (Figure 5.11A). RTs from the EEG experiment also replicated the difference in spread between picture-congruent and -incongruent RTs at low levels of predictability, with the congruency conditions showing more similar spread in RTs as predictability increases (Figure 5.11B), though again this effect was smaller for RTs in the

EEG experiment than it was for RTs in the validation experiment. Conversely, the difference in shift observed between picture-congruent and -incongruent words at high predictability in the validation experiment (Figure 5.2B) was not observed in the EEG experiment.

I similarly analysed accuracies in the picture-word task. I fit a logit-link binomial Bayesian generalised linear mixed effects model (GLMM) to accuracy data, using the same maximal mixed effects formula as that described for the planned analysis. All fixed effect prior distributions were specified to be flat, with the exception of the model intercept. As I expected overall accuracy to be very high, I specified the prior distribution for the fixed effect intercept as $\sim N(4, 1)$, where logit 4 would be equivalent to an average accuracy of .982. Priors for the *SD*s of random effects distributions were drawn from Student's *t* distributions with 3 degrees of freedom, μ of 0, and σ of 2.5. Prior distributions for all correlations were flat (between -1 and 1). The model was fit via *brms*, with 5 chains each sampling for 10,000 iterations (5,000 warmup). The *adapt_delta* parameter was set to .9, and the maximum tree depth (*max_tree_depth*) was set to 10. Results (Figure 5.12) revealed a main effect of predictability with higher accuracy at higher levels of predictability. An interaction with congruency was also observed, where predictability had a larger effect for picture-congruent than for picture-incongruent words, while accuracy remained more consistent across predictability for picture-incongruent words.

5.5.3 Exploratory Localiser Analysis

I also analysed results from localiser task, examining the full time-course of stimulus effects on ERP amplitudes, and patterns of RTs and accuracies.

Time-Course of Effects in the Region of Interest

I analysed the full time-course of stimulus effects in the localiser task, for right- and left-hemispheric occipitotemporal regions of interest. Specifically, I fit per-sample (256 Hz) linear mixed effects models via *lme4*, estimating models with the following formula:

```
amplitude ~ 1 + (false_font + noise) * hemisphere +
  (1 | participant_id) +
  (1 | participant_id:electrode_id) +
  (1 | match_set) +
  (1 | item_id)
```

Here, *false_font* and *noise* were deviation-coded variables, comparing the two nonword conditions to the null condition of words (i.e., BACS-font nonwords, and phase-shuffled words, respectively). In this way, the fixed effect slopes represented the difference between words and each non-lexical stimulus type. The deviation-coded variable, *hemisphere*, distinguished observations in the left (*hemisphere*=-.5) and right (*hemisphere*=.5) hemisphere. The *match_set* variable uniquely identified each triplet of matched items (see section 5.2.2). As

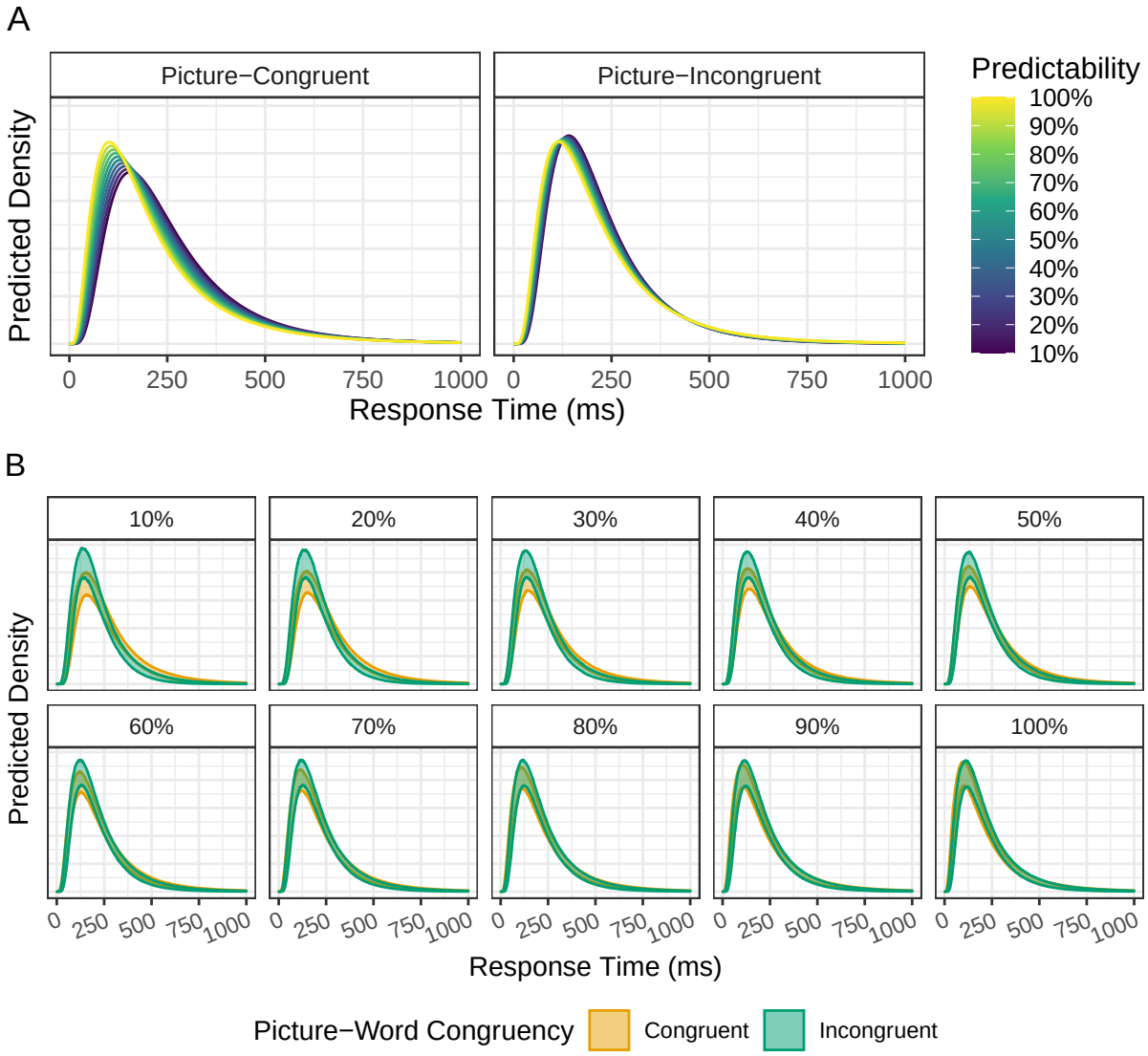


Figure 5.11: Fixed effect predictions of RT distributions in the EEG experiment. Figure layout is identical to that described for the validation experiment RTs in Figure 5.2, except that the axis limits for RTs are here limited to $\leq$1,000 ms. Unlike the validation experiment, where RTs reflect latency from stimulus presentation, RTs here reflect latency from a colour change in the stimulus, that occurred 500 ms after stimulus presentation (Figure 5.6).

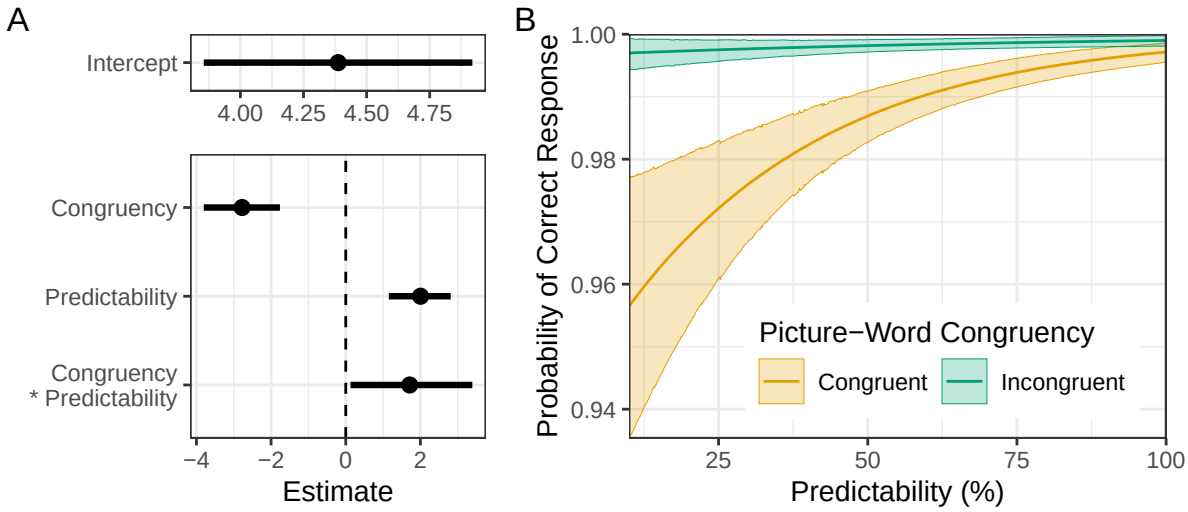


Figure 5.12: Fixed effect results for the analysis of accuracies in the picture-word task during the EEG experiment. (A) Fixed effect logit estimates, where points depict median estimates and whiskers depict the extent of 89% HDIs. (B) Model-predicted accuracies, for all levels of predictability in each congruency condition, where the central lines depict median estimates, while the shaded areas depict the extent of 89% HDIs.

in the sample-level analysis of the picture-word task, random intercepts were also estimated for each combination of participant and electrode (*participant_id:electrode_id*), and random slopes were excluded for feasibility.

Results (Figure 5.13) revealed that differences between words and phase-shuffled words emerged clearly in the P1 component, with more positive-going amplitudes observed for phase-shuffled words. Differences between words and false-font nonwords, meanwhile, remained small until later, in the N1. N1 components were more negative-going for false-font stimuli than for phase-shuffled words, for ERPs in both hemispheres. Both positive-going and negative-going ERP components elicited by words were overall more positive in amplitude for the right hemispheric occipitotemporal electrodes (i.e, the P1 was more positive-going, and the N1 less negative-going, in the right hemisphere); the N1 elicited by word stimuli was left-lateralised. An interesting stimulus-hemisphere interaction was observed, wherein ERPs elicited by words showed N1 peak amplitudes most similar to false-font stimuli in the left hemisphere, but most similar to phase-shuffled words in the right-hemisphere. Similar differences in timing were observed in the N1 peak for stimuli across both hemispheres, with phase-shuffled words peaking first, followed by false-font stimuli, and then words. Stimulus effects in occipitotemporal electrodes after the N1 were more consistent across hemispheres, with phase-shuffled words showing the most positive amplitudes, followed by false-font stimuli, which in turn elicited more positive amplitudes than words did, although the difference between words and phase-shuffled words was larger, post-N1, in the right hemisphere. The post-N! difference between words and false-font nonwords, meanwhile, did not interact with hemisphere except for a brief period around 250 ms.

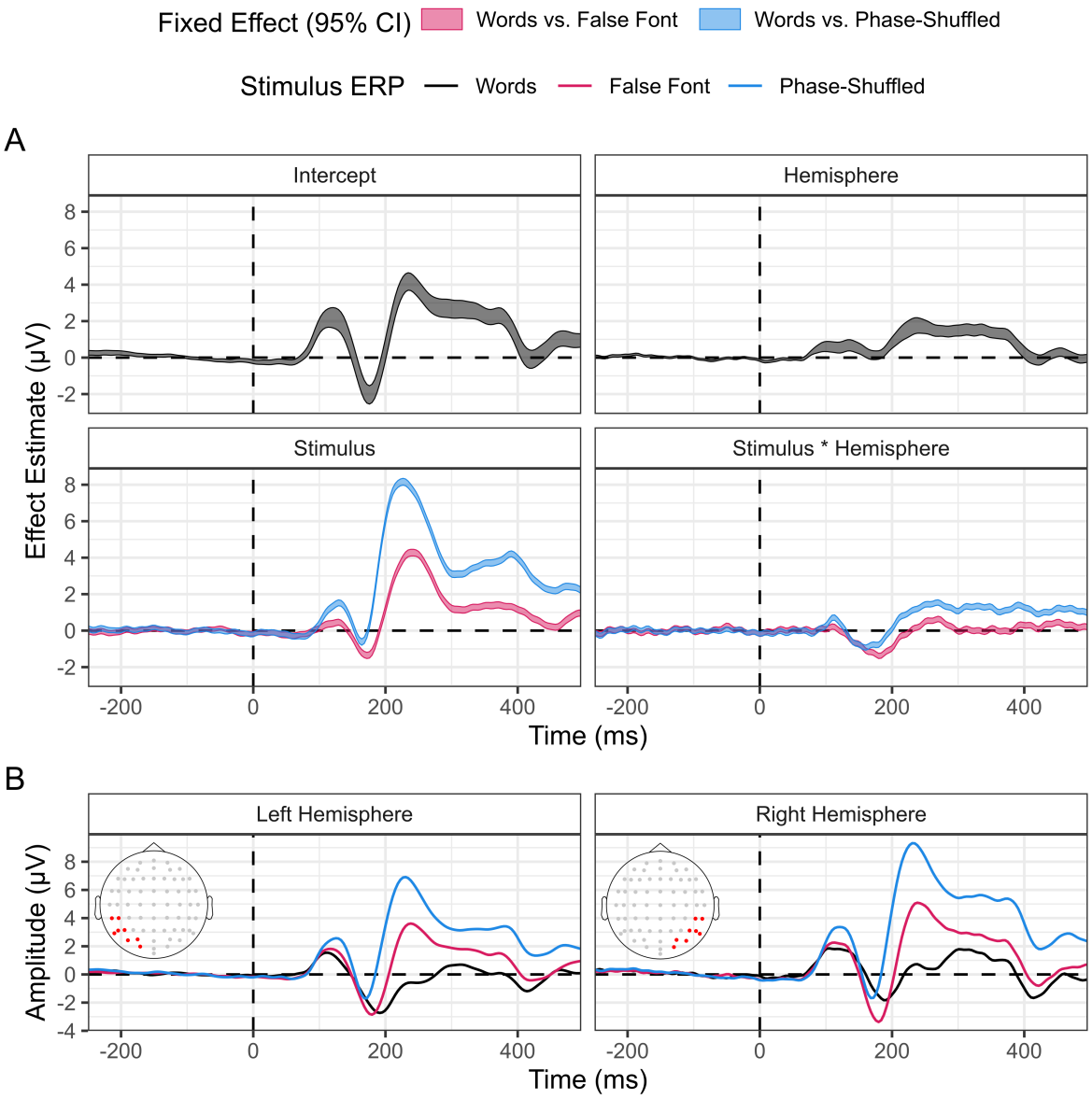


Figure 5.13: Fixed effect results for ERPs in the localiser task. (A) Fixed effects estimates for each time point, with the shaded areas depicting 95% confidence intervals. (B) Model-derived predictions for ERPs of left- (left) and right-hemispheric (right) occipitotemporal electrodes.

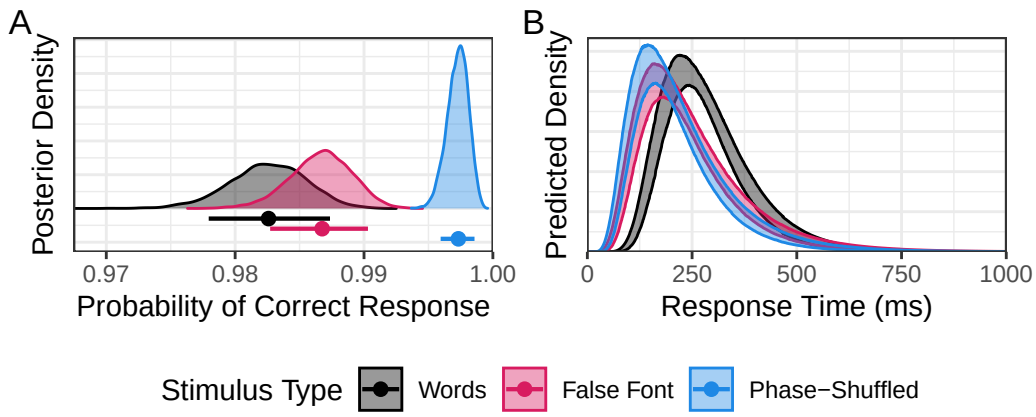


Figure 5.14: Fixed effect predictions for behavioural outcomes in the localiser task. (A) Posterior distributions for accuracies in the localiser task, where points below densities depict median posterior estimates, while whiskers depict 89% HDIs of posterior samples. (B) Predicted RT distributions, where the shaded regions depict 89% HDIs of posterior samples (density values on the *y*-axis begin at 0).

Behavioural Results

I also analysed stimulus effects on lexical decision RTs and accuracies. Specifically, I fit a logit-link binomial model to trial-level accuracies, and a distributional shifted log-normal model to RTs, with maximal random effects structures. Results (Figure 5.14) revealed that responses were fastest and most accurate for phase-shuffled words. False-font stimuli were responded to somewhat faster. Conversely, responses were slowest and least accurate for word stimuli. RT distributions were similar for false-font and phase-shuffled words, though accuracies for false-font stimuli were closer to those observed for words. Behavioural results overall suggest that participants found it easy to reject phase-shuffled words in lexical decision, but found it relatively more difficult to reject false-font stimuli. Full model details are described in Appendix C.10.

5.6 Discussion

In the present study, a clear congruency-predictability interaction was observed in the N1 ERP component, the timing of which is consistent with an account of word recognition that involves an early sensitivity to predictions, and suggestive of some form of top-down modulation. However, the pattern of effects was ostensibly inconsistent with an account of top-down modulation of orthographic processing that is based on predictive coding, suggesting that such an account may not fully explain prediction effects on the N1. Effects were also analysed in the localiser task, revealing clear effects of stimulus on N1 amplitude and timing, including a stimulus-hemisphere interaction in the N1.

5.6.1 Evidence Consistent with Top-Down Modulation

The planned and exploratory analyses of the picture-word task revealed a clear congruency-predictability interaction in the N1, the timing of which is consistent with top-down modulation. However, the pattern of effects that characterised this interaction diverged markedly from the pattern of effects hypothesised. At the highest levels of predictability, picture-congruent and -incongruent words showed a clear difference in the N1's offset, with more negative amplitudes elicited by picture-congruent words, rather than for picture-incongruent words. Furthermore, at the lowest levels of predictability, there was a smaller effect of congruency in the N1's onset and peak, with more negative-going N1s elicited by picture-incongruent words. This finding stands in contrast to findings from sentential studies, which generally suggest that less negative-going N1 components are observed for words that are orthographically congruent with readers' predictions (e.g., A. E. Kim & Gilley, 2013; Kretzschmar et al., 2015; Sereno et al., 2003).

Some studies have previously reported that predictions can elicit more negative-going N1 components when observed word forms are congruent with readers' predictions, though these studies are in the minority. For instance, Sereno et al. (2019) observed that for high frequency words, amplitudes during the N1 window were more positive over the left hemisphere for highly predictable words than for words of low predictability. However, results from this study also included a finding of more negative amplitudes for highly predictable words in the *right* hemisphere, which was not observed in the present study (Appendix C.8). More negative-going amplitudes for highly predictable items were also reported by Penolazzi et al. (2007), although there the effect was located centroparietally rather than occipitotemporally, such that the overall direction of ERPs, and likely the effects within the ERPs, was reversed.

One key difference between the present study and much previous work that has examined effects of prediction on early visual word processing is that the present study did not bias expectations via sentential stimuli. This difference may in part account for the disparity observed between the present study's results and those reported in previous investigations. However, if this is the case, it is not clear which features of the present study's design would have caused the disparity, or how they may have caused it. For instance, one possibility could be that, rather than using a sentence to bias expectations, as most previous investigations have, instead preceding the target word with an image may have altered the dynamics of visual word processing or predictive processes in some manner (see section 7.3 of the General Discussion for further discussion of this possibility).

Notwithstanding the disparities between the present study's findings and both the hypothesis and previous findings, the observed congruency-predictability interaction in the N1 is suggestive of an influence of higher-level information on early stages of word recognition that are associated with orthographic processing. The neural dynamics responsible for the observed pattern of results are difficult to delineate from the evidence presented here, though I briefly consider some possible explanations and their implications. First, the present study's findings are difficult to reconcile with an account of top-down modulation of orthographic processing during the N1 that is based on a simplistic implementation of predictive coding, according to which one would expect more negative-going N1s for picture-incongruent

words at the highest level of predictability, as less of the orthographic information would be "explained away" by the reader's predictions (A. Clark, 2013; Eisenhauer et al., 2022; Gagl et al., 2020). One possibility is that while the brain activity underlying N1 amplitude may scale with bottom-up orthographic prediction error (Gagl et al., 2020), top-down predictions of orthographic content may not be implemented by simply altering the orthographic prior but by additional mechanisms that interact with bottom-up processing and produce divergent effects in the N1. Such a process may even occur in temporal and spatial proximity to orthographic processing while itself being functionally distinct from it. Alternatively, if top-down modulation *is* implemented via on-line alteration of an orthographic prior, this may be achieved in a manner that, counterintuitively, induces stronger predictions when predictability is low.

If not implemented via predictive coding mechanisms, alternative explanations for the observed pattern of effects could include that N1 amplitude, especially in the component's offset, scales with similarity to predictions, rather than error. Such a finding could represent predictability "sharpening" neural responses, rather than "explaining away" bottom-up input. More specifically, it could be that the more certain a prediction is (i.e., the higher the predictability of an image's name), the more sensitive the neurons that generate the N1 are to the predicted word form or its features. This interpretation concords with recent findings from Eisenhauer et al. (2022), where source-localised MEG showed that prime words led to the preactivation of orthographic and lexical-semantic information for predictable, but not unpredictable, target words, leading to greater occipitotemporal activity in response to the target word's presentation. Related fMRI findings have shown that presenting primes that appear embedded in the subsequent target word (e.g., *car* within *scar*) cause the target word to elicit greater activity in vOT (Z. Zhou, Whitney, et al., 2019). Parallels can be drawn between such an explanation and related accounts of effects of task-driven modulation of activity during the N1 and in vOT, which suggests that higher task demands elicit greater occipitotemporal sensitivity to word form information (Y. Chen et al., 2013, 2015; Qu et al., 2022; Segalowitz & Zheng, 2009; Strijkers et al., 2015), perhaps due to a heightened sensitivity to orthographic information (Qu et al., 2022). According to such an account of the present study's findings, higher predictability led to more negative N1 amplitudes for picture-congruent words (matching predictions), as the predicted orthographic features, perception of which was sharpened *selectively*, were present in the observed picture-congruent word form and thus facilitated. Meanwhile, lower predictability would have reduced sensitivity to the difference between picture-congruent and -incongruent word forms or orthographic features. However, such an explanation would also need to account for why effects of predictability emerged earlier for picture-incongruent words, in the N1's peak, than they did for picture-congruent words, in the N1's offset, or why a difference between picture-congruent and -incongruent words may have emerged in the N1's peak at the lowest effect of predictability. That periods within the N1's window show differential sensitivity to higher-level information is an emerging finding in the literature, with the present study replicating previous reports that effects resulting from top-down modulation emerge and peak in the N1's offset period (F. Wang & Maurer, 2017, 2020).

Consequently, while the timing of the congruency-predictability interaction identified in this

study is consistent with a top-down influence of higher-level information on early processing in the N1, an ERP component associated with orthographic processing, the *direction* of the observed effect is more puzzling. In contrast to the hypothesised effect of predictability leading to less negative amplitudes for picture-*congruent*, and not for picture-incongruent, words, words preceded by more predictable images elicited less negative amplitudes when picture-*incongruent*, while the opposite effect was observed for picture-congruent words. There were also differences in timing of effects, with the effect of predictability emerging for picture-incongruent words during the N1's peak, but emerging for picture-congruent words later, in the N1's offset. While I have briefly speculated about possible explanations for the study's findings, it is important to note that experimental evidence alone cannot provide evidence capable of evaluating and distinguishing between possible explanations for such a pattern of effects. As Guest and Martin (2021) argue, rather than merely describing linguistically how cognitive processes relate to observations of neural activity, evaluating and comparing the explanatory power of accounts would require comparisons of how well *computational implementations* of these theories predict neural activity and behaviour. Moreover, examination of the neural and behavioural consequences predicted by computational implementations of theories could guide the design of future studies capable of providing meaningful evaluation of theories, especially in research into top-down modulation (Ramsey & Ward, 2020). As such, the findings of the present study are consistent with top-down modulation of orthographic processing, or processing that is temporally and spatially proximal to orthographic processing, and provide a basis for further research to more specifically delineate and examine evidence for the cognitive and neural mechanisms underlying such influences. In particular, I argue that more computationally explicit models of the processes involved are required to evaluate whether theories accurately account for word recognition processes.

5.6.2 Bottom-Up Sensitivity to Orthography

Exploratory analyses of the localiser task, which was included primarily to identify per-participant maximal electrodes and time-points for the picture-word analysis, replicated previous findings of sensitivity to orthography in the N1. Results revealed differences in timing, average amplitude, and (in the right hemisphere) peak N1 amplitude, in ERPs elicited by words and false-font stimuli; N1 components observed for false-font stimuli peaked earlier bilaterally, with less-negative average amplitudes in the left hemisphere, but with more negative-going amplitudes in the right hemisphere, relative to word stimuli. This finding is broadly consistent with existing evidence for such sensitivity to orthography in the N1 (Bentin et al., 1999; Brem et al., 2018; Holcomb et al., 2002; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019). In particular, previous research has generally also found that words elicit more negative-going left-hemispheric N1 components than false-font stimuli do (Appelbaum et al., 2009; Bentin et al., 1999; Eberhard-Moscicka et al., 2016; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019; J. Zhao et al., 2014), and it is this directional prediction on which the identification of maximal electrodes and time-points in the localiser task was based (subsection 5.5.1).

As expected, the word-versus-phase-shuffled difference was in the same direction as the

word-verus-false-font difference, though with a larger effect size. Robust sensitivity to the difference between words and words with scrambled phase is consistent with fMRI findings of greater activity in vOT for words relative to phase-randomised words (Rauschecker et al., 2012; Rodrigues et al., 2019; White et al., 2019; Yeatman et al., 2013). However, a difference between words and phase-shuffled words was in the present study observed prior to the N1, in the P1 component (whereas words and false-font stimuli elicited more similar P1 components). Such a finding is consistent with fMRI findings of sensitivity to the words-versus-phase-shuffled difference in regions more posterior than the typical visual word form area (VWFA) location (Rodrigues et al., 2019; Yeatman et al., 2013). It has been argued that areas posterior to the VWFA of vOT are sensitive to orthographic information, such as the mid-fusiform cortex (Woolnough et al., 2021), and indeed, even neurons in the primary visual cortex can become tuned for geometric features of shapes via top-down influences (McManus et al., 2011). However, early, posterior sensitivity to the difference between words and phase-shuffled words, observed earlier and in regions more posterior than word-nonword (or word-false-font) differences are, may arise from non-orthographic, lower-level visual differences between words and phase-shuffled words. For instance, permuting or randomising images' distributions of phase necessarily alters their phase congruency, which can be used as an effective indicator for edge and feature detection in image analysis (Kovesi, 2003). Evidence suggests that human visual processing is sensitive to phase congruency, or information that correlates with it, in areas as posterior and early as the primary visual cortex (Perna et al., 2008). In sum, sensitivity to the difference between words and phase-shuffled words in the P1 component cannot necessarily be interpreted as early sensitivity to orthographic information; if researchers wish to isolate orthographic processing, then comparisons between words and false-font or nonword stimuli may be more appropriate, as these are more closely matched on low-level features.

An interesting stimulus-by-hemisphere interaction was observed in the present study's localiser task, wherein the N1 responses to false-font stimuli were more negative-going, while responses to words were *less* negative-going, in the right hemispheric N1. Such hemispheric interactions have been observed previously. Maurer, Brandeis, et al. (2005) observed that while for the left hemisphere, average N1 amplitudes were more negative for words than they were for symbols, in the right hemisphere, the effect reversed, with more negative N1 amplitudes for symbols than for words. A related interaction was reported by Bentin et al. (1999), where orthographic stimuli (words, nonwords, pseudowords) elicited more negative-going N1 components over the left hemisphere than non-orthographic stimuli (symbols, simple shapes) did, whereas no significant difference was observed over the right hemisphere. The interaction between the word-false-font difference and hemisphere was similarly reported for typical readers in Pleisch et al. (2019), in a simultaneous EEG-fMRI analysis. As a result, the stimulus-hemisphere interaction observed in the present study's localiser task replicates similar interactions reported in existing literature.

5.6.3 On the Content of Predictions

The picture-word paradigm employed in the present study was designed to elicit predictions for specific visual word forms without relying on sentential contexts. One key issue with the paradigm, also applicable to comparable designs employed in previous studies, is that it does not provide insight into the content of predictions that participants actually form. For instance, given as context an image of a vaccum cleaner, does the participant form a specific prediction for either the word *vacuum* or *hoover*, or do they simultaneously predict both word forms? This distinction could be expected to have important implications on the cognitive mechanisms by which predictions influence early processing. For instance, a word of low cloze probability (e.g., 10%) may be expected to elicit responses functionally similar to an unpredicted word if on (e.g.) 90% of trials, readers' predictions are devoted to an entirely different word. Conversely, if readers predict multiple possible words simultaneously, a word of low cloze probability may instead simply elicit a smaller or less specific effect of predictability than a word of greater cloze probability would. That vOT could support simultaneous, parallel predictions of orthographic features for more than one word form is plausible, given recent evidence showing that vOT can represent bottom-up information of multiple presented word forms in parallel (White et al., 2019). Further research could examine the content of predictions in more detail to better understand how it relates to effects of predictability.

It is furthermore relevant to ask what information is being functionally predicted. For instance, if semantic predictions are being recoded into predictions of orthographic features, are these features predicted at the level of the word form, characters, or sub-character features? Would, as evidence from A. Kim and Lai (2012) suggests, a semantically irrelevant word that is nonetheless orthographically similar to a predicted word benefit from a top-down prediction in orthographic processing? This is an additional instance in which computational implementations and models of orthography and orthographic processing have the potential to improve insight into the processing of it, and its sensitivity to top-down modulation. For instance, with a computational description of orthography that can account for similarity at multiple levels, it could become possible to relate features like orthographic similarity more precisely to effects of top-down modulation on orthographic processing.

5.6.4 Summary

The results from this study are clearly suggestive of higher-level information influencing early visual word recognition processes during the N1 component, likely via top-down modulation. Given the N1 component's robust and replicable sensitivity to orthographic information, it is likely that top-down modulation of the N1 component reflects early interactions between orthographic and higher-level information. However, the present study and existing literature provide insufficient evidence to delineate a specific description of the manner in which such interactions may occur and produce the observed effects. Indeed, the observed pattern of effects was inconsistent with a simplistic predictive coding account of top-down modulation. I suggest that future research would benefit from being guided by hypotheses that are more directly informed by computational implementations of theories.

Chapter 6

SCOLD: Sub-Character Orthographic Levenshtein Distance

6.1 Introduction

An account of how the human reader processes orthographic information necessarily constitutes a vital component in any model of reading and visual word recognition (e.g., Besner & Smith, 1992; Coltheart et al., 1977; Coltheart et al., 2001; Dehaene et al., 2005; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1981; Reichle et al., 2003; Seidenberg & McClelland, 1989; Whitney, 2001), often constituting the model's focus (e.g., Adelman, 2011; Gagl et al., 2022; Gomez et al., 2008). This focus on orthography reflects its essential role in reading, bridging lower-level visual processing, constituting the very first processing stages in the brain, to higher-level semantic comprehension. A complete model of orthographic processing should be able to describe how visual shapes are decoded into linguistic units in computational terms, providing falsifiable predictions about the nature of such processes that can be tested via comparisons to neural or behavioural correlates of reading. Measures describing and derived from orthography are therefore vital for building and testing such theories and computational models of visual word recognition. Perhaps the most important and widely applicable measure that can be derived from orthographic descriptions is *orthographic similarity*, quantifying the distance between word forms or their components. Orthographic similarity plays a central role in testing models of orthographic processing, providing valuable insight into the nature of orthographic representations in the brain (Norris & Kinoshita, 2012). However, existing empirical measures of orthographic similarity are greatly limited in their resolution, failing to account for sub-character complexities. In this chapter, I propose and test a pixel-based measure of orthographic similarity that incorporates sub-character information: Sub-Character Orthographic Levenshtein Distance (SCOLD). I further examine whether the inclusion of geometric operations, of translation, rotation, rescaling, and mirroring, improves the measure's explanatory power for correlates of orthographic similarity. I show that the measure can capture effects of sub-character complexities on orthographic processing, and that whether SCOLD outperforms existing measures can reveal insight into cognitive mechanisms underlying orthographic processing.

Orthographic similarity refers to how alike written representations of language are, with

descriptions ranging in granularity from entire word forms to characters and sub-character features. Orthographic similarity has been shown to have far-reaching consequences in language processes spanning multiple levels of processing, strongly predicting low-level effects like letter (Mueller & Weidemann, 2012) and word confusability (Kondrak & Dorr, 2006), even extending to reported effects in *auditory* word recognition (Chéreau et al., 2007; Tanenhaus et al., 1980), purportedly due to a rapid and functionally meaningful activation of orthographic information during spoken word recognition (Muneaux & Ziegler, 2004; Perre & Ziegler, 2008; Salverda & Tanenhaus, 2010). Words' orthographic neighbourhood densities, reflecting their similarity to the lexicon, also show effects in word recognition (Andrews, 1997; Carreiras et al., 1997; Yap & Balota, 2009), that similarly permeate multiple stages of processing (Hauk et al., 2006; Holcomb et al., 2002), and similarly transcend the visual domain to impact word recognition in other modalities (Muneaux & Ziegler, 2004; Ziegler et al., 2003). Correspondingly, effects of orthographic similarity and orthographic neighbourhood density have had considerable influence on models of reading and visual word recognition. As an example, an important observation in visual word recognition is that nonwords formed by transposing two adjacent characters of a real word (e.g., *judge* - *jugde*) are often misread as, and show similar behavioural effects to, real words (Davis, 2010). To accommodate this observation, models have included elegant features such as flexible (Adelman, 2011; Whitney, 2001) or noisy (Davis, 2010; Gomez et al., 2008) coding of letter positions. In the context of this thesis, effects such as orthographic priming (e.g., Eisenhauer et al., 2022; Ferrand & Grainger, 1994; Frisson et al., 2014; Masson & MacLeod, 2002) and interactions between orthographic and higher-level lexical or semantic processes (e.g., Y. Chen et al., 2015; Dikker & Pylkkanen, 2011; A. Kim & Lai, 2012; Pecher et al., 2009; Pecher et al., 2005; Rodd, 2004; P. Yao et al., 2022; J. Zhao et al., 2019) are of particular interest, and an understanding of orthographic similarities and neighbourhoods may provide insight into the mechanisms underlying such effects. For example, the relationship between predictability and top-down modulation (see chapters 4 and 5) may have a basis in orthographic neighbourhoods densities, as low-predictability expectations for a potentially large set of orthographically diverse word forms may result in little to no top-down modulation, relative to more targeted predictions, because the predicted word forms would be orthographically similar to a larger number of irrelevant word forms.

The most widely used measure of word form similarity is orthographic Levenshtein distance (OLD), which measures the minimum number of character insertions, deletions, or substitutions required to convert one string into the other, implicitly assuming that the costs of these operations are invariant across different characters (Norris & Kinoshita, 2012; Yarkoni et al., 2008). This measure has been further applied to calculate orthographic neighbourhood density (Orthographic Levenshtein Distance 20; OLD20), which describes how similar a word form is to others in the lexicon (specifically, the OLD between the word form and its 20 closest neighbours), and which accounts well for word recognition behaviour (Siew, 2018; Yarkoni et al., 2008). Neighbourhood measures calculated from Levenshtein distance quantify a more robust description of orthographic similarity than would be permitted by the previously most widely used measure, Coltheart's N (Coltheart et al., 1977). Specifically, Coltheart's N defines

a word form's neighbourhood coarsely, as the number of words at a distance of just 1 character (i.e., number of words that can be generated by exchanging one letter in the target word). By this limited definition, it is implied that word form similarity is binary; the word forms *price* and *prize* would be orthographically similar, but *price* and *precise* would be as dissimilar as *price* and *skunk* are. The Levenshtein distance of *price* and *precise*, meanwhile, is equal to 3, reflecting their shared letters, whereas *price* and *skunk* have a Levenshtein distance of 5. Thus, Levenshtein distance provides greater discriminatory power (Yarkoni et al., 2008) and is at present accepted as the gold standard in the field. Correspondingly, computational models of visual word recognition either recognise that sub-character features must be integral to orthographic processing yet fail to explicitly describe how they are processed (e.g., Whitney, 2001), or use a simplified yet computationally convenient artificial character set (typically that proposed by Rumelhart & Siple, 1974) that only loosely relates to real-world characters.

However, as previously mentioned, because the operations that OLD is calculated from are invariant to the identities of characters which are being inserted, deleted, or substituted, the measure overlooks sub-character similarities or complexities. Consider, for example, that the word forms *price* and *pride* may be more orthographically similar than *price* and *prize* are, if the sub-character features of *c* and *d* are more similar than *c* and *z* are. According to OLD, however, these three word forms are all at an equal distance of 1 from one another. A solution could be to integrate sub-character information into the Levenshtein distance metric. For instance, character substitutions could be weighted by the distance between the characters being exchanged. While Yarkoni et al. (2008) examined the impact of altering the relative weights of all three operations globally, identifying no improvement over an equal weighting scheme, they did not examine whether actively altering weights in response to relevant character information could improve predictions of correlates of orthographic similarity. More recently, H. Kim (2021) outlined and tested the performance of such a weighting scheme for Levenshtein distance operations, including weighting character substitutions by subjective character similarity judgements (i.e., average ratings reported by Simpson et al., 2013), or by distance between keyboard keys (to reflect the probability of mistyping a character). H. Kim (2021) did not examine the resultant measures' relation to cognitive processing or its correlates, but evaluated the approach in terms of computational efficiency in tasks of string matching that OLD is typically applied to, showing that the approach outperformed traditional Levenshtein distance measures despite the increased overhead computational cost. In this chapter, I suggest that an approach extending that employed by H. Kim (2021) could be applied to word recognition research to provide a more fine-grained and powerful description of orthographic similarity.

Implementing sub-character granularity in measures of orthographic similarity should only be expected to confer more explanatory power, relative to comparable measures which ignore sub-character features, insofar as the cognitive mechanisms driving observed neural and behavioural outcomes are themselves sensitive to such fine-grained information. Evidence for such sensitivity to sub-character similarity has been observed in behavioural measures of priming studies, with effects emerging in early (pre-300 ms) stages of word recognition (Gutiérrez-Sigut et al., 2019; Marcet & Perea, 2018). Similarly, recent findings

have demonstrated that neural representations of individual characters' features can be reconstructed from functional magnetic resonance imaging (fMRI) of Visual Word Form Area (VWFA) activity and complementary electroencephalograpy (EEG) recordings of the N1 ERP component. This has been shown to be the case not only when characters are presented in isolation (Schoenmakers et al., 2013; Shen et al., 2019), but also when presented within words (Ling et al., 2019). Furthermore, orthographic prediction error, parameterised as the difference between top-down predictions and observed (bottom-up) visual word forms on the level of pixels (thereby including character-level information), has been found to account for patterns of activation seen in the occipitotemporal cortex as early as 150 ms (Gagl et al., 2020). Similarly, pixel-based descriptions of orthographic neighbourhood density have been shown to meaningfully describe orthographic similarities of Chinese characters (Sun et al., 2018), where their logographic nature makes a traditional character-based measure of OLD particularly difficult to implement.

On the other hand, some studies have reported no sensitivity to character similarity, or a sensitivity to character similarity when letters are replaced with visually similar digits, but not when replaced with visually similar letters (Kinoshita et al., 2014). Nevertheless, these divergent findings can be reconciled with an account of character similarity effects, and of orthographic processing more broadly, that stresses the importance of task context, with the observer dynamically adapting to task demands to process orthographic similarity when task- or goal-relevant (Kinoshita et al., 2015). Indeed, a possible application for a measure of orthographic similarity that is particularly relevant to the present thesis is in computational models of top-down modulation of orthographic processing: interpreting and developing further research to delineate the cognitive mechanisms underlying results observed in chapter 5 may be made possible with more powerful operationalisation of orthographic similarity, and with more computationally explicit models of orthographic processing and its sensitivity to top-down modulation (which SCOLD could support). Indeed, the two main explanations that I considered for results in that chapter, respectively, predictive coding and "sharpening" of sensitivity, would both predict that the influence of top-down modulation on neural correlates of orthographic processing varies with orthographic similarity between observed and predicted word forms (although predictions are in opposite directions). These accounts could be better evaluated and compared with more fine-grained measures of word form similarity. To summarise, a more fine-grained measure of orthographic similarity between word forms may be expected to better account for behavioural and neural correlates of orthographic processing, and could be used to provide insight into the nature of orthographic representation.

An important consideration, when calculating a measure of character and sub-character similarities, or a measure of word similarity which is sensitive to such information, is that the measure will be font-specific. This is because the similarity between characters will depend on their exact orthographic forms. For example, the similarity between the characters for *g* and *q* may be greater if the *g* were single-storey (as in Helvetica font g) rather than double-storey (as in Times font g). While objective measures of character similarity will be inherently font-specific in this way, this is not necessarily true of all neural and behavioural correlates of effects of orthographic similarity. It may be, for instance, that rather than in a

font-specific manner, orthography is processed at some levels in a font-general manner, as abstractions of character representations. This possibility is analogous to prototype accounts of perceptual categorisation, where categorisation is achieved by comparing perceptual information to a central-tendency prototype (Posner & Keele, 1968). However, the possibility of font-generality in character similarity effects does not invalidate the use of measures that are sensitive to character and sub-character information. Rather, if character similarity effects act in a font-general, or prototypic, manner, character similarity measures should try to quantify this font-general similarity, rather than a font-specific one. This could be done, for instance, by calculating an average representation for each character, and calculating similarities between these inferred prototypes. It is relevant, therefore, to examine whether measures of character similarity, or measures of word form similarity that are sensitive to such information, predict orthographic effects on behavioural or neural correlates in a font-specific or font-general manner.

In this chapter, I consider the possibility of extending the existing measure of Levenshtein distance to calculate a measure of word-level orthographic similarity that is sensitive to sub-character complexities. First, I propose a measure of character similarities that incorporates sub-character features, Jaccard similarity, and show that this can be applied to calculate character similarities using both an existing simplified framework for describing orthographies via an artificial character set (Rumelhart & Siple, 1974), and using a pixel-based approach that can be used to calculate the similarities of real-world characters. An exploratory analysis is then reported in which the proposed measure is considered as a predictor of subjective character similarity ratings (Simpson et al., 2013), showing that the measure aligns well with subjective ratings. It is also shown that these similarity effects may be font-specific, with peak model performance observed for Jaccard similarities derived from the same font that participants were judging. I then report the results of an online rating study expanding upon the Simpson et al. (2013) results, conducted to validate the finding of font specificity, though, interestingly, while the ability of Jaccard similarity to predict subjective ratings is replicated, the results were not consistent with font specificity in predicting subjective character ratings. I outline several approaches to integrating sub-character information into OLD, producing several variants of SCOLD. To compare the new word form similarity measures to classic OLD, orthographic neighbourhood densities are calculated from these sub-character-sensitive measures of orthographic similarity. I show that an orthographic neighbourhood measure which is sensitive to sub-character complexities is actually outperformed by traditional OLD20 in predicting neighbourhood effects in Lexical Decision Task (LDT) data, in the English Lexicon Project (ELP; Balota et al., 2007) and British Lexicon Project (BLP; Keuleers et al., 2012). However, results in predicting ERPs reveal differential sensitivity to sub-character information over time and space, with SCOLD neighbourhood measures particularly outperforming OLD20 values in predicting amplitudes of the occipitotemporal N1, suggesting that the calculated measures are capturing information that is relevant to orthographic processing. In sum, the proposed SCOLD measure, and the neighbourhood metrics derived from it, can provide a valuable description of sub-character orthographic similarities.

Figure 6.1: Estimates of character similarities between pairs of Rumelhart-Siple characters. (A) The fourteen segments used to form all Rumelhart-Siple characters. (B) Example fourteen-bit representations for Rumelhart-Siple characters - the presence or absence of each segment can be represented by a segment-specific bit. (C) Bit-wise Jaccard similarities for the Rumelhart-Siple characters.

6.2 Character Similarity

I estimated character similarities for all pairs of alphabetic characters, using a bit-wise approach for the Rumelhart-Siple character set (Rumelhart & Siple, 1974), and using a pixel-based approach for real-world fonts. I estimated similarity of pairs of characters as the Jaccard Similarity of their representations. For a pairs of characters, a and b, Jaccard similarity J is calculated as the size of their intersection divided by the size of their union:

$$J(a,b) = \frac{a \cap b}{a \cup b}$$

In the context of the bit-wise approach implemented for Rumelhart-Siple characters, Jaccard similarity therefore reflects the proportion of non-zero bits that occur in both characters. For the pixel-based approach, meanwhile, it reflects the proportion of non-zero pixels that are accounted for by an overlap between the characters.

6.2.1 Rumelhart-Siple Character Similarity

While many models of visual word recognition limit descriptions of orthography to the level of characters, often acknowledging the role of sub-character features but not explicitly modelling them (e.g., Whitney, 2001), many models do indeed incorporate sub-character features (e.g., Coltheart et al., 2001; Davis, 2010; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1981). However, such implementations have mostly relied on the highly limited and artificial typography first implemented by Rumelhart and Siple (1974). In this font, all alphabetic letters are, for computational convenience, represented with just fourteen segments which are binarily present or absent in any character. As a result, each character can be represented in the Rumelhart and Siple (1974) font in terms of a fourteen-bit binary sequence (Figure 6.1), and operations like calculating characters' orthographic similarity (Figure 6.1C) can be achieved via bit-wise comparisons. The character set captures the general shape of upper-case alphabetic characters, but does not include lower-case variants

The Rumelhart-Siple characters are an example of cognitive models imposing artificial restrictions to simplify real-world stimuli to make the model and its results more clearly interpretable (Johns et al., 2017). For instance, the characters are composed of a discrete set of clearly identifiable sub-character features where real-world characters are much more geometrically complex and variable. An alternative approach could be to describe the full orthographic variability observed in real-world characters, imposing no artificial constraints on character shapes or features.

6.2.2 Pixel-Based Character Similarity

One early approach to calculating the orthographic similarities between real-world letters was to find the maximum area of overlap between pairs of characters, while optionally permitting translation and rotation. For instance, Dunn-Rankin et al. (1968) showed that such an approach could be used to accurately identify neighbourhoods of characters that share key features but are distinct from one another, such as *p*, *q*, *b*, *d*, versus *n*, *u*, *m*, *h* - neighbourhoods of characters that are also reliably observed in analyses of objective effects of character similarity, such as letter confusability (Gervais et al., 1984; Kuennapas & Janson, 1969; Mueller & Weidemann, 2012; Tinker, 1928; Uttal, 1970).

In calculating pixel-based character similarity, I represented characters as binary matrices derived from TrueType font files, at 50-point font size. These matrices were constructed as images and converted to matrices using the *Pillow* (A. Clark, 2020) and *numpy* (Harris et al., 2020) libraries for Python (Van Rossum & Drake, 2009). I used a computerised method analogous to that employed by Dunn-Rankin et al., identifying the optimal overlap for any pair. I additionally controlled which geometric transformations were permitted for the optimal Jaccard similarity to be calculated. Specifically, I either permitted or did not permit translation, rescaling, rotating, and mirroring operations.

I applied this method to six separate fonts: Arial, Calibri, Consolas, DOS VGA ("More Perfect DOS VGA" - a modern recreation of the *code page 437* character set that participants were presented with in the ELP; Balota et al., 2007), Droid Sans, and Times New Roman.

Default Position Similarity

In the most simple case, pixel-based similarities were calculated for character pairs' default positions when centre-aligned horizontally (i.e., permitting neither translation, rescaling, rotation, nor mirroring; Figure 6.2). This approach provides included as a useful baseline to which the more complex geometric transformations can be compared. Default position similarities may also be psychologically meaningful however, especially for monospace fonts like Consolas where the consistent spacing provides a common reference frame for letter shapes. For example, lower-level processing may be translation-sensitive, while higher-level processing may be more translation-invariant. Jaccard similarity for default character positions is comparable to the method employed by Gagl et al. (2020), though whereas it is here calculated at the level of pairs of single characters, Gagl et al. (2020) compared entire word forms to the average of the full orthographic lexicon.

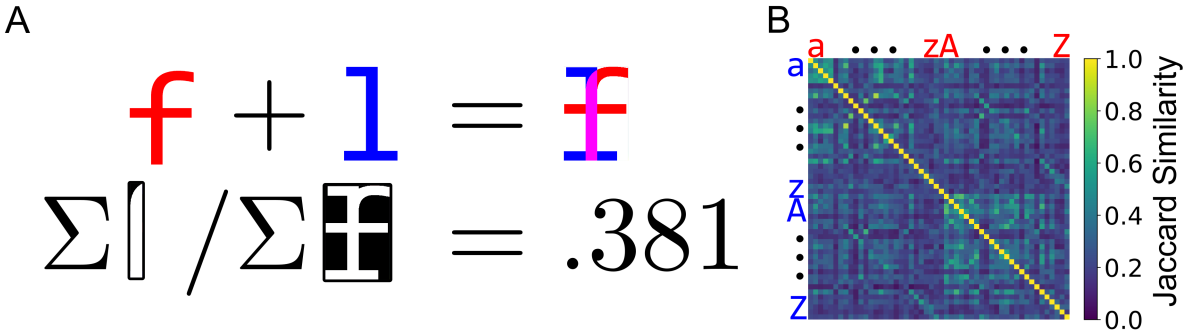


Figure 6.2: The method by which Jaccard similarities were estimated for pairs of characters from real-world fonts, using their default positions (central horizontal alignment if not monospaced), showing results from Consolas font as a monospaced example. (A) The binary matrix representations of characters are overlaid in their default positions, and Jaccard similarity is calculated as the sum of their intersection divided by the sum of their union. (B) Jaccard similarities for all alphabetic Consolas font characters.

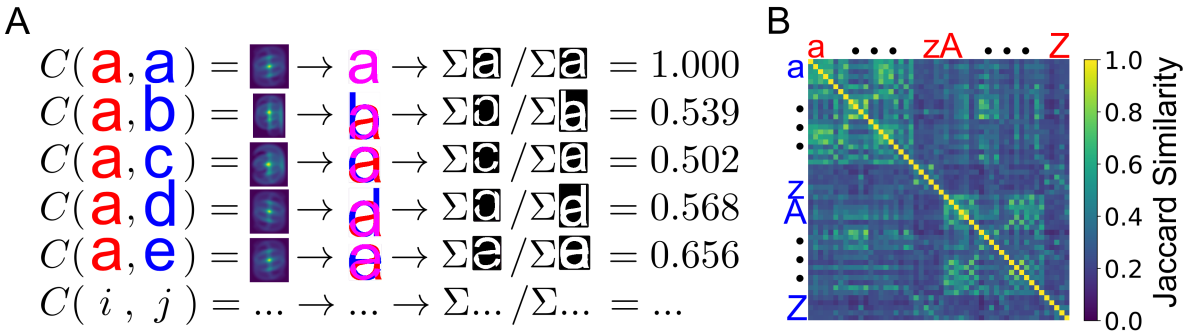


Figure 6.3: The method by which Jaccard similarities were estimated for pairs of characters from real-world fonts when translation was permitted, showing results from Arial as an example. (A) From the two-dimensional cross correlation (function C) of the binary matrix representations of characters, the translation required to achieve maximal overlap between the characters i and j could be identified as the peak coefficient. Using the position with maximal overlap, Jaccard similarity could be calculated as the sum of the intersection divided by the sum of the union. (B) Jaccard similarities for pairs of Arial characters at their positions of maximum overlap (permitting horizontal and vertical translation).

Translation Operation

To calculate the optimal Jaccard distance when translation is permitted, I estimated the two-dimensional cross correlation of each pair of characters (Figure 6.3). Two dimensional cross correlations were calculated via fast fourier transformations, using the *scipy* library for Python (Virtanen et al., 2020), where the location of the peak correlation coefficient reflects the location at which character i maximally overlaps character j. Allowing any vertical or horizontal translation in this manner allows Jaccard similarity to capture the similarity of letter features that do not align by default. For instance, rightward translation of the character *K* could allow its two diagonal segments to overlap those on the right side of the character *X*.

Rescaling, Rotating, and Mirroring Operations

In addition to translation, estimates of character similarity may also be improved by incorporating information from further geometric operations, such as rescaling, rotation, and mirroring, which are known to be relevant in objective and subjective effects of character similarity (Podgorny & Garner, 1979). For instance, many lower-case and upper-case characters differ mostly in scale (e.g., *c* and *C*), and some pairs have very similar shapes, but differ in rotation (e.g., *Z* and *N*) or flipping (e.g., *b* and *d*). Estimating the geometric transformations required to convert one character into another was made possible by combining the cross-correlation method detailed above with a non-linear optimisation approach (i.e., gradient descent; Figure 6.4). Specifically, the Nelder-Mead algorithm (Nelder & Mead, 1965) was used to identify values of rescaling and rotation that maximise Jaccard similarity. Here, rotation was represented in degrees (where 0 is the default rotation), while scale was log-transformed for symmetry around zero (i.e., the default scale of 1 is equal to zero, and rescaling to x.5 or x2 would produce values of -.69 and .69 respectively). To avoid identifying only local maxima, the algorithm was run several times for each pair of characters with varied starting values, and the maximum similarity from all runs was recorded. Starting values for rotation were varied from -180 to 180° in 6 equally spaced steps of 72°, while values for log scale were varied from $-log(3)$ to $log(3)$ (i.e., from 3x smaller to 3x larger) in 5 equally spaced steps of $\sim$.549. When both scale and rotation were optimised, this meant that Jaccard similarity was optimised with 30 distinct combinations of starting values for each pair of characters. Mirroring transformations were permitted by calculating the maximal Jaccard similarity twice, using either mirrored or unmirrored versions of the altered character.

Geometric Transformations

The maximal Jaccard similarity between pairs of characters was calculated for all possible combinations of geometric transformations (i.e., with/without translation, rescaling, rotation, and mirroring). As a result, there were 16 (2^4) variants of Jaccard similarity values for each font analysed using the pixel-based approach. In addition, each of these 16 variants was calculated twice for every non-identical character pair (i.e., the full matrix was calculated). Although I expected the matrix to be symmetrical, there were sometimes small differences between the sides of the matrix because of the non-linear optimisation approach applied for geometric transformations. As a result, I set the Jaccard similarity values to the maximum of the two estimates for a given character pair.

6.3 Character Similarity Validation

If the calculated character similarities capture perceived similarity, they should correlate with subjective judgements of perceived character similarity. To examine this, I modelled character similarity ratings reported by Simpson et al. (2013, trial-level data shared via personal correspondence). I expected this to show that all measures of character similarity would improve the quality of model fit relative to a model lacking any fixed effect of character

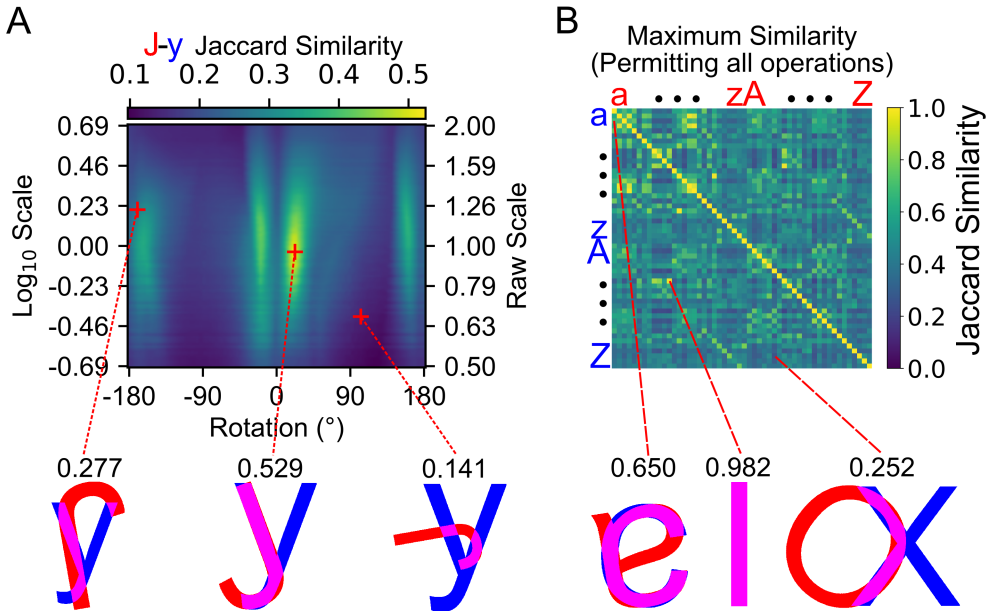


Figure 6.4: The method by which Jaccard similarities were estimated for pairs of characters from real-world fonts when translation, rescaling, rotation, and mirroring were permitted, showing Arial results as an example. (A) An example space explored by an optimiser tasked with maximising Jaccard similarity for Arial characters *J* and *y*, permitting operations of rotation and rescaling. For the *J*-*y* pair, optimal overlap is achieved by a slight decrease in *J*'s, and a rotation of around 20 degrees. (B) Jaccard similarities for pairs of alphabetic Arial characters where all geometric operations are permitted. Example pairs *a*-*c*, *l*-*I*, and *O*-*X* are highlighted.

similarity. However, I also expected the pixel-based estimates of Jaccard similarity for Arial font to provide the best improvement, as these values should best capture the features of characters that participants were judging. As expected, all values of character similarity improved model quality relative to an intercept-only model, and the pixel-based estimates for Arial font showed the best fit of the models examined. To examine whether this effect of font-specificity replicated, I then conducted an experiment similar to that conducted by Simpson et al. (2013), but where participants rated characters from one of two fonts. I expected that this would reveal a crossed interaction, where the best model for character similarity judgements would be that fit to font-congruent Jaccard similarities. The experiment comparing predictive validity of font-congruent Jaccard similarity values replicated the predictive power of Jaccard similarity, but did not find the expected font specificity effect, suggesting that the results in the Simpson et al. analysis may not reflect font specificity in the predictive power of similarities derived from the same font as that judged by participants.

6.3.1 Predicting Character Similarity Judgements

Simpson et al. (2013) collected character similarity judgements for 2,704 pairs of charcters, of which 2,356 were pairs of Arial font letters. Character pairs were rated on a seven-point Likert scale ranging from not at all similar (*1*) to very similar (*7*), and no pairs of identical characters (e.g., *k*-*k*) or case-conflicting (e.g., *g*-*F*) were presented. Unlike the original analysis, I did not exclude responses based on a ±2 *SD* cutoff, but only excluded missing (i.e., blank) or

meaningless (e.g., less than *1* or more than *7*) responses.

I calculated Jaccard similarities for the Rumelhart-Siple character set using the bit-wise approach outlined in the previous section, and used the pixel-based approach to calculate similarities for Arial, Calibri, Consolas, DOS VGA, Droid Sans, and Times New Roman versions of all upper- and lower-case alphabetic character pairs. I fit probit-link cumulative-link mixed effects models (CLMMs) with the R package *ordinal* (Christensen, 2020) to predict character similarity ratings as a function of the fixed effect of Jaccard similarity, for each of the six fonts. CLMMs account for the ordinal nature of Likert ratings, mapping responses onto ordered regions of a latent distribution (see chapter 3). For the pixel-based measures, separate models were fit for each of the 16 variants of Jaccard similarity calculated, with all possible combinations of geometric operations, for each font. The model formula, in the *lme4* syntax co-opted by the *ordinal* package, was specified as follows, where *jaccard* refers to the calculated Jaccard similarity (mean-centred), *item_id* uniquely identifies each pair of Arial characters, and *participant_id* uniquely identifies each participant:

```
rating ~ 1 + jaccard + (1 | item_id) + (1 + jaccard | participant_id)
```

Results revealed that the font producing the best-performing model (i.e., with the lowest Akaike Information Criterion; AIC) was usually fit with similarity estimates derived from Arial font, though Times New Roman values sometimes outperformed Arial when few geometric transformations were permitted (Figure 6.5A). When many geometric transformations were permitted, model performance improved, and models fit to values derived from Calibri and Droid Sans, fonts visually similar to Arial, showed the best performance of the non-Arial fonts. The best performing model overall was that fit using Arial-derived Jaccard similarities where all transformations were permitted. By far, the most informative transformation was translation, though including further transformations also improved the quality of prediction. The least informative transformation was rotation, leading to only a small improvement in AICs for the model fit to Arial when translation, rescaling, and mirroring were already permitted. Nevertheless, all Jaccard similarity variants produced CLMMs with AICs vastly lower than that observed for a model lacking the effect of Jaccard similarity (i.e., intercept-only).

Models revealed a robust effect of Jaccard similarity positively predicting participants' ratings of perceived orthographic similarity. For the best performing model, this relationship (Figure 6.5B) was estimated to be β=7.1 (*SE*=.24, likelihood-ratio $\chi^2(1)$=582.73, p<.001). As Jaccard similarity was mean-centred for the CLMMs, but was not rescaled, this is equivalent to an increase in the latent mean of .71 *SD*s for each 10% increase in Jaccard similarity.

As Rumelhart-Siple characters have only one case, the same Jaccard similarity values were used for both the upper- and lower-case characters. However, the Rumelhart-Siple characters most closely resemble typical upper-case characters. As a consequence, I re-ran the analysis reported here for lower-case and upper-case character pairs separately. The pattern of results for separate analyses of lower- and upper-case characters pairs (Appendix D.1) was broadly

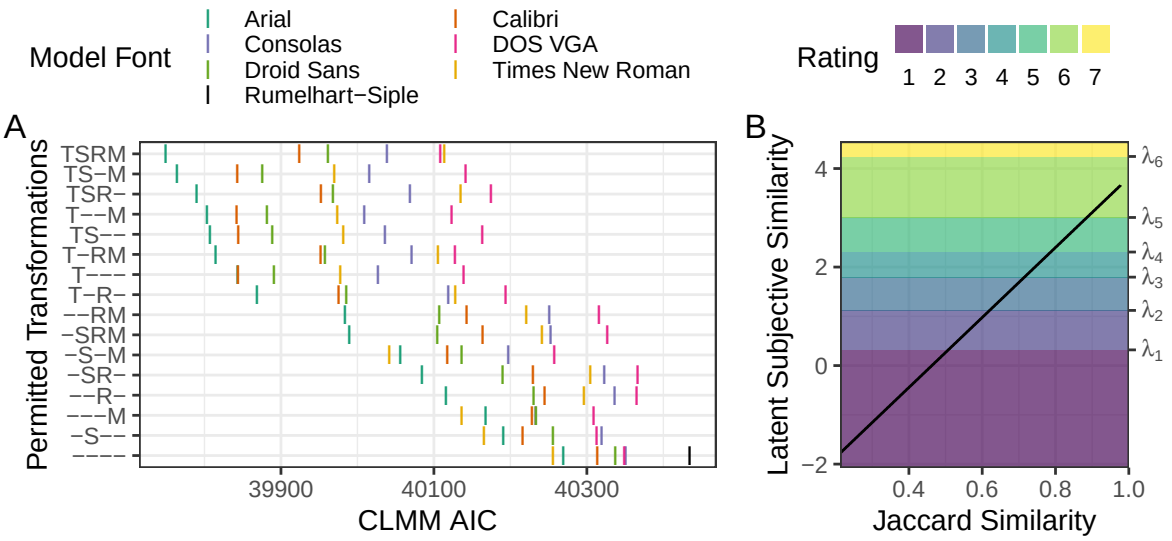


Figure 6.5: Results from the analysis of the relationship between calculated Jaccard similarity and subjective ratings of character similarity collected by Simpson et al. (2013). (A) The Akaike Information Criterion (AIC) associated with each CLMM, where line colours indicate the font for which Jaccard similarity was calculated, and the y axis indicates which geometric transformations were permitted in the calculation of Jaccard similarity: the presence of a *T* indicates that transformation was allowed, *S* that rescaling was allowed, *R* that rotation was allowed, and *M* that mirroring was allowed; - symbols indicate the absence of geometric transformations. Geometric transformations are ordered by AIC of the models fit using Arial font. (B) Fixed effects results from the optimal model, showing the estimated relationship (black line) between Jaccard similarity and character similarity ratings. The left-hand y axis depicts predicted latent means, while the right-hand y axis depicts the estimated location of thresholds (e.g., γ_1 demarcates responses *1* and *2*). Colours indicate the ordinal responses associated with latent mean values.

similar to that observed for models fit to both together.

6.3.2 Replicating Font Specificity

A key finding in the analysis of rating data from Simpson et al. (2013) was one of font specificity, with Jaccard similarity values derived from Arial quite consistently outperforming values derived from the other fonts analysed. I reasoned that this may reflect a font specificity in the relationship between Jaccard similarity and subjective perceptions of character similarities. However, as there was only evidence for one font in the Simpson et al. analysis, the superiority of Arial in predicting ratings from Simpson et al. may alternatively reflect a font-general effect, where Arial, of the fonts examined, somehow best captures font-general features that inform character similarity judgements. To examine whether the superiority of Arial in predicting ratings collected by Simpson et al. reflects font specificity in Jaccard similarities' predictions of subjective perceptions of character similarity, I conducted a replication of Simpson et al. (2013), collecting ratings for both Arial and Consolas font separately. I reasoned that if Jaccard similarity values are capturing font-specific features of characters, then models for ratings of similarities for both Arial and Consolas character pairs should more accurately predict character similarity ratings when the ratings are predicted using font-congruent Jaccard similarity values. I opted for a smaller character set to Simpson et al. (2013) for feasibility of data collection, to ensure all participants were familiar with the presented characters, and because simulations indicated that this would be sufficient to reliably detect the expected difference. I chose lower-case rather than upper-case characters because the font specificity of the relationship between Jaccard similarity and character similarity ratings appeared more robust for lower-case character pairs (Appendix D.1).

Power Analysis

To decide on a suitable minimum sample size for the experiment, I conducted a simulation-based power analysis. The experiment's central aim was to test whether font-congruent measures outperform font-incongruent measures in predicting subjective ratings of character similarity, and to show that this was true for two separate fonts. Power analyses are typically used to calculate the probability that a null hypothesis significance test will reveal a p value below a given alpha threshold for an experimental manipulation of interest. This is important for demonstrating that the proposed design and sample size have the statistical power necessary to detect the effect. Unlike studies that test the effect of a manipulation, however, the present experiment compares model fit for two possible predictors of the same data, for two separate datasets. As a result, the power analysis for this experiment focused on the expected results of a comparison in model fit, assuming that the improvement in font-specificity is equal to that observed in the Simpson et al. (2013) dataset. Specifically, I predicted a lower AIC value (or greater log-likelihood) for a CLMM which uses Arial-derived measures to predict judgements of similarity for Arial font characters, than one which uses Consolas-derived measures. Consistent with the experiment's hypothesis, it was assumed that this difference would be equal but in the opposite direction when using Consolas-derived measures to predict judgements of similarity

for Consolas font characters, rather than Arial-derived Jaccard similarity. To summarise, the power analysis' aim was to examine the distribution of AIC differences that should be expected for different sample sizes.

Based on the analysis of the Simpson et al. dataset, I decided that to use Jaccard similarities calculated with all four geometric transformations (translation, rescaling, rotation, and mirroring) permitted, as this variant of Jaccard similarity showed the best performance. I simulated data for judgements of Arial characters using estimates of the effects of Arial-derived measures in Simpson et al. (2013). Using coefficients estimated from a probit-link CLMM predicting the lower-case Arial judgements from Simpson et al., latent values were simulated as follows, where Table 6.1 presents each term's meaning and its simulated value as estimated for the lower-case Simpson et al. data:

$$L_{ip} = I_{0i} + P_{0p} + (\beta_1 + P_{1p})Jaccard_i$$

Table 6.1: Summary of the meanings of terms in the mixed effects model formula, and their simulated values. Where simulated values were drawn from random effect distributions, $\sim N(\mu, \sigma)$ indicates that the respective variable's values were drawn from a normal distribution with mean μ and standard deviation σ.

Term	Meaning	Simulated Value
L_{ip}	Trial-level latent means, for the item i and participant p	
I_{0i}	Item random intercept for item i	$\sim N(0, .69)$
P_{0p}	Participant random intercept for participant p	$\sim N(0, .79)$
β_1	Fixed effect of Jaccard Similarity	$= 7.12$
P_{1p}	Participant random slope for Jaccard similarity for participant p	$\sim N(0, 1.85)$
$Jaccard_i$	Jaccard similarity values for item i	

The correlations of random and fixed effects, and the variance-covariance matrices, were also drawn directly from the CLMM fit to the Simpson et al. data. As in the CLMM fit to the Simpson et al. data, per-item Jaccard similarity values ($Jaccard_i$) were those calculated for all lower-case Arial and Consolas letters when all geometric transformations were permitted. Latent means were used to calculate trial-level latent values by drawing from a normal (latent) distribution, $N(L_{ip}, 1)$, reflecting the *SD* of 1 that is imposed on CLMMs by default. Following simulation of latent values, these were recoded to Likert responses on a 7-point Likert scale by separating the latent distribution into 7 ordered regions, the locations of which were determined by the 6 response thresholds estimated by the CLMM which informed the simulation.

I decided that because participants would complete the experiment online with no monetary compensation, each participant should only rate 30 items. I varied the number of ratings collected for each item (from 2 to 8, with 250 iterations each), assigning simulated participants to items using the method described in section 6.3.2. CLMMs were fit using the same model formula as that described in section 6.3.1. On each iteration, I fit a CLMM with the *ordinal* package for R, using the same formula as that applied in the Simpson et al. analysis:

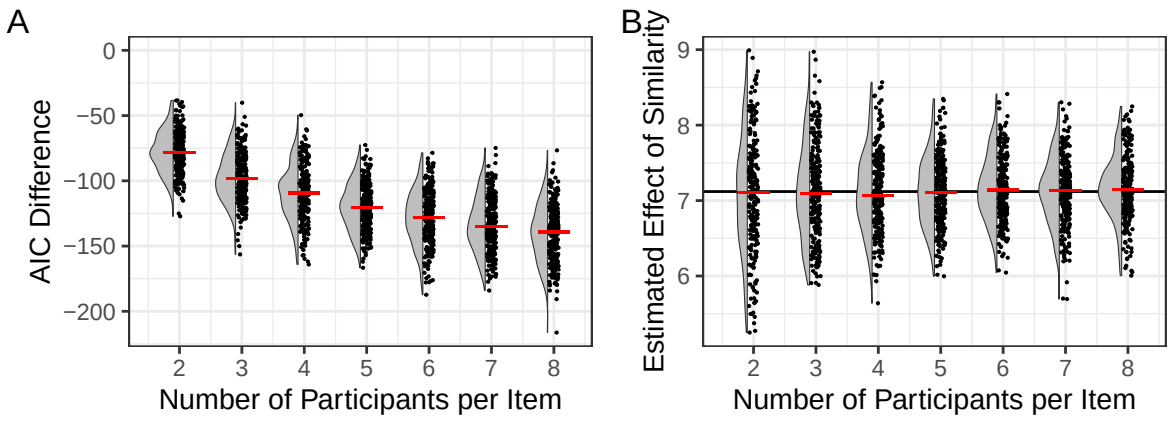


Figure 6.6: Results from the power simulations. (A) The distribution of AIC differences between CLMMs fit to simulated datasets using Arial- and Consolas-derived Jaccard similarities. A negative value indicates a superior fit for Arial-derived values. (B) The distribution of estimates of the effect of Jaccard similarity in CLMMs fit using Arial-derived Jaccard similarity values. In both panels, points depict values from individual simulation iterations, grey regions depict values' densities, and red lines depict median values. The solid black line in panel B depicts the simulated effect.

```
rating ~ 1 + jaccard + (1 | item_id) + (1 + jaccard | participant_id)
```

In analysing the results of the simulations, I examined the AIC difference between the CLMMs fit using Arial- and Consolas-derived fonts, and the variability of this value between iterations. Results (Figure 6.6A) showed that even with just one rating for each item, AIC values consistently and clearly detected the AIC difference between Arial- and Consolas-derived Jaccard similarity values. I also examined variability in the estimated coefficient for the effect of Jaccard similarity, with the direction and approximate magnitude of the effect also observed with only a small number of ratings per items. If the effect observed in the results of the analysis for the Simpson et al. data do reflect a font congruency effect, then a similar AIC difference should be observed for judgements of Consolas characters, and should be expected to be similarly clear.

Based on the simulations and on feasibility within the available period of time, I decided to collect data from at least 132 participants, which would provide at least 6 ratings for each unique pair of characters in Arial and Consolas (N=325 per font). I decided to then collect additional data from subsequent participants for additional accuracy in estimation of effects, with the experiment scheduled to stop automatically at the end of the pre-decided data collection period (July 2021 - April 2022).

Methods

The experiment was run using a custom-made web application using the Shiny package (W. Chang et al., 2018) for R, hosted on a publicly available Shiny web server (hosted by the

University of Glasgow School of Psychology and Neuroscience). The experiment was designed to be methodologically comparable to Simpson et al. (2013).

Participants each rated 30 different pairs of lower-case alphabetic letters, all drawn from either Arial or Consolas font, on a Likert scale from *1* ("not at all similar") to *7* ("very similar"). Participants were instructed to focus on the visual appearance of the letters rather than their sound: "It is important that you ignore the sounds of the letters, and just rate them purely on their visual appearance".

Trials were allocated to participants in the following way. The first 132 participants (66 Arial, 66 Consolas) were pre-allocated trials pseudorandomly such that each participant judged a random set of 30 different pairs of characters drawn from either Arial or Consolas, and so that each pair of characters was judged by at least 3 different participants. Participants after the first 132 were randomly assigned either Arial or Consolas font, and were then assigned trials pseudorandomly during the experiment by sampling on a per-trial basis from character pairs that currently had the lowest number of responses, that had not yet been responded to (e.g., if at the start of the experiment, 27 pairs had been responded to 5 times, and all other pairs had been responded to at least 6 times, then the first 27 trials would be sampled randomly from those that had been responded to 27 times, and the last 3 trials would be sampled from those that had at the start of the experiment only been responded to 6 times).

On each trial, participants were presented with their two characters side-by-side, in 72-point font, an equal distance from the horizontal centre of the screen (Figure 6.7). The side that each character was displayed on was randomised on each trial. The characters were presented as pre-rendered png images to ensure that participants' operating systems or browsers did not replace Arial and Consolas characters with characters from different fonts. Above characters was shown text, reading, "How similar do these letters appear, on a scale of 1 *(Not at All Similar)* to 7 *(Very Similar)*?". Below characters were presented seven buttons, labelled with the numbers 1 to 7. To encourage participants to judge each trial, these buttons only appeared 1 second after the current trial's characters were displayed. At the top of the page, text indicated the current trial number (e.g., "Trial 23/30"). Participants progressed through trials by clicking one of the seven buttons to indicate their choice.

Participants

In total, 247 participants completed the experiment (180 female, 58 male, 9 non-binary), of which 130 judged Arial character pairs, and 117 judged Consolas character pairs. All participants reported having no learning disabilities or disorders that impair their reading, and reported that in their first language, all (N=165) or most (N=82) characters were in the Latin alphabet. The mean age was 20.09 (SD=5.24), and most participants (N=206) were aged 20 or younger. I decided prior to data analysis to exclude any participants who responded to all 30 items with only one response, but all participants responded with at least two different responses. As a result, no participants were excluded from the analysis.

Data collection was approved by the University of Glasgow School of Science and Engineering Ethics Committee (application number: 300200293).

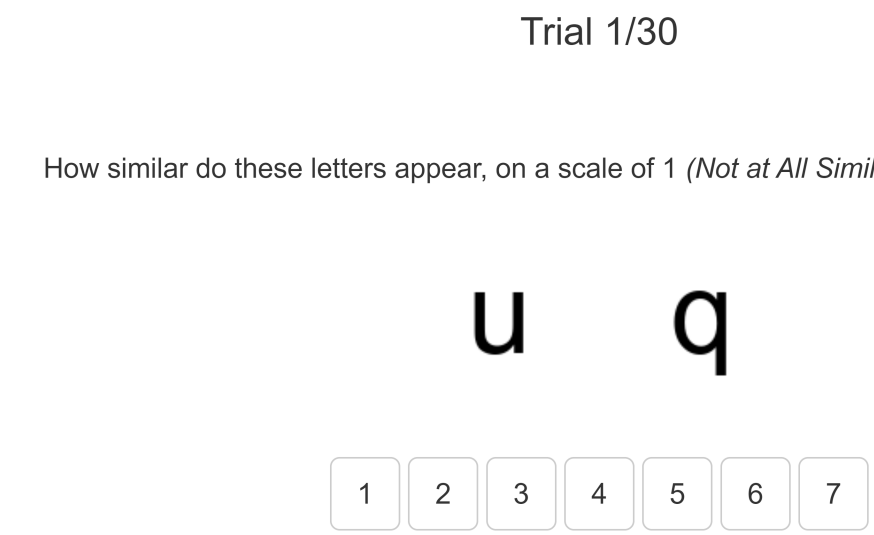


Figure 6.7: An example trial for the Arial characters *u* and *q*.

Results

Arial character pairs were judged by between 10 and 17 participants (mean=12, *SD*=1.45), while Consolas character pairs were judged by between 10 and 14 (mean=10.8, *SD*=.91) participants. In total, there were 3,900 observations for Arial font, and 3,510 observations for Consolas font.

I fit four separate CLMMs with the *ordinal* package, varying the font that participants were judging (Arial/Consolas) and the font that Jaccard similarities were derived from (Arial/Consolas). The model formula was identical to that used in the SImpson et al. analysis and in the power analysis. All Jaccard similarity values were derived permitting all four geometric transformations (translation, rescaling, rotation, mirroring). Model AICs are presented in Table 6.2. Surprisingly, results showed that while the superiority of font-congruent Jaccard similarity values observed for judgements of Arial characters in Simpson et al. (2013) data was replicated, this font congruency effect was not observed for judgements of Consolas characters.

Table 6.2: AICs for the CLMMs fit predicting character similarity judgements using Arial- or Consolas-derived Jaccard similarity values, when the presented font was either Arial or Consolas.

Presented Font	Arial Model AIC	Consolas Model AIC
Arial	10889.87	11026.97
Consolas	9815.42	9838.41

I also examined whether the same conclusions would be drawn using a more Bayesian approach, making use of parameters estimated in the analysis of the Simpson et al. data. I fit four Bayesian CLMMs with prior distributions informed by the analysis of Simpson et al.,

using Arial-derived Jaccard similarities. However, I also added more uncertainty in the prior distributions than I would have were I completing a full-scale replication of Simpson et al. (2013), as I anticipated that differences in the experimental design from Simpson et al. (e.g., online vs. in-person, 30 vs. 108-120 trials per session, etc.) may lead to different patterns of effects. I specified priors for threshold locations as $\sim N(\gamma_i, 1)$, where γ_i were the estimated threshold locations from the analysis of Simpson et al.. I specified the prior for the fixed effect of Jaccard similarity as $\sim N(7.1, 1)$. Prior distributions for standard deviations of random effects distributions were specified with Student's *t* distributions, as $\sim t(10, \sigma, 1)$, where σ was the the estimated standard deviation from the Simpson et al. data analysis for that random effect. Finally, the prior distribution for the random correlation between per-participant intercepts and per-participant effects of Jaccard similarity, was specified to be flat, between -1 and 1.

All four Bayesian models were fit using *brms*, with the same model formula as that applied in the analysis using CLMMs fit with the *ordinal* package. Each model was fit with 5 chains, each with 10,000 iterations (split equally between warm-up and sampling). The *adapt_delta* parameter was set to .8, and the *max_treedepth* parameter was set to 10.

I examined differences in the expected log probability densities (ELPDs) of the Bayesian models, using the *loo* package for R (Vehtari et al., 2017). ELPD quantifies the expected predictive accuracy of a given model for out-of-sample observations from the same population (Vehtari et al., 2017). I calculated the leave-one-out ELPD for each model, using the Pareto-smoothed sampling approach. In each model, over 99.4% of the Pareto distribution's *k* parameters were <.5, which is indicative of high reliability in the ELPD estimates (Vehtari et al., 2017). Results revealed the same pattern of results as the AICs of the CLMMs fit via the *ordinal* package, with models using Jaccard similarities derived from Arial outperforming those using models derived from Consolas, for both sets of judgements. However, while the difference between the predictive accuracy of the models fit to judgements of Arial characters was quite robust (ELPD difference = -46.5, *SE* = 10.2), the difference between the models fit to judgements of Consolas characters was less robust, with the ELPD difference showing a high standard error relative to its size (ELPD difference = -8.4, *SE* = 8.3). As a result, this experiment does not provide conclusive results on font specificity in the predictions that Jaccard similarities provide of subjective similarity ratings, although this analysis does suggest that if font-congruent Jaccard similarity values better predict character similarity ratings, then the improvement in predictive accuracy is likely smaller than I first anticipated.

I also examined the posterior distributions for the fixed effect of Jaccard similarity (Figure 6.8). Differences in the effect of Jaccard similarity revealed a surprising finding, that for both Arial and Consolas characters, the effect of Jaccard similarity was estimated to be larger when the similarity values were derived from Consolas font than when they were derived from Arial font. This finding is difficult to interpret, but may be an artefact of the poorer predictions provided by Consolas font.

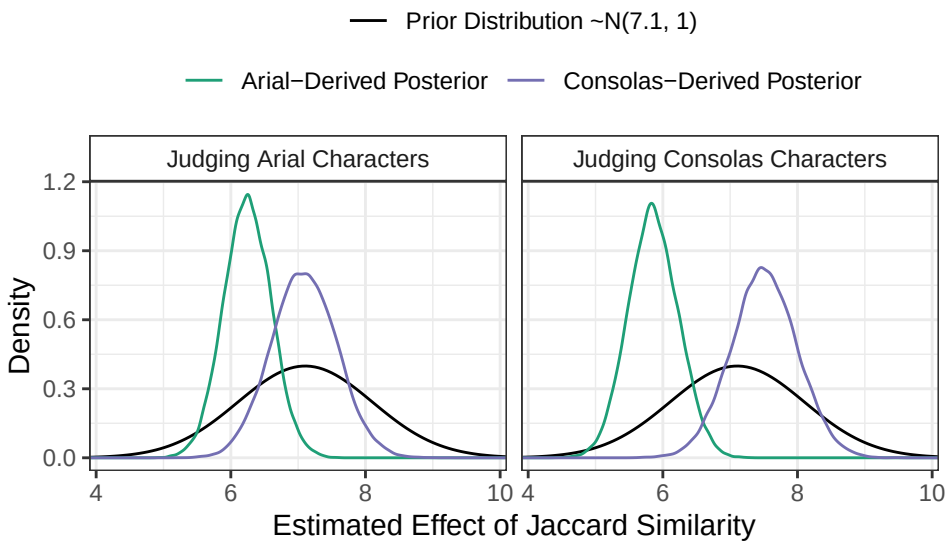


Figure 6.8: Posterior distributions for the effect of Jaccard similarity, for judgements of Arial characters (*left*) and Consolas characters (*right*), where Jaccard similarity values were derived from Arial font (*green*) or Consolas font (*purple*). The distribution shown in black reflects the prior distribution for the effect of Jaccard similarity, which was identical in all four models.

6.4 Sub-Character Orthographic Similarity and Orthographic Neighbourhood Density

After calculating objective estimates of sub-character letter similarities, which show a clear (though not necessarily font-specific) relationship with subjective ratings of letter similarities, I was interested in applying these estimates to quantify the sub-character orthographic similarity of entire word forms. From these estimates descriptions of sub-character orthographic neighbourhood density could then be derived. The approach I took was to calculate metrics computationally similar to OLD and OLD20, weighting substitution, insertion, and deletion operations by sub-character information, or else that were conceptually similar to OLD and OLD20, capturing similar features of words but using a different approach. For both approaches, the aim was to calculate measures of orthographic similarity between word forms that are sensitive to sub-character information. I refer to these measures as variants of SCOLD (sub-character orthographic Levenshtein distance), and the associated neighbourhood metrics as variants of SCOLD20.

6.4.1 SCOLD

I calculated two variants of SCOLD: SCOLD$_s$ and SCOLD$_{sid}$. Both measures could be calculated using any combination of permitted geometric transformations. For feasibility, however, only three combinations of geometric transformations were examined: (1) default positions, (2) translations only, and (3) all four transformations. These correspond to ----, *t*---, and *tsrm* in Figure 6.5, respectively. All variants of SCOLD were calculated for all 7 character sets examined (6 fonts, plus the Rumelhart-Siple character set), but, as in prior analyses in this chapter, only

default character position SCOLD values were calculated for the Rumelhart-Siple character set.

Example SCOLD values are presented in Table 6.3, showing the differences between SCOLD values and traditional OLD values. The selected examples demonstrate (1) that all variants of SCOLD show sensitivity to sub-character features in substitutions (e.g., the reduced cost for substitution *o-a* in *pocket-packet*, relative to *o-i* in *pocket-picket*), (2) that the variants differ considerably in cases where geometric transformations permit high overlap between characters (e.g., the reduced cost for substitution *p-d* in *pocket-docket*, relative to *p-r* in *pocket-rocket* for *tsrm* variants of SCOLD), (3) that inclusion or exclusion of geometric transformations only affects SCOLD values when substitutions are involved (e.g., unchanged values for the addition of characters *ed* in *pocket-pocketed*), and (4) that SCOLD can capture the orthographic similarity between words and nonwords (e.g., *pocket* is more similar to *drdhbl* than it is to *vlfczm*).

Table 6.3: Example SCOLD values for the orthographic distances between *pocket* and 8 other strings, calculated from matrix representations of 50-point Arial font, and their corresponding OLD values for comparison. Variants of SCOLD reflect the method by which character similarities were calculated, permitting, respectively, no geometric transformations (----), only translation transformations (*t*---), and translation, rescaling, rotation and mirroring transformations (*tsrm*). For comparability with the later SCOLD20 metrics, all $SCOLD_s$ values are calculated using only lower-case character similarity matrices.

String A	String B	OLD	$SCOLD_s$			$SCOLD_{sid}$		
			----	t---	tsrm	----	t---	trsm
pocket	pocket	0	.000	.000	.000	0	0	0
pocket	packet	1	.616	.703	.815	160	152	140
pocket	picket	1	1.124	1.209	1.313	309	269	217
pocket	docket	1	1.073	.749	.393	450	192	22
pocket	rocket	1	.947	.960	1.083	270	214	117
pocket	pocketed	2	2.000	2.000	2.000	666	666	666
pocket	chainsaw	8	6.901	7.041	7.352	1929	1733	1286
pocket	drdhbl	6	5.310	4.725	4.972	1531	1029	793
pocket	vlfczm	6	5.954	6.152	6.208	1755	1519	1305

$SCOLD_s$

The first measure I calculated was $SCOLD_s$, where the only change from OLD is that substitution operations were weighted by Jaccard similarity. Specifically, Levenshtein distance was calculated using a typical Wagner-Fischer algorithm (Vintsyuk, 1968; Wagner & Fischer, 1974), but where the cost assigned to character substitutions was the scaled Jaccard similarity between the characters being exchanged. Jaccard similarity was scaled such that it had a range of 1 (between the maximum and minimum), and was mean-centred on a value of 1. Because of large differences between lower- and upper-case characters, due to size differences, this rescaling approach can result in within-case substitutions, which are by far the

most common in everyday written language, being under-weighted. To avoid this, I calculated SCOLD$_s$ using only lower-case character similarity matrices (as it was only data for lower-case words that were analysed in the later validation of SCOLD and SCOLD20).

SCOLD$_{sid}$

I also calculated SCOLD$_{sid}$, where all OLD operations (substitution, insertion, and deletion) were conducted using sub-character information. Specifically, again using the Wagner-Fischer algorithm, the cost of inserting or deleting a character was equal to (for the pixel-based approach) the number of pixels in the character being inserted or deleted, or (in the bit-wise approach) the number of bits being inserted or deleted. Correspondingly, substitutions were in this approach weighted by the number of pixels or bits that needed to be inserted or deleted to transform one character into the other (i.e., $a \cup b - a \cap b$). SCOLD$_{sid}$ is similar to an approach recently implemented to describe orthographic similarities of Chinese characters (Sun et al., 2018), though the geometric transformations permitted in the present implementation, as well as the application to alphabetic script, are novel.

6.4.2 SCOLD20

SCOLD20 values were calculated for each SCOLD variant: SCOLD20$_s$ and SCOLD20$_{sid}$, for all three combinations of geometric transformations: (1) default positions, (2) translations only, and (3) all four transformations (translation, rescaling, rotation, and mirroring), for each of the 6 fonts and for the Rumelhart-Siple character set. For comparability to Yarkoni et al. (2008), these neighbourhood metrics were calculated from words in the ELP (Balota et al., 2007), although unlike Yarkoni et al., I only calculated the measures for words that only contained lower-case alphabetic characters without diacritical marks (i.e., no spaces, hyphens, accented letters, etc.; *N*=37,432). For comparison, OLD20 values were calculated from the same pool of words, using equal costs for each Levenshtein distance operation (i.e., addition, deletion, and substitution costs were all set to 1). SCOLD20 values were calculated for each word as the mean SCOLD of that word's 20 closest neighbours, for each SCOLD variant separately. All SCOLD20 values showed high correlations with one another, and with OLD20 (Figure 6.9 shows example results for Arial font).

6.4.3 Validation

To assess the degree to which the inclusion of sub-character information in metrics of orthographic neighbourhood density is cognitively meaningful or useful, I compared the performance of models using OLD20 and variants of SCOLD20 to predict behavioural and neural correlates of lexical decision.

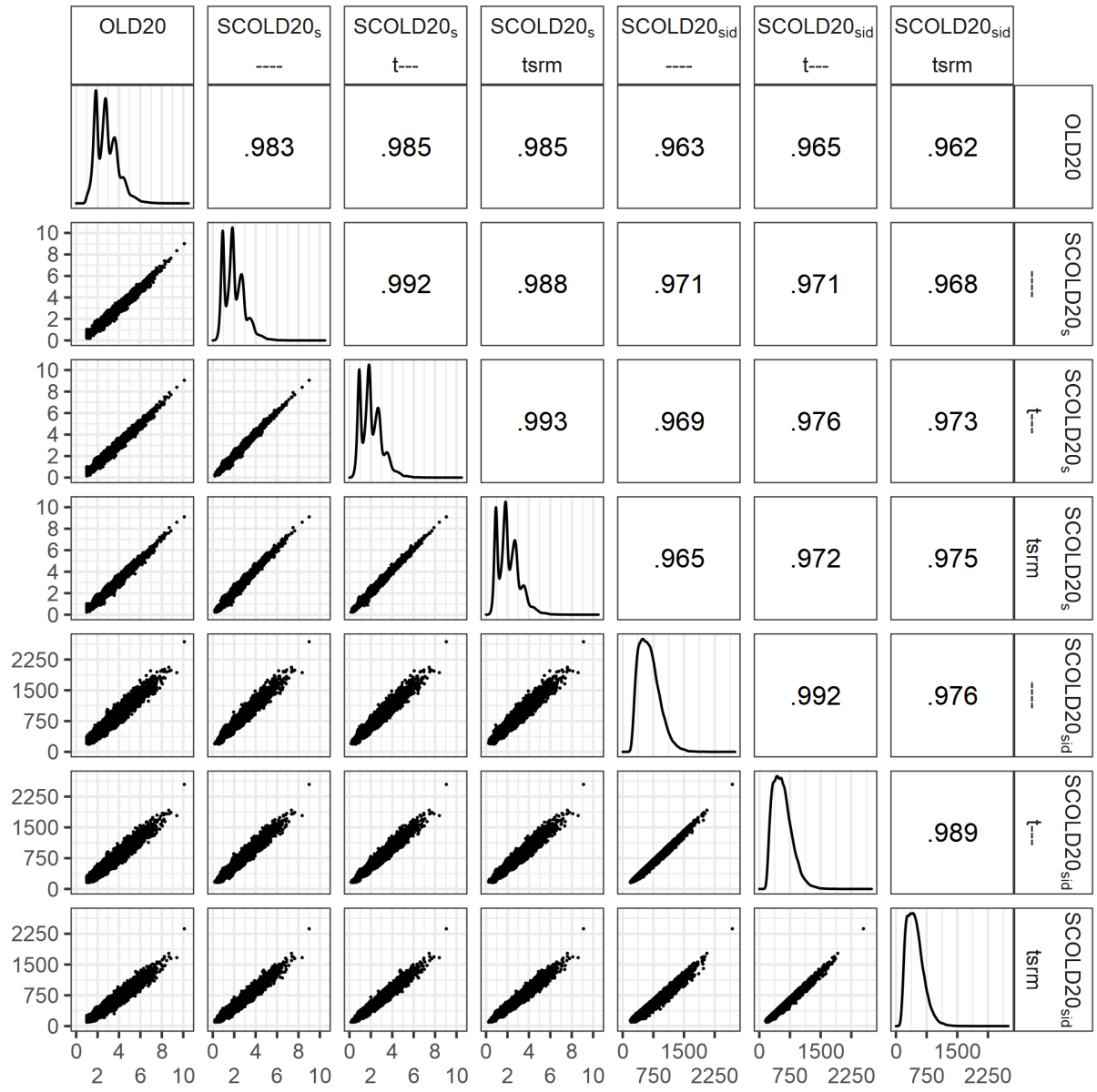


Figure 6.9: Correlations between all pairs of SCOLD20 variants for Arial font. Correlations with OLD20 are also included for comparison. Panels in the lower triangle of the figure depict the correlations between words' values in each variable. Numbers in the upper triangle of the figure are Spearman correlations for each pair of variables. Panels on the figure's diagonal depict individual variables' densities.

Predicting Lexical Decision Behaviour

I predicted lexical decision response times (RTs) and accuracies for all words in the ELP lexical decision dataset. I fit trial-level generalised linear mixed effects models (GLMMs) where the only fixed effect predictor was either a SCOLD20 variant or OLD20. Trials were excluded from both the RT and accuracy analyses if either the RT was greater than 4000 ms or recorded in the dataset as less than 1 ms. Trials were additionally excluded from the RT analysis if responded to incorrectly. Trials were only included if the presented word was a member of the pool of the 37,432 lower-case words for which OLD20 and SCOLD20 values were calculated. Following these exclusions, the accuracy data consisted of 1,278,171 trials, while the RT data consisted of 1,072,162 trials. Random effects comprised per-participant and per-word random intercepts, and per-participant random slopes for the effect of the measure being assessed. The models were fit using *lme4* (Bates et al., 2015), with the following model formula:

```
outcome ~ 1 + measure + (1 | word_id) + (1 + measure | participant_id)
```

Accuracies were modelled via logit-link binomial GLMMs, while RTs were modelled via linear mixed effects models (Gaussian link) predicting *inverse* RTs. Inverse RTs were calculated as -1000/RT, where RTs were measured in ms, such that models of inverse RTs can be understood as estimating the number of items, or information units, that could be processed or responded to within one second (Brysbaert & Stevens, 2018). While analysing effects on inverse RTs necessarily forgoes much rich distributional information in RTs (Heathcote et al., 1991; Lo & Andrews, 2015), it represents an approach that captures robustly effects on RTs' central tendencies, while its efficiency makes it appropriate for describing effects in large datasets like lexical decision megastudies, of which the ELP is one.

Results (Figure 6.10) showed that for both RTs and accuracies, the model fit using OLD20 values outperformed models fit using all variants of SCOLD20. This finding suggests that the inclusion of sub-character information in the OLD20 metric of orthographic neighbourhood density did not improve predictions of lexical decision behaviour. Furthermore, although the best-performing SCOLD20 model was fit using DOS VGA values (where DOS VGA was the font presented in ELP trials), the AIC differences did not show as clear a superiority of font-congruent models as was observed in the analysis of subjective character similarity judgements.

I also analysed RTs and accuracies from the British Lexicon Project (BLP; Keuleers et al., 2012), which showed a similar pattern of effects as was observed for the ELP, with the OLD20 model outperforming all SCOLD20 models, and any font-congruency superiority was less clear than in the subjective character similarity results (Appendix D.2).

Predicting ERPs

I compared the explanatory power of the SCOLD20 metrics in describing ERP data, with an exploratory analysis of data collected in Chapter 4. I fit sample-level models (at 512 Hz) to the

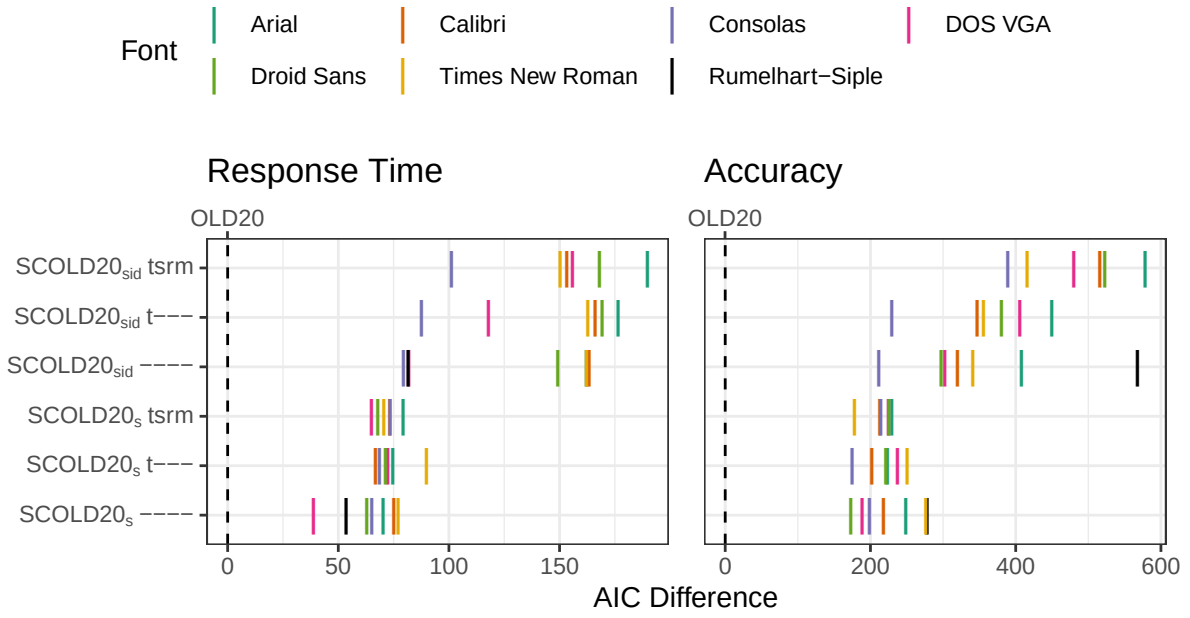


Figure 6.10: AIC differences between models predicting ELP lexical decision behaviour using OLD20, and using all calculated SCOLD20 variants. AIC difference values reflect the difference between AICs from each SCOLD20 model and the OLD20 model, where a positive AIC difference indicates superior performance in the OLD20 model.

dataset from the task modulation experiment. Models were fit as described in section 4.3.1, except that models did not estimate random slopes, and additionally estimated the fixed effect of orthographic neighbourhood, represented by the OLD20 and SCOLD20 metrics, scaled by standard deviation for comparability. All possible fixed effect interactions with the fixed effect structure were additionally included, to capture any interactions with the stimulus and task variables manipulated in the experiment. This decision was motivated by evidence for interactions between orthographic similarities and lexicality (Baeck et al., 2015). I analysed data from three separate regions of interest: a left-lateralised occipitotemporal region (matching that used in chapter 4), a right-lateralised occipitotemporal region, and a centroparietal region.

I first examined the effect of OLD20 - the results of this analysis are summarised in Appendix D.3. To summarise the results here, OLD20 showed effects in all three regions examined, and may have interacted with task and stimulus variables. One notable finding was that OLD20 values (larger orthographic neighbourhoods) elicited less negative-going left hemispheric N1 components.

I then examined the relative performances of the OLD20 and SCOLD20 models. Comparing model AICs over time (Figure 6.11) revealed that some SCOLD20 variants outperformed OLD20 at key points. In particular, SCOLD20$_{sid}$ without any geometric transformations outperformed OLD20 in predicting right hemispheric N1 components, while this difference was smaller in the left occipitotemporal N1. This difference may reflect that the effect size of orthographic neighbourhood is larger overall in the N1 over the right hemisphere than over the left (Appendix D.3). For the right hemispheric occipitotemporal region, SCOLD20$_{sid}$ without any geometric transformations showed sustained superiority in predicting ERP amplitudes.

All SCOLD20 variants also outperformed OLD20 models at later time points, especially in the centroparietal region after 300 ms. Here, the best model was again that fit using SCOLD20$_{sid}$ without any geometric transformations. The variant of SCOLD20$_{sid}$ that was calculated permitting *all* geometric transformations showed a sustained superiority over other variants in predicting centroparietal ERP amplitudes later than 400 ms, while all other variants showed performance more similar to OLD20. At several other time points, OLD20 showed a superiority over SCOLD20, especially centroparietally between 150 and 300 ms. OLD20 also largely outperformed SCOLD20 variants in predicting N1 amplitudes during the component's offset period, suggesting that sub-character feature similarity neighbourhoods may be more influential during processing that occurs during the earlier portions of the N1.

6.5 Discussion

I have developed and evaluated a measure of word form similarity with sub-character granularity, in the form of SCOLD; Sub-Character Orthographic Levenshtein Distance. I have shown two implementations of this measure, weighting substitutions by Jaccard similarity (which I show to predict well subjective ratings of character similarity), or weighting all operations by a bit-wise or pixel-wise cost. I suggest that while SCOLD and SCOLD20 do not represent a universal improvement over OLD and OLD20, they may provide greater insight into the information that is represented at different stages of visual word processing. I additionally suggest that SCOLD and SCOLD20 are well-placed for inclusion in computational models of visual word recognition theories, which can be used to form computationally informed hypotheses.

I calculated several variants of Jaccard similarity and of SCOLD, designed to capture variability between fonts and the influence of geometric transformations. For SCOLD I implemented both SCOLD$_s$, which weights character substitutions by Jaccard similarities, and SCOLD$_{sid}$, which weights all operations by the number of pixels involved. SCOLD$_s$ and SCOLD$_{sid}$ therefore differed in both the operations that were weighted and the manner in which the weighting was implemented, with the SCOLD$_{sid}$ approach functionally similar to the approach that has previously been applied to calculate similarities for word forms in more logographic orthographies like Chinese (Sun et al., 2018), though it was here implemented more flexibly, supporting geometric transformations, and in a manner more efficient for alphabetic orthographies. It is important to note that the SCOLD and SCOLD20 variants calculated here were not intended to be exhaustive, but rather to provide proof-of-concept validation. Indeed, SCOLD variants may be implemented in a more bespoke manner, such as permitting horizontal but not vertical translation, to preserve the vertical locations of ascenders and descenders that evidence shows are particularly informative in reading processes (Beech & Mayall, 2005). Similarly, anticipating that the *degree* of geometric transformation required may provide more nuanced descriptions of character similarities, additional predictors could be integrated as coefficients into models of character and word similarities capturing features like the geometric distances of translations, size changes in rescaling, and angles of rotations, that are required to optimise character similarity.

In validating SCOLD and SCOLD20, I first showed that Jaccard similarities between

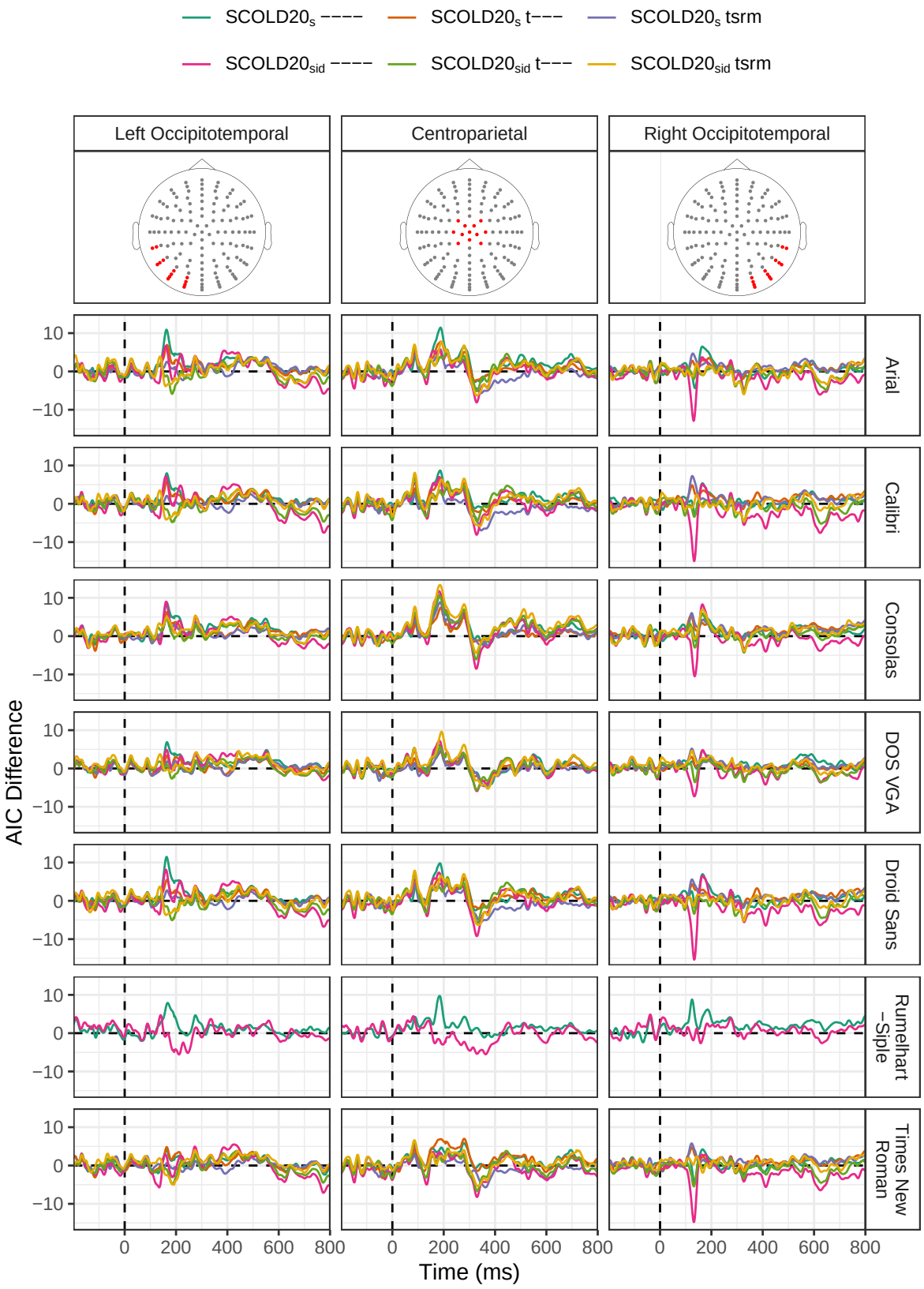


Figure 6.11: AIC differences between models predicting ERP amplitudes from Chapter 4 using OLD20, and all calculated SCOLD20 variants. Positive AIC differences indicate superior performance in the OLD20 model, while negative differences indicate superior performance for SCOLD20.

characters concord well with subjective ratings of orthographic character similarity, and that optimising the measure using geometric transformations further improves predictions of subjective ratings. Results initially indicated a possible superiority of measures calculated from fonts that match those presented, though a follow-up study suggested this may not be true. I then calculated measures of word form similarity that aimed to integrate sub-character similarity information with the traditional measure of OLD, in the form of SCOLD variants. From these measures I derived measures of orthographic neighbourhood density, SCOLD20, comparable to the current gold standard measure, OLD20. I showed that while OLD20 outperformed SCOLD20 in predicting behavioural outcomes in lexical decision, SCOLD20 may better account for neural correlates of some word recognition processes, including the orthographic processing indexed by the N1. These differences were small, though this is consistent with the high correlations observed between OLD20 and SCOLD20 values (Figure 6.9). One key finding was that while geometric transformations improved predictions of subjective ratings of character similarities, objective effects of neighbourhood density were best predicted by measures that did not permit geometric transformations, especially in the early occipitotemporal effects. The superiority of SCOLD20 values, and effects of neighbourhood density more generally, were also heterogeneous across the N1's window, both peaking in the N1's onset, with the effect of orthographic neighbourhood and difference in predictive power of OLD20 and SCOLD20 then peaking in the opposite direction during the N1's offset. These results reflect the heterogeneous and often hierarchical nature of orthographic processing in the ventral occipitotemporal cortex (vOT), with processing becoming progressively more abstracted from visual input (Hannagan & Grainger, 2013), in particular showing greater invariance to information like retinal position in more anterior subregions (Dehaene et al., 2004) that are likely involved later than more posterior regions. Such posterior-to-anterior, early-to-late, progressive abstraction from visual input may account for why early, but not late, periods of the N1 are better explained by SCOLD, which captures sub-character features, than by OLD values, which is restricted to the character level. Tentative evidence for a superiority of font-congruent SCOLD20 was observed in the modelling of ERPs, with measures derived from Droid Sans (the predicted font) and visually similar Calibri font, among the SCOLD20 variants providing the greatest improvements over OLD20. To summarise the validations, SCOLD does not always account for behavioural and neural correlates of orthographic and word recognition processes better than OLD, but whether it does can provide valuable insight into whether such information is relevant to the analysed outcome, for instance identifying time points at which processing is sensitive to sub-character information.

One area not explored here is in the effects of similarities between individual word forms on behavioural and neural correlates of word recognition. While neighbourhood metrics for SCOLD and OLD, that average over the *N* smallest pairwise values, correlate very highly (Figure 6.9), differences between individual pairwise SCOLD and OLD values may be more variable, such that SCOLD could provide greater improvement over OLD than was observed at the level of the neighbourhood. For instance, consider effects of repetition priming where the orthographic similarity of prime and target can vary (e.g., Dehaene et al., 2001; Eisenhauer et al., 2022; Huang et al., 2022): evidence suggests that priming effects scale with the

degree of similarity between prime and target (Kinoshita et al., 2014; Lien et al., 2021). The degree to which behavioural and neural correlates of repetition priming are affected by the orthographic similarity of word forms could be described more fully by SCOLD than by OLD. Indeed, sub-character descriptions of orthographic similarity may be particularly useful for describing and modelling effects of top-down modulation in orthographic processing - whether predictions are hypothesised to "explain away" bottom-up orthographic input (A. Clark, 2013; Gagl et al., 2020; Lien et al., 2021; Rao & Ballard, 1999), or "sharpen" sensitivity to predicted orthographic features (Eisenhauer et al., 2022, see chapter 5), the degree to which top-down modulation functionally influences orthographic processing should be expected to scale with the orthographic similarity between the predicted and observed word form (though in opposite directions). It follows that measures like SCOLD could also be applied to provide greater insight into orthographic processing in biasing contexts, where higher-level information influences early orthographic processing via top-down modulation, as was examined in chapters 4 and 5. SCOLD could be included in models like those reported in chapter 5 to more accurately estimate, and more fully describe, how orthographic similarity relates to top-down modulation. The stimuli in the picture-word verification experiment employed in chapter 5 were designed specifically to minimise orthographic similarity between congruent and incongruent word forms, such that any similarity-dependent effects of predictability and congruency were maximised. Nevertheless, future research could examine whether the congruency-predictability interaction observed in this experiment scales with the degree of dissimilarity between predicted and observed word forms, which results here suggest could be better captured by SCOLD than OLD. In addition, this could more directly relate the observed pattern of top-down effects to modulation of orthographic processing specifically.

Relatedly, SCOLD also provides an opportunity for the development of more advanced computational models of orthographic processing. Rather than assuming characters to be functionally interchangeable orthographic units, incapable of being further divided into their constituent parts, computational models integrating SCOLD could be expected to describe and predict in greater detail exactly how orthographic neighbourhood and similarity effects relate to behavioural and neural correlates, providing falsifiable, computationally derived hypotheses that experiments could be conducted to examine. Such computational models can play a vital role in testing and comparing cognitive models (Guest & Martin, 2021), and may be one route to providing the more specific hypotheses that Ramsey and Ward (2020) argue is lacking in current research into top-down modulation. Computational models integrating SCOLD could also suggest possible mechanisms that might reconcile seemingly contradictory findings consistent with predictive coding accounts of orthographic processing (Gagl et al., 2020; A. Kim & Lai, 2012; Kretzschmar et al., 2015; Lien et al., 2021; Sereno et al., 2003) or with "sharpening" of predicted features (Eisenhauer et al., 2022; Sereno et al., 2019, chapter 5).

I finally note that, although, in the analyses reported here, the estimates of Jaccard similarity and SCOLD that were derived, using a bit-wise approach, from the Rumelhart-Siple character set, were consistently outperformed by estimates from the pixel-based approach, this does not entirely invalidate the use of Rumelhart-Siple characters. Indeed, while the characters were originally developed to model empirical behaviour (Rumelhart & Siple, 1974), they may

be integrated into models more as a computationally convenient and easily interpretable representations of orthography that is *analogous* to real-world orthography and orthographic processing, rather than as a veridical description of it. A model built on Rumelhart-Siple characters could predict broad patterns of effects that can be compared to observed behavioural and neural correlates of visual word recognition (e.g., predicting the relationship between word frequency and lexical decision latency; Coltheart et al., 2001) to evaluate model performance. However, it should also be noted that pixel-based representations of characters, and measures like pixel-based SCOLD that can be derived from such representations, may produce computational models with greater ecological validity. By better describing empirical stimuli in this way, such models could produce item-level predictions of orthographic processing that could be compared more directly to behavioural and neural correlates.

To summarise, SCOLD can capture, at the sub-character level, the orthographic similarity of word forms. The measure, and the neighbourhood metrics derived from it, can provide more nuanced descriptions of orthography and orthographic processing, and could be integrated into models to describe and test specific mechanisms by which top-down contributions affect orthographic processing.

Chapter 7

General Discussion

The work in this thesis examines orthographic processing in visual word recognition, and the degree to which it may be sensitive to higher-level predictions in biasing contexts via top-down modulation. In particular, I have focussed on the N1: an early occipitotemporal event-related potential (ERP) component that peaks earlier than 200 ms after word presentation. In addition to replicating the N1's early sensitivity to orthographic information, findings suggest that, indeed, these processes are likely sensitive to higher-level predictions, most probably via top-down modulation. However, the specific mechanisms underlying this sensitivity to top-down contributions on activity indexed by the N1, and that are capable of accounting for the observed patterns of effects, remain unclear. I suggest that the dynamics of top-down modulation of orthographic processing can be more clearly delineated through the testing of more specific hypotheses, informed by and derived from computational implementations of candidate theories, and methodological improvements such as in the operationalisation of orthographic similarity.

7.1 Sensitivity to Orthography in the N1

The N1 ERP component shows clear sensitivity to orthographic information (Appelbaum et al., 2009; Bentin et al., 1999; Brem et al., 2018; Eberhard-Moscicka et al., 2016; Gagl et al., 2020; Holcomb et al., 2002; Ling et al., 2019; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019; J. Zhao et al., 2014). This finding was replicated in chapters 4 and 5, with the left-hemispheric N1 showing more negative (average) amplitudes for nonwords and false-font stimuli than it does for words. Conversely, no robust difference was observed in the left hemispheric N1 between words and orthographically (and phonologically) plausible pseudowords - this finding was consistent with some existing evidence (Holcomb et al., 2002; Maurer, Brem, et al., 2005) but inconsistent with some other findings (Eberhard-Moscicka et al., 2016; Hauk et al., 2012; Hauk et al., 2006; Segalowitz & Zheng, 2009). This inconsistency in findings may reflect variability in the plausibility of pseudowords, such that previous reports of word-pseudoword sensitivity reflect not necessarily lexical access, as they have sometimes been interpreted (e.g., Eberhard-Moscicka et al., 2016), but rather a sensitivity to orthographic features that can differ between words and pseudowords to a varying extent. Demonstration of a sensitivity

to lexicality, via comparisons between the factorial conditions of words and pseudowords, would require, in addition to pseudowords being orthographically *plausible*, careful and precise item-wise matching of orthographic variables such as N-gram probabilities. A comparable argument for early lexical access has been made on the basis of early sensitivity to word frequency in components that include the N1 (Assadollahi & Pulvermüller, 2003; Dambacher et al., 2006; Hauk & Pulvermüller, 2004; Sereno et al., 2003; Sereno et al., 1998; Simon et al., 2007), with similar sensitivity to word frequency observed in functional magnetic resonance imaging (fMRI) of the ventral occipitotemporal cortex (vOT; Kronbichler et al., 2004) - the likely generator of the N1 component. However, word frequency necessarily correlates with sublexical orthographic features like character N-gram probabilities and orthographic neighbourhood size, such that a sensitivity to word frequency could emerge from processing of only sublexical constiuents of words. If precise matching of confounding variables across such factorial (word vs. pseudoword) or continuous (word frequency) predictors is not possible, then correlating such experimental manipulations with neural correlates may be a method that is, when applied in isolation, incapable of disentangling the mechanisms underlying such correlates. As I have argued is true of research into top-down modulation of orthographic processing, an approach that may elucidate cognitive mechanisms underlying neural correlates is the implementation and testing of computational models. To summarise, evidence in this thesis and elsewhere is consistent with the N1 indexing orthographic processing. However, I argue that neither sensitivity to lexicality in the N1, nor to lexical frequency, necessarily index lexical access.

7.2 Sensitivity to Top-Down Modulation in the N1

In chapter 4, where I examined effects of category-level semantic predictions in a task modulation paradigm, no robust interaction between category relevance and task was observed in the N1. The earliest that task interacted with category relevance, in occipitotemporal electrodes, was late into the N1's offset, more than 50 ms after the N1 peak and closer to the succeding P2 component's onset, with the difference peaking centroparietally around 400 ms. I suggested that the lack of evidence for top-down modulation in the N1 at the level of semantic categories may be unsurprising, if the N1 indeed indexes orthographic processing, as even if the participant does predict orthographic features, a semantic category will encompass so many orthographically diverse word forms as to be functionally uninformative once recoded to an orthographic representation. I did, however, find evidence for a small stimulus-general task modulation of the N1, with more negative-going N1 components in the SCT, consistent with existing evidence for such early effects of task (Y. Chen et al., 2013; Rahimi et al., 2022; Segalowitz & Zheng, 2009).

In chapter 5, I examined effects of prediction for more specific word forms in a picture-word verification task. A congruency-predictability interaction was observed in the left, occipitotemporal N1. This finding demonstrates an interaction between context (image predictability) and stimulus (presented word form), which provides evidence for an influence of higher-level predictions of specific word forms on the N1. If bottom-up processing during the

N1 is restricted to orthographic information, this congruency-predictability interaction reflects top-down modulation of orthographic processing.

Nevertheless, the direction of effects observed in chapter 5 was counter to the pattern of effects that I hypothesised. Based on previous findings from sentential studies, where prediction-congruent word forms elicited less negative-going N1 components (A. E. Kim & Gilley, 2013; Kretzschmar et al., 2015; Sereno et al., 2003), I hypothesised that, similarly, picture-congruent word forms would elicit less negative-going N1s. Further, I expected that this effect would be largest at the highest level of predictability, and close to zero at the lowest level of predictability. This hypothesised pattern of effects would be consistent with a predictive coding account of processing during the N1, if such a mechanism additionally permits a short-term online influence of predictions for specific word forms, according to which predicted orthographic features are "explained away" to minimise overall prediction error (A. Clark, 2013; Gagl et al., 2020; Rao & Ballard, 1999; J. Zhao et al., 2019). The pattern of effects observed was in the opposite direction to that hypothesised: an effect of predictability emerged in the N1's peak for picture-incongruent and not picture-congruent words, with N1 components becoming less negative-going for picture-incongruent words as predictability increased. In addition, during the N1's offset the size of this interaction grew, with the opposite effect emerging for picture-congruent words, with N1 amplitudes in the offset period becoming *more* negative for picture-congruent words as predictability increased, not less.

Based on the discrepancy between the hypothesised and observed pattern of effects, a simplistic implementation of top-down modulation via predictive coding is insufficient to account for effects of prediction in the picture-word verification task. One alternative explanation could be that rather than predictions "explaining away" bottom-up input, they lead to a preactivation of orthographic information that induces a "sharpening" of responses (Eisenhauer et al., 2022). Ostensibly, this interpretation could account well for the results of the present work, but, as I have argued, is in need of reconciliation with findings from sentential studies. It is also unclear why timing differences were observed, between the congruency conditions, for the effect of predictability, which emerged for picture-incongruent words before it did for picture-congruent words. I have suggested that, to better construct, discriminate between, and evaluate hypotheses of orthographic processing and its sensitivity to top-down contributions, there should be a greater focus on computational implementations of theories. Results of computational simulations could guide research by highlighting patterns of neural correlates of orthographic processing that arise as consequences of theories, with such insights informing the direction of future experimental investigations. I developed one important contribution to such computational models in chapter 6, in the form of Sub-Character Orthographic Levenshtein Distance (SCOLD). The SCOLD approaches I implemented could be integrated into models of orthographic processing to describe and predict more fully effects of orthographic similarity and the sensitivity of orthographic processing to top-down modulation.

7.3 Generation and Recoding of Predictions

An important consideration, if orthographic processing is influenced by top-down modulation, is the manner in which the participant constructs and maintains internal predictions of upcoming content. In predictive coding terms, such a mechanism could be referred to as a generative model, capable of continually updating in response to prediction error and changes in context. Clearly the neural mechanism that underlies participants' predictions is of central importance to any model of top-down modulation of orthography, as it will determine what information is predicted, and the latencies and locations at which predictions can interact with bottom-up processes.

One possibility is that readers are using the language *production* system (Adank, 2012; Dell & Chang, 2014; Momma & Phillips, 2018; Pickering & Garrod, 2007) to construct predictions. Here, it is suggested that the reader or hearer covertly emulates language production as linguistic input is comprehended, and predicts upcoming content from perceived contexts and intentions (Pickering & Gambi, 2018; Pickering & Garrod, 2013). Indeed, there is a high degree of overlap between regions implicated in language production and comprehension systems (Giglio et al., 2022), and vOT shows both direct and indirect anatomical and functional connections to such shared regions (Bouhali et al., 2014; Vogel et al., 2012; Woodhead et al., 2014; W. Zhou et al., 2016). If participants are utilising language production systems to construct predictions, then this may result in differences between paradigms that use linguistic stimuli and non-linguistic stimuli to bias participants' expectations (see section 1.5.2). For instance, more naturalistic sentential stimuli may automatically induce the production system to predict upcoming content, whereas more artificial non-linguistic (e.g., pictures) or single-word stimuli may require more conscious control to translate semantic and contextual information into linguistic predictions. Indeed, evidence suggests that naturalistic, self-paced reading, where task demands are minimal, elicit disparate dynamics of orthographic processing, with more activity for words than pseudowords (Schuster et al., 2015), whereas most studies that use more artificial tasks and paradigms reveal the opposite pattern (e.g., Kronbichler et al., 2004; Woolnough et al., 2021) that would be more consistent with predictive coding accounts (Gagl et al., 2020) according to which, if their is a word-pseudoword difference, prediction error should be higher for pseudowords. Such stimulus differences could also underlie the distinct patterns of prediction effects observed in sentential and non-sentential studies, which the results of chapter 5 further highlight. An alternative account of prediction generation posits that linguistic predictions are indistinct from from other prediction processes, sharing a single prediction system with non-linguistic prediction processes (Altmann & Mirković, 2009), or integrating multiple mechanisms to predict language, of which a domain-general production system is simply one component (Huettig, 2015). However, even if prediction processes are separate from language production mechanisms, this does not discount the possibility that the dynamics of communication between language and predictive processing systems could result in vastly different patterns of neural activity for predictions that are derived linguistically versus nonlinguistically.

It is additionally pertinent to ask by what mechanisms top-down contributions interact

with bottom-up signals in vOT. Indeed, exactly how information is transmitted between levels of processing remains a key question in research into, and models of, word recognition more generally (Norris, 2013). Kay and Yeatman (2017) have suggested a model in which contributions from the intraparietal sulcus (IPS) are largely responsible for task-driven modulation of processing in the VWFA, showing that stimulus-related activity in IPS scales with the degree of top-down modulation observed in vOT. Consistent with a functional role of IPS in controlling word recognition processes is the finding that anatomical connections between vOT and parietal regions like IPS strengthen with literacy acquisition (Moulton et al., 2019). However, in addition to task-driven modulation of early occipiotemporal activity (Bentin et al., 1999; Y. Chen et al., 2013, 2015; Qu et al., 2022; Strijkers et al., 2015; F. Wang & Maurer, 2017), which requires only broad modulation of word form processing, evidence suggests a more targeted influence of predictions, such as for specific word forms (chapter 5). In their model of vOT computations, Kay and Yeatman (2017) implement a simplified model where Gabor filters for local orientations are projected onto category templates and vOT responses scale with the result of a dot-product template-matching computation. The most parsimonious implementation of more targeted predictions for specific word forms in the general model structure outlined by Kay and Yeatman would be targeted alteration of the word form template, perhaps differentially weighting orthographic features to facilitate processing of features present in the predicted word form.

If predictions are implemented via weighting of orthographic features, then predictions must be recoded into orthographic representations. Research is indeed consistent with the representation of and decoding of such sub-word-form orthographic features in early occipitotemporal activity. Training a classifier to decode presented word form identities from bilateral occipitotemporal EEG activity, Ling et al. (2019) have shown that orthographic features can be reconstructed with accuracy well above chance, and that confusability of word identities in the classifier's representations concord well with word forms' orthographic similarities. Discrimination accuracy of this classifier additionally peaked at 200 ms, within the typical offset period of the N1. Sensitivity to the visual features of orthography appears to contradict findings of invariance in vOT (e.g., insensitivity to letter case; Lu et al., 2021). Nevertheless, this disparity could be reconciled with reference to the heterogeneity of processing within the N1 and vOT (see section 7.4 below), with invariance *emerging* in vOT, and even abstracted orthographic representations retaining some degree of low-level information (Rauschecker et al., 2012; J. S. Taylor et al., 2019). The recoding of semantic predictions into targeted orthographic representations is likely to involve a pattern of activation distinct from that involved in more general task-driven top-down modulations, which Kay and Yeatman (2017) show can be described well with reference to just IPS and vOT activity. For instance, the pattern of neural activity involved in targeted semantic-to-orthographic top-down modulation may involve more frontal and perisylvian regions (Eisenhauer et al., 2022; J. Wang et al., 2019; Woodhead et al., 2014), as well as more proximal regions like that identified by Purcell et al. (2014), a region anterior to the VWFA that appears to be selectively involved in interfacing between orthographic and semantic information. As I have argued in this work, recoding of predictions from semantic to orthographic representations would be computationally lossy, entailing a loss of specificity.

For instance, predictions for the word form *chair* may also confer facilitation for features in semantically unrelated yet orthographically similar word forms, such as *stair*, *chain*, or *char*, and even orthographically similar nonwords like *cbeln*. In addition to testing this predicted effect as a general hypothesis (e.g., A. E. Kim & Gilley, 2013), future investigations could provide insight into the mechanisms by which top-down modulations influence occipitotemporal orthographic processing by testing how well different computational models, like that utilised by Kay and Yeatman (2017), could account for effects like cross-word-form facilitation when predictions target specific word forms. The calculation of sub-character word form similarity could be implemented using a measure similar to or derived from the SCOLD measure that I have proposed and tested in this work.

However, further research is required to describe more clearly how orthographic information is represented in vOT and during the N1. One barrier that must be overcome is that models of top-down modulation of orthographic processing that account for effects entirely through reference to computations operated upon visual descriptions of stimuli need to be reconciled with evidence for the meta-modal and flexible nature of processing in vOT (Price & Devlin, 2011). Further to this, it should be considered whether top-down influences causally modulate the perception of non-visual linguistic, or even non-linguistic, processing in vOT. For instance, Willems et al. (2016) reported an effect of predictability, specifically of surprisal, on vOT activity during *spoken* language comprehension. Whether such findings reflect a role of phonological-to-orthographic recoding (Dehaene et al., 2002; Madec, Le Goff, Anton, et al., 2016) or direct representation of phonological information in vOT (Pattamadilok et al., 2019; Qu et al., 2022; L. Zhao et al., 2017) is of central relevance to computational models of orthographic processing and its sensitivity to top-down modulation, as it will constrain the manner in which orthographic information is encoded and the types of representational codes that predictions can be transcoded into, which, as argued above, constrain the functional implications of top-down modulation. How to infer and understand neural representations remains a key and current question in cognitive neuroscience more generally, especially as computational models (and data-driven classifiers) can achieve high accuracy in predicting neural activity even when models' representations are entirely unalike those in the brain (Guest & Martin, 2021; Popov et al., 2018). Yet, the development and testing of more formal descriptions of the content of orthographic representations are vital for research to delineate the manner in which bottom-up processing interacts with top-down predictions.

7.4 Processing during the N1 is Heterogeneous

A common finding in the current work was that processing within the N1 is not homogeneous. Rather, sensitivity to different features emerged at different time points across the component's window and differed between hemispheres. The heterogeneities I observed in the N1 were generally consistent with previous findings, and align well with research that takes a more spatial perspective.

7.4.1 Timing Differences

Effects of stimuli, including word-versus-false-font differences and the effect of orthographic neighbourhood density, generally emerged earlier, in the component's onset or peak, although effects were largest in the component's offset. Such early effects of stimuli are consistent with evidence that bottom-up features of word forms can be accurately decoded from occipitotemporal N1 amplitudes as early as 150 ms (Ling et al., 2019, although it is notable that classifier accuracy similarly peaks later, during the component's offset). Meanwhile, consistent with previous findings (F. Wang & Maurer, 2017, 2020), I found that sensitivity to higher-level information and predictions, likely to rely on top-down contributions, emerged later, in the N1's peak or offset. Temporal heterogeneity within the N1 is consistent with comparable *spatial* heterogeneity in vOT, the N1's likely generator, with processing becoming increasingly abstracted from information in visual input such as retinotopy and image contrast (although not completely; Kay & Yeatman, 2017; Rauschecker et al., 2012; J. S. Taylor et al., 2019), as it progresses anteriorly through a feedforward hierarchy (Dehaene et al., 2004; Hannagan & Grainger, 2013; J. S. Taylor et al., 2019; Vinckier et al., 2007).

7.4.2 Hemispheric Differences

In chapter 4, the sensitivity to the word-nonword difference was larger over the right hemisphere during the N1 than it was over the left hemisphere. The ERP analysis of these data in chapter 6 similarly suggested that the effect of orthographic neighbourhood density (OLD20) and the superiority of a metric that is additionally sensitive to sub-character complexities (SCOLD20) were larger in the right hemispheric N1 than they were in the left. Results from chapter 5 also showed that sensitivity to the difference between words and false-font stimuli was larger over the right hemisphere than it was over the left hemisphere. These findings are consistent with previous reports of stimulus-hemisphere interactions in the N1 component (Bentin et al., 1999; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019), as well as evidence for divergences between the right and left vOT in the effect of orthographic similarity (Dehaene et al., 2004; Krafnick et al., 2016; McCandliss et al., 2003; Vinckier et al., 2007; Woolnough et al., 2021) and possibly corresponding with differences in the sensitivity to orthographic information presented in the right and left visual hemifield (Parker et al., 2021; Rauschecker et al., 2012, c.f., Takamiya et al., 2020). Such hemispheric interactions may be related to evidence that orthographic processing occurs primarily in the left vOT (and in the visual word form area; VWFA), with the region's right-hemispheric homologue conveying left-visual-field input to the left vOT via the corpus callosum (Bouhali et al., 2014; Cohen et al., 2000; McCandliss et al., 2003; Molko et al., 2002).

7.5 Orthographic and Predictive Processing Prior to the N1 and vOT

While this work focuses on the N1 ERP component, some evidence has suggested that activity earlier than the N1, in regions posterior to vOT, may also index orthographic processing. For instance, magentoencephalography (MEG) evidence reported by Solomyak and Marantz (2010) shows sensitivity in an occipital component peaking around 130 ms, dubbed the M130, to the frequency of word forms' affixes (e.g., affixes *able* and *ity*, in word form *probability*). Evidence for this component being distinct from the M170 has been reported by Gwilliams et al. (2016), with the finding that levels of visual noise in images of word forms modulated a "type two response" (a term used to distinguish pre-orthographic, visual processing from word form processing; Tarkiainen et al., 1999) in a posterior M130 component, while the later M170 shows sensitivity to differences between letter strings and symbols. The timing of this component is consistent with the earliest observed sensitivity to visual differences between lower- and upper-case characters, at around 120 ms (Madec, Le Goff, Riès, et al., 2016). However, the degree to which this M130 component indexes orthographic processing, rather than just late and more anterior effects of low-level features such as luminancy, as are observed in the M100 (Gwilliams et al., 2016; Helenius et al., 1999; Tarkiainen et al., 1999), is here unclear.

The M100 and associated P100 component observed in EEG are mostly associated with low-level features of stimuli like luminance or stimulus size (Helenius et al., 1999; Kurita-Tashima et al., 1991; Tarkiainen et al., 1999; Tobimatsu et al., 1993; Wicke et al., 1964), or features of word forms like word length (Hauk & Pulvermüller, 2004) or number of character strokes (Hsu et al., 2011) that correlate with such low-level visual information. However, some studies have suggested that effects of orthographic or post-orthographic information are even observed in the early stage of processing indexed by the M/P100. For instance, Segalowitz and Zheng (2009) identified an effect of lexicality in the P100, indexed by a difference between words and orthographically (and phonologically) plausible pseudowords that were matched in length, with more positive-going P100 components elicited over both hemispheres by words than by pseudowords. A similarly early effect of lexicality was reported by Sereno et al. (1998), with more positive-going P100 components elicited by pseudowords and nonwords than by words (although these differences were observed with a parietal topography unusual for this primarily occipital component). Related to such reports are findings of early, pre-N1 effects of frequency, which, as previously mentioned, are often interpreted as a proxy indicator of lexical access. In their study of prediction effects elicited by sentential contexts, Sereno et al. (2019) reported a main effect of word frequency in the P100, with a more positive-going posterior P100 elicited by low-frequency than by high-frequency words. It was further reported that this effect interacted with predictability, present for words with high but not low predictability. Strijkers et al. (2015) also reported a context-dependent sensitivity to frequency, with a left-hemispheric posterior effect of word frequency at 120 ms during a semantic task, but not during a colour categorisation task, with more positive-going amplitudes observed for high-frequency than for low-frequency words. Scott et al. (2009) similarly reported an interaction between *emotion* and word frequency, with more positive average P100 amplitude for positively valenced high

frequency words than for negatively valenced high-frequency words, but no effect of emotion for low-frequency words. Similar effects have been reported in the M100 component of MEG, mostly comprising early influences of predictability. In a study with a picture-word verification design similar to that I employed in chapter 4, Dikker and Pylkkanen (2011) reported an effect of picture-word congruency in the M100 for predictive pictures (designed to elicit a prediction for a specific word form), but not for non-predictive pictures, with more extreme amplitudes for picture-incongruent words. In a related study, Dikker et al. (2009) reported effects of syntactic violations on the M100, where syntactically unexpected word forms elicited more extreme amplitudes. Such early effects of syntactic predictions have also been reported in Arabic script: Matar et al. (2019) found that syntactically unpredictable Arabic word forms elicited greater activation in posterior M100 activity. As highlighted by Nieuwland (2019), in their review of word form prediction literature, some early effects of word form prediction could be related to artefacts of high-pass filtering that, for non-causal filters, result in differences in later components like the N400 being pushed backwards in time (Tanner et al., 2015; VanRullen, 2011). However, while high-threshold (1 Hz) high-pass filtering is a common feature of some studies listed here that report early prediction effects in the M/P100 (Dikker & Pylkkanen, 2011; Dikker et al., 2009; Matar et al., 2019), many employ lower high-pass thresholds, at .1 Hz or below.

If pre-N1 activity indexes orthographic (or post-orthographic) processing, what neural structures could be generating these signals? Research has generally pointed to vOT, and specifically the VWFA, as the site of the earliest specifically orthographic processing (Centanni et al., 2018; Cohen et al., 2000; Cohen et al., 2002; Dehaene & Cohen, 2011; Price, 2012). However, information processed in vOT must be necessarily present, and to some extent decodable, in regions that provide input to vOT (Petersen et al., 1988; Pugh et al., 1996) - even if such regions encode information that is less orthographically specific (and more involved in domain-general visual processing), they could still be sites of early sensitivity to orthographic features, or even candidate sites for interactions between bottom-up, visual and top-down, higher-level information. Recently, Woolnough et al. (2021) identified via intracranial EEG a mid-fusiform region posterior to the VWFA that shows early sensitivity to lexicality and word frequency - possibly consistent with M/EEG findings of early sensitivity to these features. This region showed an initial sensitivity to low-level orthographic information, indexed by the difference in response to words and strings of infrequent letters, which was sustained and later joined by a sensitivity to more high-level orthographic and lexical information, sensitive to differences between words and, successively, strings of frequent letters, strings of frequent bigrams, and finally, strings of frequent quadrigrams. The role of this mid-fusiform region in font-invariant orthographic processing is supported by fMRI evidence (Z. Zhou, Vilis, et al., 2019), and may be the same as the occipitotemporal "letter form" area identified by Thesen et al. (2012) via fMRI and intracranial EEG, which is similarly mid-fusiform and directly posterior to the VWFA. However, linking the activity of this letter form area to the M/P100 is made difficult by mismatches in timing and localisation. Specifically, the timing and localisation of the pre-VWFA activation identified by Woolnough et al. and Thesen et al. do not concord well with the M/P100, peaking instead closer to 170 ms than 100 ms, and showing a more

occipitotemporal source than would be suggested by the mostly occipital topography of the M/P100. Indeed, a source localisation of the M100 prediction effects identified by Matar et al. (2019) localised effects mostly to the primary visual cortex (i.e., V1), while effects were much smaller or absent in M100 activity localised to Brodmann areas 18 (i.e., V2) and 19 (comprising V3, V4, V5, and V6), suggesting that the reported M/P100 sensitivity to syntactic violations is indeed related to processing in more posterior visual areas. That such posterior regions could show early selective sensitivity to differences between word forms is at first surprising, given the sensitivity of V1 to low-level information like retinotopy (Engel et al., 1997; Tootell et al., 1998) and visual contrast (Avidan et al., 2002) that could be expected to exclude it as a candidate for orthographic processing, as such regions lack the invariance across viewing conditions that has been argued to be vital for the shape recognition that orthographic processing involves (Avidan et al., 2002; Cohen & Dehaene, 2004; Rust & DiCarlo, 2010). Nevertheless, evidence shows that, to a limited extent, the low-level processing of early visual areas is actively implicated in object and shape recognition processes (Fischer & Whitney, 2009; Hsieh et al., 2010; Kok & De Lange, 2014). Indeed, recent findings have demonstrated that early representations of individual letters, in V1 and V2, are enhanced when presented within the context of words relative to nonwords (Heilbron et al., 2020), and that regions of the middle and inferior occipital gyrus may have a response to orthographic information that is font-invariant (Z. Zhou, Vilis, et al., 2019). Relatedly, emerging research shows that task demands or context can influence such early processing in word recognition, probably via top-down modulation, with effects of task to as early as 60 ms post-stimulus, and as posterior as V1 (Rahimi et al., 2022).

In sum, orthography-relevant information is necessarily processed prior to the N1 and vOT. Immediately prior to VWFA processing, there is evidence for less holistic, letter-specific processing, though the timing and localisation of activation is here very similar to that of the N1. It is possible, however, that although not selective for orthography, very early visual processing could also be considered orthographic. In particular, emerging findings point to a top-down influence on such early visual processing of words, possibly corresponding to reports of effects in the M/P100. These findings therefore further highlight the far-reaching influence of top-down contributions to visual perception and word recognition.

7.6 Methodological Contributions

The present work also makes several contributions to methodology in the research area, especially in chapters 2 and 3. In chapter 2, I describe *LexOPS*, an R package I developed to aid in robust and reproducible item-wise stimulus matching. The package is aimed at language researchers, and provides code-based and graphic interfaces for users to construct formal pipelines to generate stimulus lists that can be reproduced by other researchers and altered to generate new stimulus lists that fit the same criteria. I applied LexOPS in chapters 4 and 5 to investigate top-down influences on orthographic processing. I also describe in chapter 2 an assumption-free distribution-wise approach to matching variables, and show that this method can be integrated with the item-wise approach implemented in LexOPS to design stimuli more flexibly. I applied this integrative approach in chapter 5 to generate a list of picture-word stimuli

that were matched on some variables in an item-wise manner, and on other variables in a distributional manner. In addition, stimuli for the localisation task of that chapter were matched distribution-wise, on several variables, to the population of words they were drawn from.

In chapter 3, I demonstrated that a hierarchical, ordinal modelling framework commonly applied to data from Likert rating experiments, the cumulative-link mixed effects model (CLMM), can be applied to Likert rating norming studies, conferring several advantages over traditional approaches of means and standard deviations. Specifically, I showed that via Monte-Carlo simulations that norming items by their random effects estimates of latent variables from CLMMs removes artefacts of nonlinearities in responses that have previously been identified as problematic in norming study datasets, while shrinkage and accounting for between participant variability confer additional accuracy in estimation. I also showed that distributional CLMMs, estimating differences in both location and spread of latent distributions, can provide more complete descriptions of items' ratings, from which an analogue of ratings' standard deviations can be calculated which does not treat responses as continuous.

I applied the modelling approaches described and evaluated in chapter 3 in subsequent chapters. The example CLMM analysis of data from Simpson et al. (2013), described in section 3.3, is expanded upon in chapter 6, with the inclusion of a fixed effect of Jaccard similarity estimates for orthographic similarity. This approach allowed me to more accurately estimate and describe the strong relationship between the objective estimates of character similarities that I calculated, and participants' subjective perceptions. Additionally, the distributional modelling approach outlined and applied in the CLMM framework, in sections 3.2.3 and 3.3, is later applied to model response time data in chapters 4 and 5, providing a more full description of response time distributions by modelling effects on all parameters of a shifted log-normal distribution. Relatedly, Monte-Carlo methods like those utilised in chapter 3, accounting for multiple hierarchical effects and relationships, were applied to conduct power analyses in chapters 5 and 6.

Finally, I note that the contributions from chapters 2 and 3 are complementary: word stimuli are frequently matched on normed variables. If specifying a tolerance within which a variable, from a normed ratings dataset, should be matched, the researcher is essentially assuming that the variable being matched is continuous and linear. However, given the nonlinearities inherent to the traditional norming approach of calculating averages of ratings, the specified tolerance will effectively vary across the normed distribution. The tolerance will fail, for instance, to account for floor and ceiling effects, and the accuracy of tolerances would be reduced by a failure of raw means to account for hierarchical variability. Matching conditions by norms calculated from the random effects of CLMMs will therefore improve the quality of item matches that methods like LexOPS can produce.

In chapters 2 and 3 I outlined methodological and statistical approaches that can benefit psycholinguistic research, including research into the processing of orthography, and its top-down modulation, which is the focus of the present thesis. However, this methodological work has potential to improve Psychological research more generally, as matching of factorial conditions and norming of participants or items on Likert scale responses are applied across a range of research areas.

7.7 Summary and Conclusions

In this work, I have provided evidence that, in alphabetic scripts, the N1 ERP component, which indexes early orthographic processing, is not sensitive to broad predictions of semantic categories, but is sensitive to targeted predictions for specific word forms. Such an early influence of higher-level information on low-level processing is evidence for top-down modulation of orthographic processing. I suggest that testing hypotheses informed by computational models will allow future research to better disentangle the mechanisms by which top-down contributions influence orthographic processing. I have provided an important contribution to the implementation of such models with the development and validation of sub-character measures of orthographic similarity. I have additionally developed, validated, provided, and applied methodological tools and approaches that can improve the quality of scientific inferences in this area.

A Chapter 3 Appendices

A.1 CLMMs offer no Additional Accuracy in Estimating Rank Order

I conducted an additional analysis of the results of Simulation 1, to assess whether the use of raw means of Likert ratings is appropriate if researchers are only interested in the rank order of items. Here, I examined the relationship between the rank positions of each iteration's simulated latent mean, and the estimate of its rank position from (A) raw means, and (B) estimated latent means (Figure A.1). This revealed that, indeed, rank order is relatively unaffected by differences in response patterns, and CLMMs offer no additional gain in accuracy of estimating rank positions. However, as with any continuous variable, it should be noted that ranking considerably increases noise in the relative distances between items. This is because a rank difference of 1 could be a very large difference in the original units if at a position where items are sparsely spread, or a very small difference if at a point where items are densely clustered. Therefore, when researchers are interested in the relative distances between items, I recommend the usage of ordinal models like CLMMs to appropriately account for the ordinal nature of Likert scales.

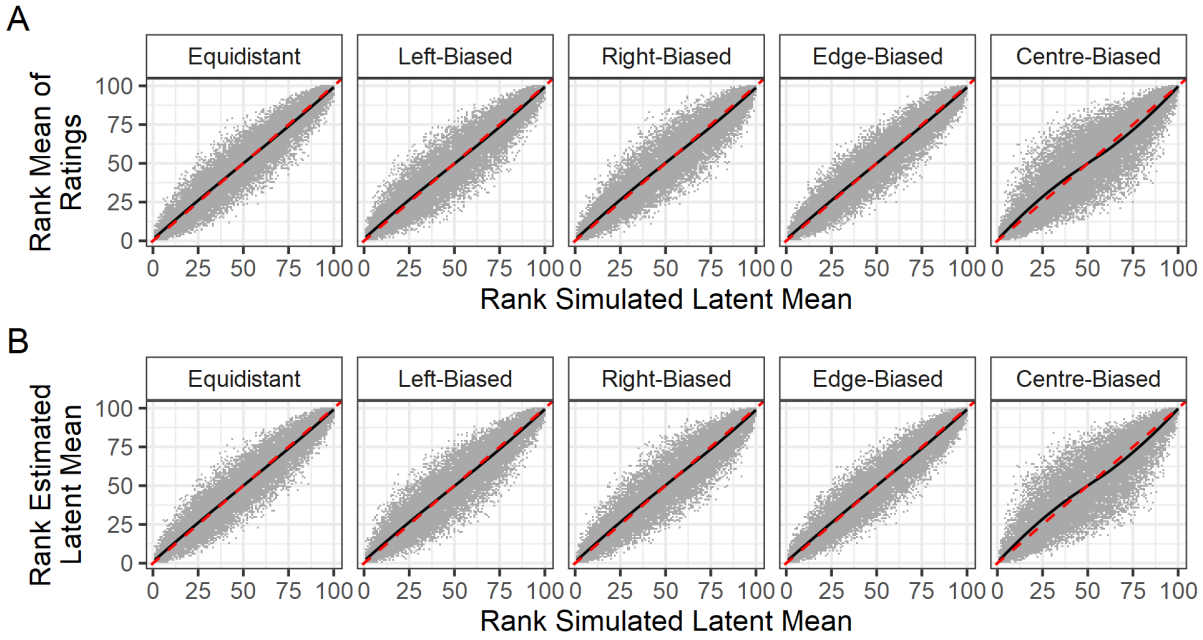


Figure A.1: The relationship between rank simulated latent means, and (A) rank raw means, or (B) rank estimated latent means (from CLMMs). Grey points show the results for individual items (ranked within their iterations). The relationships shown with the black lines were estimated via locally estimated scatterplot smoothing (LOESS), with a span parameter of .75. The dashed red lines show an expected linear relationship for reference, identical across all response patterns.

A.2 Within-Participant Z-Scores of Raw Responses do not Account for the Ordinal Nature of Likert Responses

One approach which researchers may consider when norming items on Likert ratings, where responses are provided by multiple participants, is to firstly *z*-score responses within each participant. However, I argue that such an approach still fails to account for the ordinal nature of Likert ratings, and consequently still entails the distortion in norming estimates which I have identified for averages of raw responses.

A possible justification for *z*-scoring responses within participants may be that per-item averages should be less biased by individual participants' response styles. For instance, participants who consistently respond with extreme (i.e., very high or low) ratings, will exhibit correspondingly extreme averages and low *SD*s. Within-participant z-scores for such a participant would thus align, more closely than raw responses would, with the *z*-scores of participants who responded with ratings less extreme and more variable. However, this approach still assumes that the raw Likert responses are continuous, rather than ordinal. In this way, researchers applying this approach to norming items should still expect the norms to be distorted by nonlinear response styles. To demonstrate this, I re-analysed the results from Simulation 2, comparing the performance of the CLMM approach to that of within-participant z-scores in norming items. The results (Figure A.2) showed that in assuming responses are continuous, the *z*-scoring approach results in a distortion of item norms very similar to that observed for averages of raw responses.

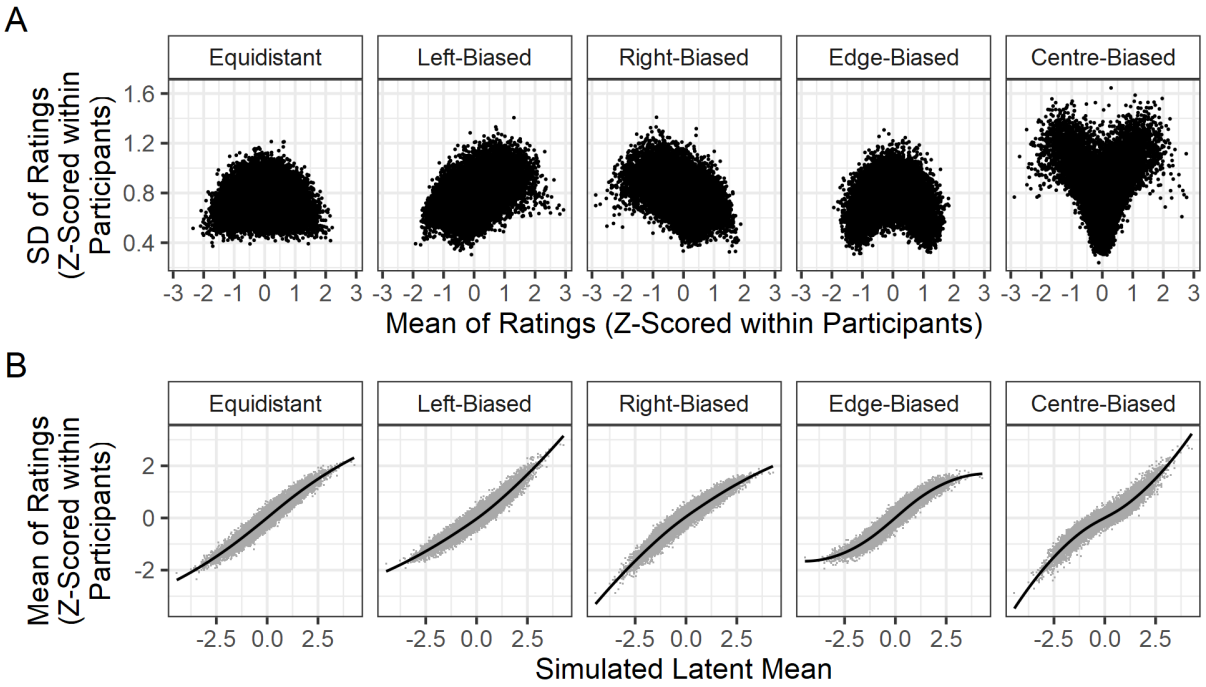


Figure A.2: The relationship between rank simulated latent means, and (A) rank raw means, or (B) rank estimated latent means (from CLMMs). Grey points show the results for individual items (ranked within their iterations). The relationships shown with the black lines were estimated via locally estimated scatterplot smoothing (LOESS), with a span parameter of .75.

The rationale behind *z*-scoring responses within participants, namely that raw means of ratings will be biased by variability between participants, is well-considered. Nevertheless, the random effects structures of CLMMs are better placed to account for between-participant variability in ordinal models, as they estimate these differences in the latent distribution, rather than assuming raw responses scale linearly. Furthermore, accounting for participant variability via random effects allows the impact of item and participant variability to be estimated simultaneously, and thus more accurately. This is preferable to the separate steps (i.e., *z*-score within participants, then calculate per-item averages) of the *z*-scoring approach. As an example, consider a design where participants each rate only a small random subset of a pool of items: here, it is likely for a single participant to be presented with items that all happen to be extremely high or low in the feature being rated. Accounting for participant variability in a separate step before calculating item norms would result in such a participant's ratings being adjusted away from the factually extreme responses, thereby reducing the magnitude of the average ratings for the items they rated. Item and participant random effects estimated in a single model, in contrast, would more accurately reflect both sources of variability. To use the example again, the ratings of participants who were presented with only extreme items would align with the ratings provided by participants who were not presented with such a biased sample of items. By estimating crossed random effects simultaneously, the partially pooled model would be able to disentangle item and participant variability, without requiring each participant to be presented with a necessarily representative sample of items. Finally, it should again be noted that pooling of the data allows the random effects structure to confer additional accuracy via shrinkage, where unlikely extreme observations are appropriately adjusted towards more likely estimates.

B Chapter 4 Appendices

B.1 LDT and SCT Stimuli

Table B.1: All stimuli presented in the experiment in chapter 4. Stimuli are ordered by semantic category (drawn from Van Overschelde et al., 2004), with category-irrelevant words, pseudowords, and nonwords presented alongside the category-relevant words they were matched with. Participants were presented with half of these 6 blocks in the LDT, and half in the SCT. *Cat.*=Semantic Category, where *1*=Relative, *2*=Four-Footed Animal, *3*=Part of the Human Body, *4*=Thing that Flies, *5*=Fruit, and *6*=Musical Instrument. Items (*String*), words frequencies (*Zipf*; van Heuven et al., 2014), average character bigram probabilities (*BG*; calculated from van Heuven et al., 2014), and orthographic neighbourhood density (*OLD20*; Yarkoni et al., 2008) are presented for all stimulus types (except for word frequency, where word frequencies are only reported for Category-Relevant and -Irrelevant words; no pseudowords or nonwords occurred in the SUBTLEX-UK corpus). Stimulus types are labelled as follows: *CR*=Category-Relevant Words, *CI*=Category-Irrelevant Words, *PW*=Pseudowords, and *NW*=nonwords. Length (number of characters) is presented as a single column, as all items for a given set had identical length. Rows are numbered for ease of reference.

	Cat.	String				Length	Zipf		BG				OLD20			
		CR	CI	PW	NW		CR	CI	CR	CI	PW	NW	CR	CI	PW	NW
1	1	uncle	video	clope	jrxvl	5	4.56	4.56	.0037	.0026	.0029	<.0001	1.75	1.50	1.00	2.95
2	1	aunt	barn	penk	kskq	4	4.32	4.32	.0050	.0048	.0058	.0004	1.00	1.00	1.00	2.00
3	1	cousin	runner	insten	mvwvsj	6	4.26	4.26	.0120	.0065	.0118	<.0001	1.70	1.20	1.85	3.05
4	1	mother	friend	protal	zjbjbt	6	5.29	5.29	.0204	.0069	.0057	<.0001	1.00	1.65	1.90	3.30
5	1	grandmother	supermarket	flispoilant	vmvgbflgpwx	11	4.33	4.24	.0140	.0054	.0064	<.0001	3.50	2.80	4.85	6.90
6	1	father	garden	lesten	mgmqtb	6	5.26	5.28	.0208	.0060	.0097	<.0001	1.35	1.00	1.45	2.95
7	1	grandfather	electricity	gristlewell	mvqsgkkwplp	11	4.46	4.44	.0142	.0051	.0063	.0003	3.40	2.70	3.90	7.00
8	1	sister	island	lanker	xxjccd	6	4.91	4.96	.0109	.0099	.0099	<.0001	1.35	1.75	1.00	3.70
9	1	brother	bedroom	flittoe	llbkqcg	7	5.01	5.00	.0177	.0054	.0062	.0015	1.60	2.25	2.10	4.00
10	1	niece	stool	meast	wjvxx	5	3.70	3.71	.0033	.0078	.0092	<.0001	1.40	1.05	1.40	2.95
11	1	nephew	galaxy	ortler	nvmppp	6	3.81	3.81	.0086	.0030	.0085	.0009	2.50	2.00	1.80	2.95
12	1	son	ice	tid	wcp	3	5.17	4.98	.0095	.0045	.0049	<.0001	1.00	1.00	1.00	1.00
13	1	daughter	industry	tratcher	zlmfzjmn	8	4.97	4.95	.0055	.0088	.0107	<.0001	2.10	2.65	1.90	4.90
1	2	dog	sun	poy	ynq	3	5.17	5.01	.0027	.0028	.0015	<.0001	1.00	1.00	1.00	1.85
2	2	cat	hat	ost	stw	3	4.83	4.76	.0074	.0129	.0065	.0057	1.00	1.00	1.00	1.00
3	2	horse	dream	penth	xjcsw	5	4.99	5.00	.0071	.0081	.0164	.0001	1.00	1.25	1.45	2.80
4	2	lion	bite	pund	crnx	4	4.45	4.45	.0080	.0083	.0058	.0009	1.00	1.00	1.00	1.70
5	2	bear	boat	voke	zqpz	4	4.88	4.89	.0088	.0044	.0023	<.0001	1.00	1.00	1.00	2.00

	Cat.	String				Length	Zipf		BG				OLD20			
		CR	CI	PW	NW		CR	CI	CR	CI	PW	NW	CR	CI	PW	NW
6	2	tiger	photo	flamp	qyhzt	5	4.36	4.42	.0086	.0063	.0024	<.0001	1.00	1.55	1.50	3.00
7	2	cow	lad	jat	ltw	3	4.44	4.43	.0059	.0039	.0050	.0008	1.00	1.00	1.00	1.00
8	2	elephant	cupboard	durledge	ppchpsxf	8	4.34	4.27	.0087	.0027	.0039	.0010	2.60	2.50	2.85	4.75
9	2	deer	chip	quib	mykc	4	4.26	4.26	.0100	.0051	.0009	.0005	1.00	1.00	1.20	1.90
10	2	mouse	opera	plail	kbdwz	5	4.41	4.41	.0088	.0073	.0038	<.0001	1.00	1.60	1.25	2.90
11	2	pig	map	nid	vbw	3	4.51	4.52	.0023	.0031	.0027	<.0001	1.00	1.00	1.00	1.30
12	2	rat	pit	dar	sgh	3	4.23	4.20	.0069	.0080	.0064	.0017	1.00	1.00	1.00	1.00
13	2	giraffe	yoghurt	bruddle	ftgvhvf	7	3.78	3.78	.0021	.0037	.0020	.0001	2.55	2.45	1.80	4.05
14	2	squirrel	ballroom	scanther	hzchrrsz	8	3.94	3.93	.0045	.0053	.0183	.0016	2.05	2.45	2.35	4.00
15	2	rabbit	button	tellow	qmhblb	6	4.39	4.46	.0046	.0077	.0067	.0004	1.35	1.00	1.25	3.05
16	2	goat	poem	spog	chbz	4	4.12	4.13	.0050	.0020	.0016	.0017	1.00	1.15	1.00	1.80
17	2	zebra	fibre	spalt	xyvlx	5	3.69	3.69	.0016	.0060	.0035	<.0001	1.80	1.55	1.55	2.90
18	2	moose	dryer	pober	hmjnh	5	3.35	3.35	.0045	.0066	.0075	<.0001	1.00	1.35	1.25	2.00
19	2	sheep	chair	glate	lkljx	5	4.67	4.66	.0109	.0068	.0058	.0002	1.00	1.00	1.35	2.90
20	2	cheetah	brewery	quintle	yzcqqcy	7	3.45	3.45	.0090	.0086	.0078	<.0001	2.20	1.85	1.90	4.00
21	2	raccoon	cutlass	gromple	wqdmybd	7	2.57	2.57	.0055	.0039	.0044	.0003	2.45	2.20	2.20	4.00
22	2	wolf	jury	pess	thjg	4	4.17	4.17	.0023	.0035	.0054	.0135	1.00	1.00	1.00	1.75
23	2	fox	jam	vay	tyf	3	4.46	4.34	.0025	.0018	.0023	.0007	1.00	1.00	1.00	1.00
24	2	hamster	missile	cromdel	nldccqn	7	3.67	3.66	.0098	.0057	.0041	.0007	1.70	1.70	2.00	3.95
25	2	donkey	boiler	altess	cjymfs	6	3.99	4.00	.0057	.0074	.0062	<.0001	1.60	1.50	1.80	2.95
26	2	elk	jig	het	stf	3	3.16	3.13	.0031	.0017	.0191	.0054	1.00	1.00	1.00	1.00
27	2	lizard	banker	baffen	yqlxyg	6	3.73	3.73	.0038	.0095	.0036	<.0001	1.60	1.00	1.85	3.70
28	2	turtle	cement	vamine	tgkcdh	6	3.64	3.64	.0039	.0074	.0080	<.0001	1.65	1.65	1.65	3.00
1	3	leg	bus	rea	dzx	3	4.88	4.74	.0044	.0041	.0140	<.0001	1.00	1.00	1.00	1.25
2	3	arm	pub	wub	pfm	3	4.70	4.71	.0058	.0008	.0003	<.0001	1.00	1.00	1.00	1.00
3	3	finger	toilet	gellor	cnqckq	6	4.53	4.54	.0128	.0066	.0069	.0005	1.00	1.65	1.75	3.00
4	3	head	room	fune	tgvn	4	5.61	5.60	.0150	.0056	.0041	<.0001	1.00	1.00	1.00	1.90
5	3	toe	bug	dut	vxl	3	4.08	4.07	.0068	.0023	.0034	<.0001	1.00	1.00	1.00	1.00
6	3	eye	bed	hix	rrp	3	5.13	5.12	.0028	.0073	.0050	.0008	1.00	1.00	1.00	1.00
7	3	hand	door	fope	rync	4	5.44	5.26	.0160	.0067	.0035	.0018	1.00	1.00	1.00	1.75

	Cat.	String				Length	Zipf		BG				OLD20			
		CR	CI	PW	NW		CR	CI	CR	CI	PW	NW	CR	CI	PW	NW
8	3	nose	moon	bist	dtzn	4	4.72	4.74	.0057	.0072	.0083	<.0001	1.00	1.00	1.00	1.90
9	3	ear	toy	zet	mfw	3	4.41	4.39	.0097	.0067	.0027	<.0001	1.00	1.00	1.00	1.00
10	3	mouth	cloud	pench	yfrtc	5	4.78	4.77	.0172	.0064	.0056	.0014	1.20	1.30	1.00	2.00
11	3	heart	phone	shiff	vlhfv	5	5.30	5.19	.0139	.0070	.0043	<.0001	1.10	1.10	1.40	2.85
12	3	knee	pipe	flen	sbpf	4	4.26	4.26	.0048	.0019	.0067	<.0001	1.35	1.00	1.00	1.85
13	3	neck	bath	hisk	scfm	4	4.65	4.65	.0044	.0175	.0075	.0004	1.00	1.00	1.10	1.40
14	3	brain	stone	spont	pqnrk	5	4.84	4.84	.0086	.0113	.0070	.0003	1.00	1.00	1.55	2.25
15	3	hair	star	weck	qcdk	4	5.03	5.04	.0074	.0086	.0043	<.0001	1.00	1.00	1.00	1.90
16	3	elbow	hatch	quind	tcdnw	5	3.85	3.85	.0035	.0078	.0101	.0002	1.85	1.20	1.50	2.60
17	3	lip	pea	weg	sjf	3	3.91	3.91	.0035	.0064	.0039	<.0001	1.00	1.00	1.00	1.00
18	3	thigh	basin	slint	bmrvz	5	3.48	3.48	.0142	.0100	.0106	.0002	1.60	1.00	1.00	2.50
19	3	ankle	wheat	trond	lynjd	5	3.94	3.95	.0074	.0143	.0091	.0012	1.40	1.50	1.25	1.95
20	3	face	city	rupe	csnz	4	5.44	5.40	.0032	.0059	.0023	<.0001	1.00	1.00	1.00	1.90
21	3	liver	diary	shipe	xsfgp	5	4.13	4.13	.0095	.0047	.0045	<.0001	1.00	1.60	1.25	2.75
22	3	lung	cart	tesh	zplg	4	3.81	3.77	.0055	.0064	.0074	.0010	1.00	1.00	1.00	1.90
23	3	tongue	pastry	wackle	ddqdsp	6	4.36	4.37	.0081	.0057	.0040	.0006	1.75	1.55	1.65	3.00
24	3	tooth	sword	adelp	gvqny	5	4.09	4.09	.0159	.0041	.0036	.0003	1.20	1.50	1.85	2.45
25	3	torso	igloo	chiff	vdrql	5	2.93	2.91	.0081	.0033	.0046	.0003	1.50	1.90	1.70	2.60
26	3	wrist	candy	serth	swzkt	5	3.84	3.83	.0073	.0095	.0178	.0001	1.10	1.00	1.45	2.00
27	3	hip	pen	rog	kpx	3	4.37	4.37	.0051	.0077	.0036	<.0001	1.00	1.00	1.00	1.00
28	3	muscle	cheque	brance	xcblvj	6	4.11	4.11	.0034	.0081	.0062	.0005	1.80	1.80	1.40	3.20
1	4	bird	ship	plub	fvdg	4	4.85	4.77	.0021	.0047	.0015	<.0001	1.00	1.00	1.05	1.90
2	4	plane	grass	sterd	jrqxk	5	4.58	4.58	.0085	.0042	.0101	<.0001	1.00	1.00	1.30	3.00
3	4	helicopter	auctioneer	hotchsting	ppkpwgvfff	10	4.38	4.40	.0096	.0070	.0081	.0005	3.45	3.05	3.20	6.15
4	4	bee	gym	lon	gcd	3	4.19	4.19	.0063	.0002	.0094	<.0001	1.00	1.00	1.00	1.00
5	4	kite	arch	gope	vrgl	4	3.89	3.88	.0084	.0054	.0035	.0004	1.00	1.00	1.00	2.00
6	4	butterfly	catalogue	rothespan	flbmmtvpz	9	4.03	3.99	.0056	.0045	.0148	.0002	2.45	2.15	3.45	5.00
7	4	mosquito	cauldron	hoshdess	zvfbgvpd	8	3.29	3.29	.0051	.0050	.0043	<.0001	2.55	2.35	2.85	4.90
8	4	bat	pin	lim	fhl	3	4.30	4.26	.0060	.0135	.0046	<.0001	1.00	1.00	1.00	1.00
9	4	eagle	canal	meach	wrprc	5	4.11	4.18	.0048	.0090	.0066	.0009	1.10	1.00	1.20	2.00

	Cat.	String				Length	Zipf		BG				OLD20			
		CR	CI	PW	NW		CR	CI	CR	CI	PW	NW	CR	CI	PW	NW
10	4	dragonfly	ballerina	swothpold	gnhlfrtjt	9	3.20	3.21	.0039	.0102	.0074	.0008	3.15	2.45	3.70	5.00
11	4	ladybird	waitress	naprator	bhlnpfbb	8	3.35	3.38	.0022	.0083	.0063	<.0001	2.75	2.10	2.70	5.00
12	4	moth	monk	fung	xcpd	4	3.64	3.63	.0165	.0064	.0055	<.0001	1.00	1.00	1.00	1.90
13	4	wasp	mast	ploe	pprm	4	3.45	3.51	.0052	.0080	.0028	.0017	1.00	1.00	1.00	1.70
1	5	apple	guest	velay	lncsw	5	4.58	4.58	.0034	.0056	.0061	.0006	1.35	1.35	1.65	2.00
2	5	banana	dancer	strile	qxmqht	6	4.21	4.22	.0091	.0094	.0065	.0004	1.60	1.20	1.45	3.85
3	5	grape	carpet	clane	vlvqt	5	3.54	4.25	.0028	.0051	.0081	<.0001	1.00	1.65	1.00	2.65
4	5	pear	tyre	trun	mrfm	4	3.83	3.83	.0078	.0069	.0027	.0002	1.00	1.00	1.00	1.75
5	5	peach	strap	flube	xdcpz	5	3.62	3.62	.0054	.0049	.0023	<.0001	1.00	1.00	1.45	2.75
6	5	strawberry	wheelchair	entrophite	rhgpxddbmv	10	3.99	4.01	.0055	.0086	.0073	<.0001	3.45	3.70	3.60	6.00
7	5	kiwi	lard	feep	htzr	4	3.27	3.28	.0022	.0056	.0029	.0007	1.45	1.00	1.00	2.00
8	5	pineapple	magazine	essendial	pwmhhlgtg	9	3.84	4.29	.0071	.0059	.0073	<.0001	3.30	2.45	3.00	4.90
9	5	plum	maid	bamp	bvsl	4	3.79	3.77	.0016	.0035	.0024	.0002	1.00	1.00	1.00	1.90
10	5	mango	wagon	prind	tcxxv	5	3.74	3.73	.0102	.0065	.0117	.0001	1.00	1.60	1.20	2.80
11	5	cherry	lawyer	lerine	qwvlpj	6	4.23	4.23	.0124	.0054	.0133	<.0001	1.55	1.70	1.45	3.65
12	5	lemon	storm	inkle	lphcq	5	4.46	4.46	.0070	.0089	.0090	.0002	1.00	1.00	1.35	2.75
13	5	blueberry	letterbox	pracement	vmpwrnhlv	9	3.17	3.15	.0044	.0058	.0060	.0005	2.55	2.90	2.40	5.00
14	5	cantaloupe	physician	rambration	ptblqmbvdj	10	2.26	3.22	.0086	.0040	.0054	.0004	3.15	2.90	3.00	5.95
15	5	raspberry	warehouse	nidgeward	vgnkllrqz	9	3.81	3.80	.0057	.0092	.0036	.0015	2.85	1.85	3.50	5.00
16	5	lime	lamp	wurn	rsdj	4	4.09	4.09	.0061	.0030	.0023	.0012	1.00	1.00	1.00	1.90
17	5	tangerine	beanstalk	prosplard	pxntlbnwk	9	3.03	3.00	.0123	.0078	.0041	.0013	2.45	2.95	3.00	4.60
18	5	melon	broth	pluff	qqjng	5	3.49	3.49	.0084	.0135	.0014	.0030	1.00	1.30	1.65	2.75
19	5	nectarine	paperclip	spimstick	fsvrvvwhc	9	2.53	2.55	.0084	.0045	.0040	.0008	2.70	3.75	3.45	5.85
20	5	papaya	folder	quanch	kqjlzt	6	2.90	2.94	.0021	.0071	.0056	<.0001	1.80	1.00	1.85	3.40
21	5	apricot	toaster	slought	wrhdcks	7	3.29	3.29	.0045	.0102	.0052	.0006	2.40	1.60	1.75	2.85
1	6	drum	desk	crun	skpm	4	4.33	4.34	.0010	.0051	.0020	.0002	1.00	1.00	1.00	1.65
2	6	guitar	pastry	shrout	lmwdmp	6	4.39	4.37	.0063	.0057	.0073	.0004	1.80	1.55	1.70	3.00
3	6	flute	apron	weast	mtfzs	5	3.48	3.48	.0041	.0062	.0088	<.0001	1.05	1.05	1.45	2.25
4	6	piano	shell	punce	jnwxl	5	4.36	4.39	.0072	.0129	.0027	<.0001	1.10	1.00	1.35	2.80
5	6	trumpet	desert	swoggle	zrdvnvw	7	3.82	4.30	.0026	.0091	.0023	.0004	1.70	1.60	1.90	4.00

	Cat.	String				Length	Zipf		BG				OLD20			
		CR	CI	PW	NW		CR	CI	CR	CI	PW	NW	CR	CI	PW	NW
6	6	clarinet	hosepipe	glutcher	wlsrcbnw	8	3.26	3.25	.0086	.0034	.0094	.0003	2.10	2.80	2.10	4.90
7	6	saxophone	scarecrow	beetdouse	kknwrbjjc	9	3.25	3.35	.0040	.0066	.0068	.0002	2.75	3.35	3.40	5.85
8	6	violin	oyster	lacket	rxvhlv	6	3.82	3.82	.0081	.0080	.0040	<.0001	1.75	1.55	1.45	3.00
9	6	trombone	armchair	incolent	bkjbbslv	8	3.27	3.31	.0058	.0056	.0095	.0002	2.50	2.80	2.60	4.15
10	6	tuba	wart	lidy	zdrm	4	2.95	2.88	.0016	.0064	.0032	.0007	1.00	1.00	1.00	1.95
11	6	cello	wedge	clush	bgcbd	5	3.46	3.45	.0060	.0048	.0028	<.0001	1.00	1.25	1.45	2.45
12	6	oboe	toga	crim	ggmv	4	2.70	2.58	.0015	.0049	.0034	.0001	1.55	1.00	1.00	1.95
13	6	bass	nail	kile	vrkz	4	4.18	4.18	.0045	.0034	.0049	.0004	1.00	1.00	1.00	2.00
14	6	viola	clove	thack	vdwqq	5	3.07	3.07	.0032	.0046	.0157	<.0001	1.35	1.00	1.15	2.95
15	6	harp	slug	snam	vdlr	4	3.40	3.39	.0090	.0010	.0018	.0002	1.00	1.00	1.00	1.85
16	6	keyboard	fountain	swentern	nfzrbhlr	8	3.64	3.67	.0034	.0101	.0083	.0001	2.40	2.15	2.40	4.00
17	6	piccolo	thicket	gruzzle	tfszjhg	7	2.56	2.55	.0035	.0112	.0018	<.0001	2.30	1.60	1.85	4.00
18	6	banjo	kayak	tassy	mbrcl	5	3.31	3.27	.0057	.0016	.0040	.0009	1.45	1.30	1.00	2.00
19	6	harmonica	projector	strebbler	jtwjmvvvq	9	3.05	3.07	.0072	.0049	.0079	<.0001	2.15	2.05	2.90	6.00
20	6	cymbal	amulet	fergle	psshtr	6	2.59	2.54	.0025	.0043	.0062	.0024	2.15	1.95	1.95	2.30
21	6	tambourine	paintbrush	sproughlay	nmkvtqsnpk	10	3.13	3.17	.0083	.0059	.0048	<.0001	2.55	3.50	3.75	6.00

B.2 Task-Stimulus Interaction over Right-Hemispheric Occipitotemporal Electrodes

Considering that effects may differ between hemispheres, I examined the pattern of effects observed over a right hemispheric occipitotemporal cluster of electrodes (Figure B.3), mirroring the locations of the left-hemispheric cluster in the main analysis. I estimated sample-level (512 Hz) linear mixed effects models to right-hemispheric occipitotemporal electrodes, as described for left-hemispheric electrodes (section 4.3.1). Model estimates (Figure B.4) and fixed-effect predictions (Figure B.5) were similar to those observed over the left hemisphere, and did not reveal robust N1 effects of the task-stimulus interaction of central interest for the study, nor a main effect of task. As was observed for left-hemispheric occipitotemporal electrodes, the interaction between task and category relevance emerged late, after the N1, and was observed later, as a smaller effect, over the right hemisphere than it was over the left hemisphere. Effects of stimulus were also similar between the left and right hemisphere, although the estimated difference between words and pseudowords in the N1 was around .5μV larger in the right hemisphere.

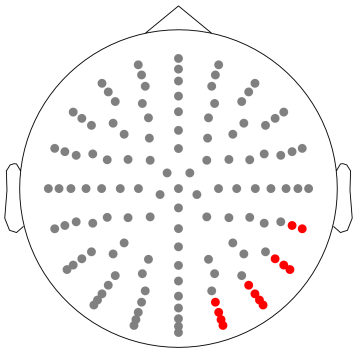


Figure B.3: The locations of the 13 right-hemispheric occipitotemporal electrodes (*red*).

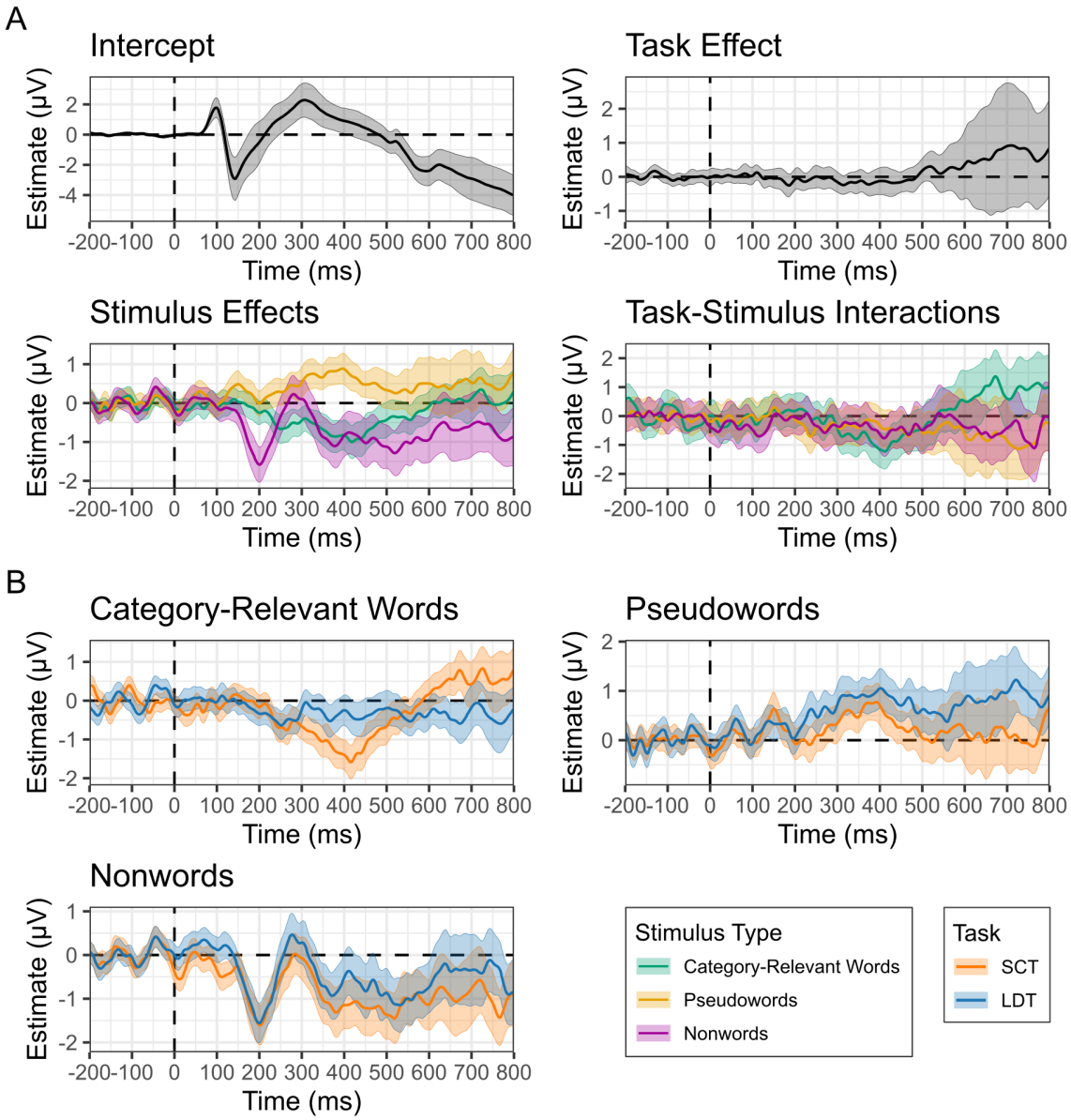


Figure B.4: Time-course of fixed effects estimates from per-sample linear mixed effects models of right-hemispheric occipitotemporal electrode voltages. (A) Fixed effects estimates from a model with all variables deviation-coded. (B) Simple effects of task for each stimulus type. In both panels, solid lines depict estimates for each sample, while shaded regions depict 95% confidence intervals.

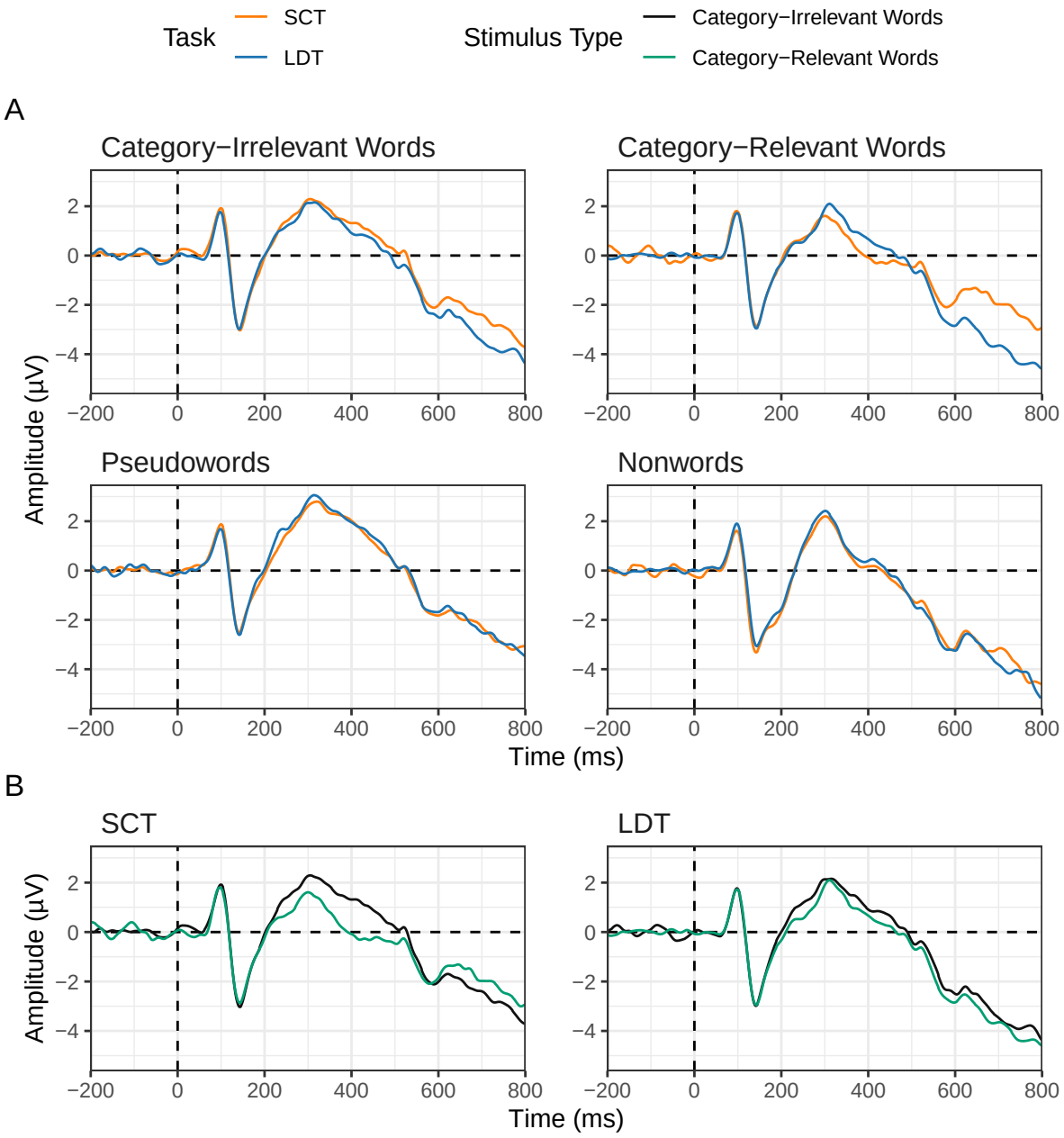


Figure B.5: Fixed-effect predictions of right-hemispheric occipitotemporal ERPs for each factorial cell, using estimates depicted in Figure B.4. These predictions are equivalent to overall average ERPs, but with the influence of random intercepts and slopes removed. Panels focus on (A) the effect of task for each stimulus type, and (B) the effect of category relevance in each task.

B.3 Task-Stimulus Interaction over Centroparietal Electrodes

In an exploratory analysis, I examined effects in a centroparietal cluster of electrodes (Figure B.6), reflecting the typical topography of the N400. I estimated sample-level (512 Hz) linear mixed effects models to centroparietal electrodes, as described for left-hemispheric occipitotemporal electrodes (section 4.3.1). Model estimates (Figure B.7) and fixed-effect predictions (Figure B.8), consistent with the scalp-wide analysis (section 4.3.2), revealed a large interaction between category relevance and task, peaking around 400 ms. Here, less negative-going amplitudes, and more positive-going amplitudes, were observed for category-relevant than -irrelevant words in the SCT after 200 ms, while these differences were not obsered in the LDT. Effects of stimulus revealed an interesting pattern, where nonwords and category-irrelevant words elicited similar ERPs, while amplitudes were more negative-going for pseudowords. Finally, it is notable that although effects peaked at around 400 ms centroparietally, the observed components did not - rather, a negative-going component peaked around 300 ms, and a positive-going component (possibly a P600) peaked between 550 and 600 ms.

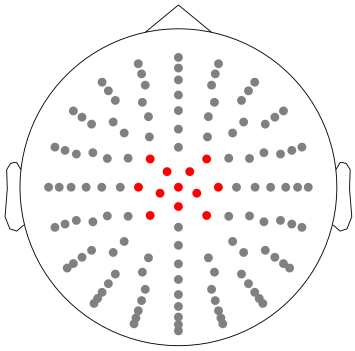


Figure B.6: The locations of the 12 centroparietal electrodes (*red*).

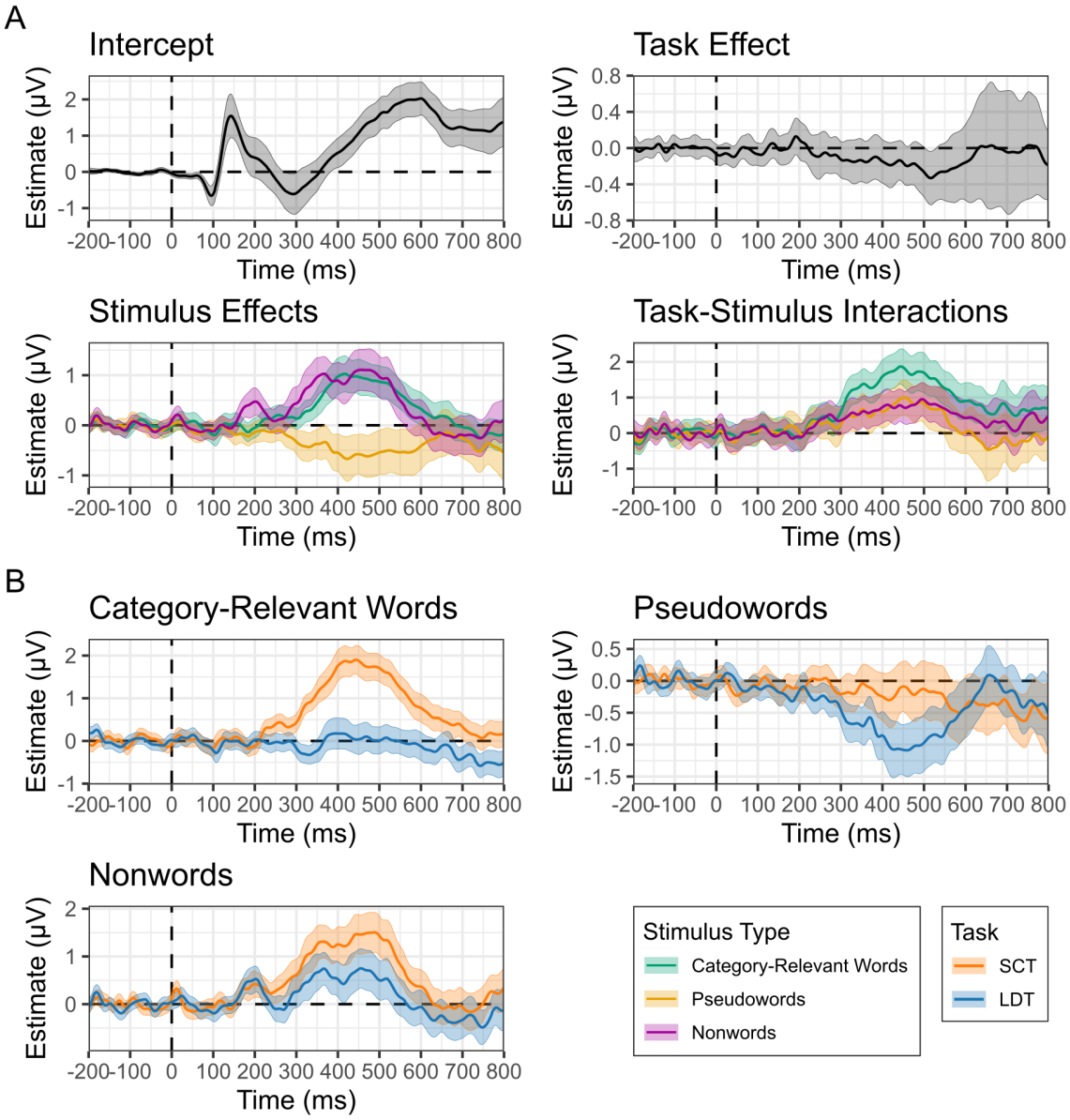


Figure B.7: Time-course of fixed effects estimates from per-sample linear mixed effects models of centroparietal electrode voltages. (A) Fixed effects estimates from a model with all variables deviation-coded. (B) Simple effects of task for each stimulus type. In both panels, solid lines depict estimates for each sample, while shaded regions depict 95% confidence intervals.

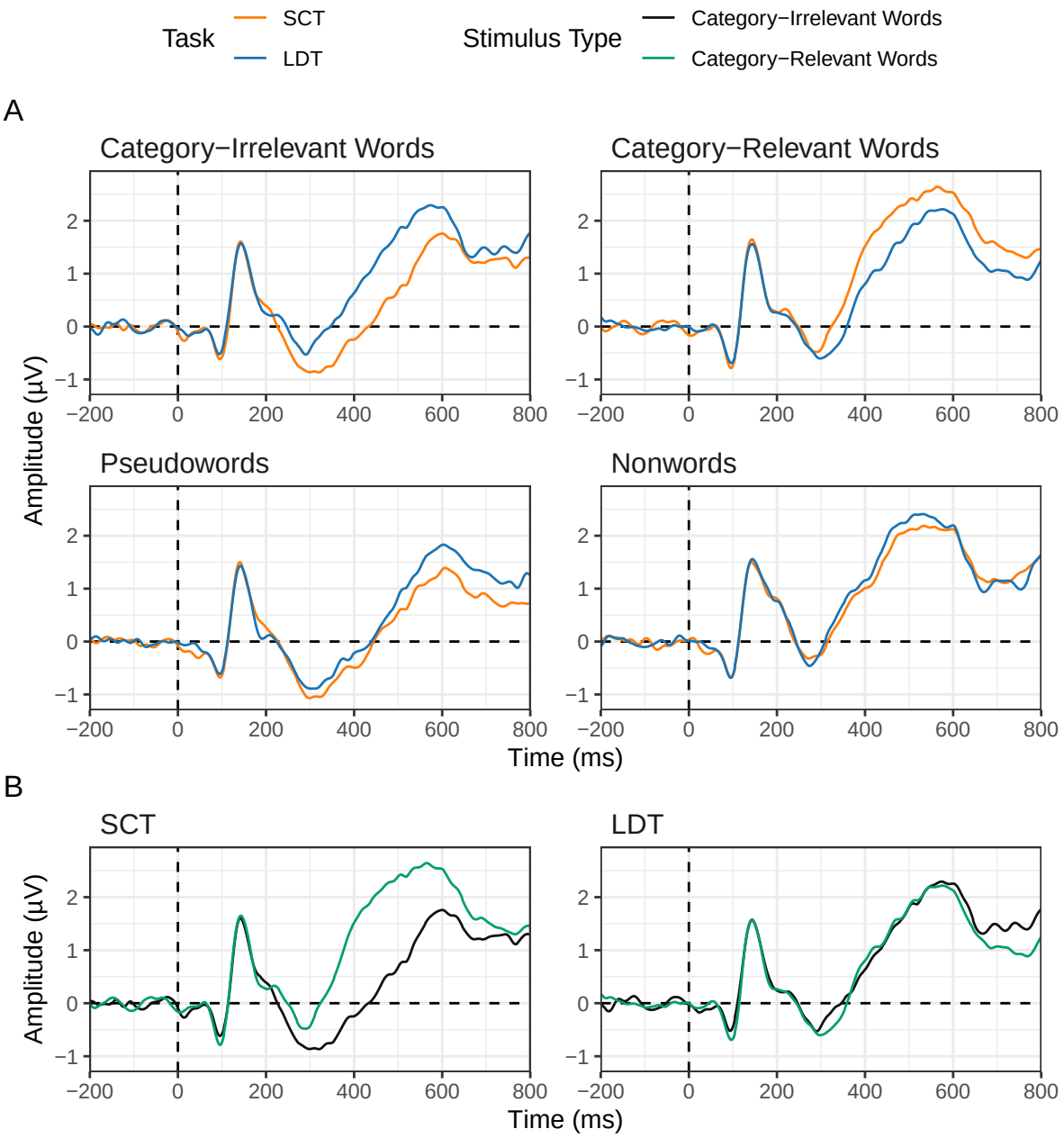


Figure B.8: Fixed-effect predictions of centroparietal ERPs for each factorial cell, using estimates depicted in Figure B.7. These predictions are equivalent to overall average ERPs, but with the influence of random intercepts and slopes removed. Panels focus on (A) the effect of task for each stimulus type, and (B) the effect of category relevance in each task.

C Chapter 5 Appendices

C.1 Picture-Word Stimuli

Table C.2: All stimuli for the picture-word task. Column Image IDs are unique file names given to each image in the BOSS, while %Agree reports the percentage of modal name agreement for the image in the BOSS. Set refers to the assigned stimulus sets. Column *Word* contains the matched congruent (*C*) and incongruent (*I*) words associated with each image. The remaining columns are as follows, separating values into those for the congruent (*C*) and incongruent (*I*) words where possible: Length = number of characters; *Zipf* = Zipf frequency in SUBTLEX-UK; *OLD20* = OLD20 values in the LexOPS dataset; *BG* = mean character bigram probabilities in SUBTLEX-UK; *CNC* = mean concreteness ratings in Brysbaert et al. (2014); *Cosine PPMI* = cosine positive pointwise mutual information values of semantic associative similarity between matched congruent and incongruent words from the Small World of Words. Rows are numbered for ease of reference.

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
1	joustingspear	7%	2	spear	porch	5	5	3.42	3.47	1.20	1.15	.0062	.0047	5.00	4.92	.0071
2	cabasa	10%	1	shaker	trough	6	6	3.35	3.32	1.00	1.65	.0092	.0067	4.11	4.17	.0073
3	powerchair	10%	1	scooter	missile	7	7	3.63	3.66	1.60	1.70	.0077	.0057	4.96	4.83	.0065
4	pottery	12%	1	pottery	rainbow	7	7	4.14	4.18	1.65	2.40	.0070	.0069	4.72	4.57	.0093
5	lbracket01	13%	1	bracket	tornado	7	7	3.40	3.49	1.75	2.10	.0036	.0059	4.43	4.53	.0003
6	flail	14%	2	mace	knob	4	4	3.37	3.50	1.00	1.35	.0042	.0029	4.81	4.75	.0017
7	plastictube	16%	1	tube	chip	4	4	4.28	4.26	1.00	1.00	.0031	.0051	4.82	4.71	.0008
8	paintscraper	16%	2	scraper	nightie	7	7	2.80	2.89	1.75	1.85	.0052	.0037	4.23	4.30	.0002
9	pillar	19%	2	pillar	sewage	6	6	3.54	3.58	1.60	1.95	.0060	.0041	4.77	4.52	.0016
10	bazooka	19%	1	bazooka	sunburn	7	7	2.76	2.86	2.55	1.85	.0015	.0026	4.66	4.57	.0037
11	chocolatecroissant	21%	1	pastry	weapon	6	6	4.37	4.29	1.55	1.90	.0057	.0067	4.97	4.76	.0011
12	solderingwire	21%	2	wire	pond	4	4	4.29	4.20	1.00	1.00	.0089	.0097	4.72	4.90	.0084
13	hedgeshears	24%	2	shears	tendon	6	6	2.94	2.98	1.40	1.55	.0120	.0103	4.61	4.47	.0013
14	pouch01b	26%	1	pouch	ledge	5	5	3.36	3.38	1.05	1.10	.0069	.0051	4.50	4.72	0
15	ram	27%	2	ram	pup	3	3	3.65	3.74	1.00	1.00	.0037	.0014	4.55	4.61	.0083
16	oats	28%	1	oats	lice	4	4	3.33	3.29	1.00	1.00	.0041	.0051	4.78	4.73	.0059
17	bandage	28%	2	bandage	whisker	7	7	3.22	3.10	1.80	1.65	.0071	.0087	4.85	4.70	.0087
18	bastingbrush	28%	2	brush	stamp	5	5	4.29	4.20	1.35	1.30	.0029	.0050	4.54	4.70	.0021
19	rug01	29%	1	rug	soy	3	3	3.57	3.64	1.00	1.00	.0014	.0028	4.79	4.70	0
20	radio01	29%	1	radio	smile	5	5	4.82	4.71	1.40	1.00	.0036	.0040	4.74	4.50	.0012

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
21	tonfa	29%	2	baton	yeast	5	5	3.53	3.64	1.00	1.50	.0097	.0076	4.64	4.72	.0036
22	salsa	29%	1	salsa	trunk	5	5	3.82	3.93	1.20	1.15	.0038	.0025	4.70	4.71	.0010
23	smokedsalmon	29%	2	salmon	tunnel	6	6	4.34	4.23	1.25	1.65	.0058	.0039	4.81	4.82	.0007
24	handmixer01d	31%	2	mixer	wedge	5	5	3.45	3.45	1.40	1.25	.0056	.0048	4.33	4.41	.0019
25	videotape01b	31%	1	cassette	revolver	8	8	2.94	2.85	1.95	1.75	.0058	.0079	4.60	4.69	.0019
26	woodboard	31%	2	wood	ship	4	4	4.78	4.77	1.00	1.00	.0034	.0047	4.85	4.87	.0097
27	jar03	33%	2	jar	lip	3	3	3.96	3.91	1.00	1.00	.0055	.0035	5.00	4.96	.0039
28	cuttingpliers02	33%	2	pliers	beanie	6	6	2.73	2.80	1.65	1.35	.0070	.0082	4.93	4.74	.0019
29	kalashnikov	33%	2	rifle	altar	5	5	3.62	3.58	1.65	1.00	.0042	.0063	4.85	4.85	.0057
30	overalls	33%	1	overalls	mongoose	8	8	3.01	2.94	2.00	2.70	.0079	.0070	4.74	4.89	.0025
31	towel01	34%	1	towel	spine	5	5	3.87	3.91	1.30	1.00	.0077	.0090	4.86	4.88	.0090
32	branch02	36%	1	branch	powder	6	6	4.10	4.17	1.15	1.55	.0064	.0065	4.90	4.76	0
33	ribbon03a	36%	2	lace	beak	4	4	3.73	3.83	1.00	1.00	.0041	.0059	4.85	4.96	.0023
34	yarn	36%	1	yarn	twig	4	4	3.14	3.22	1.00	1.20	.0041	.0028	4.93	4.75	0
35	napkin	36%	1	napkin	weasel	6	6	3.31	3.33	1.90	1.60	.0062	.0076	4.93	4.74	.0028
36	bag	36%	1	bag	oil	3	3	4.89	4.98	1.00	1.00	.0021	.0033	4.90	4.93	.0026
37	mussel	36%	2	clam	sash	4	4	3.35	3.37	1.00	1.00	.0029	.0050	4.89	4.67	.0033
38	tray	37%	1	tray	sail	4	4	4.15	4.17	1.00	1.00	.0038	.0035	4.74	4.59	.0032
39	brainmodel	38%	1	brain	river	5	5	4.84	4.93	1.00	1.00	.0086	.0094	4.69	4.89	.0014
40	megaphone	38%	1	megaphone	billiards	9	9	2.89	2.89	2.85	2.30	.0050	.0046	4.76	4.61	.0018
41	foodprocessor	38%	2	blender	javelin	7	7	3.32	3.35	1.45	1.85	.0098	.0087	5.00	4.90	.0022
42	slide02	38%	2	slide	trail	5	5	4.17	4.27	1.15	1.10	.0036	.0038	4.48	4.46	.0078
43	turnip	38%	2	turnip	nickel	6	6	3.36	3.27	1.70	1.35	.0024	.0040	4.79	4.79	.0013
44	oyster02	38%	1	oyster	canvas	6	6	3.82	3.95	1.55	1.80	.0080	.0068	4.85	4.78	.0017
45	giftbow02b	39%	1	bow	jam	3	3	4.22	4.34	1.00	1.00	.0040	.0018	4.61	4.71	.0072
46	mask02a	39%	1	mask	pony	4	4	4.04	3.96	1.00	1.00	.0046	.0059	4.96	4.90	.0045
47	bulldozer	40%	1	bulldozer	pepperoni	9	9	2.91	2.95	2.50	2.70	.0055	.0067	4.90	5.00	0
48	iceberglettuce	41%	1	lettuce	pyramid	7	7	3.81	3.71	2.40	2.50	.0038	.0022	4.97	4.96	.0031
49	leek	42%	1	leek	moat	4	4	3.56	3.69	1.00	1.00	.0047	.0045	4.92	4.69	.0013
50	scalpel	43%	1	scalpel	tequila	7	7	3.10	3.19	1.85	2.60	.0043	.0034	4.86	4.77	0

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
51	pipe	43%	2	pipe	taxi	4	4	4.26	4.30	1.00	1.00	.0019	.0016	4.88	4.93	.0011
52	glassescase	44%	1	wallet	brandy	6	6	3.81	3.83	1.20	1.25	.0074	.0077	4.81	4.81	.0038
53	coaster	44%	2	tile	mast	4	4	3.56	3.51	1.00	1.00	.0068	.0080	4.68	4.92	.0057
54	lectern01	45%	2	podium	liquid	6	6	4.17	4.25	1.85	1.75	.0018	.0022	4.89	4.72	.0032
55	doorlock	46%	2	lock	rail	4	4	4.42	4.40	1.00	1.00	.0028	.0040	4.65	4.90	.0045
56	puzzle	48%	1	puzzle	sketch	6	6	3.93	3.85	1.65	1.70	.0018	.0031	4.75	4.56	.0069
57	rhinoceros02	48%	2	rhinoceros	aftershave	10	10	3.07	3.14	3.55	3.35	.0084	.0075	4.75	4.56	.0012
58	box01a	49%	1	box	sun	3	3	5.12	5.01	1.00	1.00	.0015	.0028	4.90	4.83	.0033
59	star	50%	2	star	wall	4	4	5.04	5.05	1.00	1.00	.0086	.0079	4.69	4.86	.0015
60	scanner	50%	1	scanner	bedding	7	7	3.46	3.55	1.75	1.35	.0089	.0093	4.79	4.61	0
61	mug05	50%	2	mug	wax	3	3	3.93	3.87	1.00	1.00	.0014	.0027	4.80	4.97	.0069
62	ladle02a	51%	1	ladle	tiara	5	5	3.21	3.16	1.50	1.45	.0041	.0059	4.90	4.89	.0050
63	humanskeleton	52%	1	skeleton	tortoise	8	8	3.78	3.82	2.05	2.60	.0073	.0089	4.97	4.87	.0025
64	gecko	52%	1	lizard	barley	6	6	3.73	3.61	1.60	1.00	.0038	.0051	4.68	4.59	.0009
65	boxtrailer	52%	2	trailer	receipt	7	7	3.68	3.68	1.70	2.20	.0071	.0049	4.79	4.86	.0058
66	mechanicalpencil02	53%	1	pencil	kidney	6	6	3.98	3.94	1.90	1.70	.0047	.0030	4.88	4.96	.0086
67	spatula03	54%	2	spatula	airship	7	7	2.95	2.83	2.05	2.35	.0039	.0040	4.96	4.92	.0002
68	fusilli03a	54%	1	pasta	motor	5	5	4.19	4.25	1.00	1.60	.0066	.0082	4.86	4.84	.0022
69	bracelet01	54%	2	bracelet	postcard	8	8	3.79	3.72	2.60	2.60	.0047	.0048	4.96	4.93	.0047
70	riverotter	55%	2	otter	wrist	5	5	3.80	3.84	1.00	1.15	.0090	.0073	4.86	4.93	.0042
71	grandpiano	55%	2	piano	salad	5	5	4.36	4.38	1.10	1.00	.0072	.0049	4.90	4.97	0
72	canoepaddle02	55%	2	paddle	buzzer	6	6	3.73	3.73	1.30	1.70	.0029	.0046	4.80	4.66	.0039
73	suitcase	56%	2	suitcase	pavement	8	8	3.78	3.68	2.85	2.10	.0057	.0070	4.97	4.72	.0014
74	aquarium	57%	1	aquarium	textbook	8	8	3.33	3.38	2.45	2.75	.0028	.0028	4.77	4.86	.0026
75	trombone	57%	2	trombone	mosquito	8	8	3.27	3.29	2.50	2.55	.0058	.0051	4.90	4.88	0
76	spaghetti01	57%	1	spaghetti	underwear	9	9	3.79	3.72	3.25	2.60	.0071	.0084	5.00	4.96	.0061
77	thimble	58%	2	thimble	oregano	7	7	2.98	3.00	1.80	2.15	.0106	.0096	5.00	4.81	0
78	syringe01	58%	2	syringe	mascara	7	7	3.12	3.05	1.85	1.80	.0080	.0058	4.81	4.93	0
79	antenna	59%	2	antenna	sirloin	7	7	3.01	2.95	1.95	2.70	.0087	.0067	4.75	4.66	0
80	notebook03a	59%	1	notebook	pendulum	8	8	3.32	3.30	2.75	2.70	.0044	.0049	4.92	4.69	.0001

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
81	cleaver01	59%	2	knife	album	5	5	4.49	4.55	1.75	1.75	.0021	.0033	4.90	4.69	.0037
82	honeydewmelon	59%	2	melon	timer	5	5	3.49	3.40	1.00	1.00	.0084	.0097	4.78	4.69	.0003
83	platypus	60%	2	platypus	campfire	8	8	2.82	2.94	2.75	2.55	.0035	.0050	4.83	4.79	.0014
84	shelf	60%	1	shelf	trout	5	5	4.02	3.96	1.50	1.00	.0108	.0087	4.96	4.72	<.0001
85	macaroni01	60%	2	macaroni	bookcase	8	8	3.08	3.02	1.95	2.70	.0068	.0044	4.97	4.93	.0014
86	apricot	61%	1	peach	valve	5	5	3.62	3.53	1.00	1.55	.0054	.0050	4.90	4.83	0
87	seaturtle	62%	1	turtle	pelvis	6	6	3.64	3.53	1.65	1.75	.0039	.0048	5.00	4.93	.0012
88	triangle	62%	1	triangle	lighting	8	8	3.93	4.00	1.85	1.35	.0072	.0086	4.52	4.38	.0037
89	vulture	62%	1	vulture	measles	7	7	3.20	3.12	1.80	1.90	.0052	.0074	4.73	4.69	.0024
90	balcony02	64%	1	balcony	seaweed	7	7	3.83	3.83	1.90	1.85	.0056	.0063	4.68	4.89	.0033
91	adjustablewrench01b	64%	2	wrench	blouse	6	6	3.15	3.08	1.60	1.75	.0076	.0078	4.93	4.96	.0048
92	cane	64%	2	cane	reef	4	4	3.75	3.79	1.00	1.00	.0107	.0087	4.87	4.70	.0053
93	shield02	64%	2	shield	packet	6	6	3.80	3.90	1.70	1.45	.0051	.0036	4.66	4.46	0
94	tank	64%	1	tank	seed	4	4	4.34	4.24	1.00	1.00	.0086	.0071	4.80	4.71	.0066
95	straw	66%	2	straw	badge	5	5	4.17	4.06	1.00	1.00	.0047	.0025	4.77	4.93	.0020
96	pickle01a	66%	2	pickle	magnet	6	6	3.66	3.70	1.10	1.70	.0034	.0039	4.64	4.70	.0081
97	axe01	67%	1	axe	rum	3	3	3.85	3.91	1.00	1.00	.0001	.0010	5.00	4.93	.0023
98	boat	67%	1	boat	card	4	4	4.89	4.89	1.00	1.00	.0044	.0059	4.93	4.90	.0093
99	bowl01	67%	1	bowl	neck	4	4	4.69	4.65	1.00	1.00	.0027	.0044	4.87	5.00	.0063
100	plunger02	67%	2	plunger	caribou	7	7	3.03	2.93	1.65	1.95	.0072	.0073	4.96	4.92	.0069
101	panda	67%	1	panda	lever	5	5	3.73	3.66	1.00	1.00	.0091	.0100	4.75	4.77	.0008
102	toothpick02	67%	2	toothpick	periscope	9	9	2.79	2.81	3.35	2.65	.0090	.0069	4.93	4.78	.0011
103	kettle01	67%	2	kettle	picnic	6	6	4.02	4.04	1.45	1.90	.0041	.0027	4.75	4.83	0
104	lime	67%	2	lime	swan	4	4	4.09	3.98	1.00	1.00	.0061	.0083	4.96	4.96	.0031
105	razor01	68%	1	razor	strap	5	5	3.69	3.62	1.75	1.00	.0038	.0049	4.90	4.79	.0021
106	sailboat	69%	1	sailboat	knapsack	8	8	2.08	2.04	2.50	3.00	.0034	.0020	4.89	4.90	.0089
107	ribbon04	69%	1	ribbon	bunker	6	6	3.58	3.63	1.85	1.30	.0047	.0065	4.89	4.79	.0015
108	barn	69%	2	barn	menu	4	4	4.32	4.36	1.00	1.00	.0048	.0070	4.79	4.67	.0009
109	moon	69%	2	moon	seat	4	4	4.74	4.78	1.00	1.00	.0072	.0088	4.90	4.78	.0001
110	parrot01	69%	2	parrot	sleeve	6	6	3.84	3.88	1.65	1.70	.0053	.0054	5.00	4.84	.0016

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
111	bacon	71%	1	bacon	photo	5	5	4.34	4.42	1.00	1.55	.0067	.0063	4.90	4.93	0
112	americangoldfinch	71%	2	bird	cake	4	4	4.85	4.81	1.00	1.00	.0021	.0039	5.00	4.81	.0035
113	cheetah	71%	1	cheetah	stopper	7	7	3.45	3.39	2.20	1.55	.0090	.0083	4.70	4.83	0
114	seagull	71%	2	seagull	apricot	7	7	3.30	3.29	2.50	2.40	.0053	.0045	5.00	4.97	.0008
115	nail	72%	2	nail	sofa	4	4	4.18	4.22	1.00	1.00	.0034	.0047	4.93	4.90	.0079
116	starfish01	72%	2	starfish	armchair	8	8	3.27	3.31	2.15	2.80	.0065	.0056	4.90	5.00	.0054
117	pill	72%	1	pill	knot	4	4	3.81	3.74	1.00	1.05	.0050	.0045	4.72	4.87	.0006
118	acorn	73%	1	acorn	bugle	5	5	3.13	3.01	1.65	1.25	.0056	.0033	4.96	4.84	.0065
119	shorts01	74%	1	shorts	needle	6	6	3.82	3.93	1.35	1.55	.0052	.0058	4.82	4.93	.0018
120	tripod01	74%	1	tripod	seesaw	6	6	3.04	2.97	1.85	1.95	.0029	.0053	4.72	4.92	0
121	cabbage	74%	2	cabbage	uniform	7	7	4.07	4.16	1.65	2.00	.0027	.0043	4.75	4.67	.0053
122	raccoon	74%	2	raccoon	notepad	7	7	2.57	2.55	2.45	2.80	.0055	.0046	4.67	4.70	.0004
123	dormer	76%	1	window	letter	6	6	4.84	4.85	1.40	1.00	.0106	.0087	4.86	4.70	.0094
124	volleyball	76%	1	volleyball	chimpanzee	10	10	3.31	3.19	3.80	3.70	.0050	.0052	4.93	4.96	0
125	cocktailshrimp02	76%	2	shrimp	tablet	6	6	3.63	3.54	1.80	1.65	.0030	.0045	4.80	4.82	.0042
126	bowrake	76%	1	rake	yolk	4	4	3.40	3.52	1.00	1.20	.0035	.0040	4.84	4.78	.0043
127	tulip02	76%	1	tulip	llama	5	5	3.21	3.12	1.70	1.60	.0031	.0054	5.00	4.78	.0022
128	tie02	79%	1	tie	map	3	3	4.58	4.52	1.00	1.00	.0051	.0031	4.81	4.93	.0008
129	popcorn	79%	1	popcorn	luggage	7	7	3.68	3.61	2.60	2.55	.0038	.0017	5.00	4.83	0
130	pigeon	79%	1	pigeon	muscle	6	6	4.03	4.11	1.70	1.80	.0048	.0034	4.71	4.50	.0027
131	honeybee	79%	1	bee	lid	3	3	4.19	4.16	1.00	1.00	.0063	.0044	4.88	4.96	0
132	callbell	79%	2	bell	oven	4	4	4.54	4.54	1.00	1.00	.0073	.0079	4.96	4.97	.0066
133	teapot	79%	1	teapot	mousse	6	6	3.78	3.76	1.90	1.35	.0054	.0076	4.96	4.83	.0039
134	rope03	79%	1	rope	text	4	4	4.30	4.40	1.00	1.10	.0040	.0035	4.93	4.93	0
135	marble	80%	2	marble	puppet	6	6	3.86	3.77	1.50	1.70	.0051	.0027	4.85	4.64	.0094
136	boot02b	82%	2	boot	page	4	4	4.43	4.52	1.00	1.00	.0043	.0028	4.96	4.90	.0009
137	plum01	82%	1	plum	ramp	4	4	3.79	3.67	1.00	1.00	.0016	.0030	4.85	4.69	.0047
138	tampon	82%	1	tampon	poncho	6	6	2.30	2.41	1.80	1.65	.0051	.0059	4.86	4.97	.0076
139	slipper01b	82%	2	slipper	warship	7	7	3.19	3.09	1.40	1.85	.0054	.0056	4.86	4.86	0
140	chalk	82%	2	chalk	organ	5	5	3.87	3.99	1.30	1.00	.0077	.0080	4.90	4.77	.0018

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
141	banjo	83%	2	banjo	scalp	5	5	3.31	3.28	1.45	1.35	.0057	.0040	4.90	4.82	.0050
142	peanut01	83%	2	peanut	bumper	6	6	3.65	3.53	1.95	1.40	.0078	.0058	4.89	4.96	.0091
143	pillow01a	84%	2	pillow	beetle	6	6	3.73	3.72	1.60	1.60	.0050	.0053	5.00	4.83	.0068
144	cigar	85%	2	cigar	stump	5	5	3.59	3.52	1.75	1.25	.0041	.0037	4.93	4.78	0
145	jellyfish	86%	2	jellyfish	sunflower	9	9	3.56	3.55	3.10	2.95	.0049	.0054	4.93	4.80	.0096
146	calendar	86%	2	calendar	medicine	8	8	4.33	4.31	2.30	2.10	.0085	.0084	4.62	4.79	.0017
147	bull	86%	1	bull	cave	4	4	4.28	4.19	1.00	1.00	.0052	.0064	4.85	4.96	.0058
148	daddylonglegs	86%	1	spider	tongue	6	6	4.24	4.36	1.25	1.75	.0059	.0081	4.97	4.93	.0083
149	chimney	86%	2	chimney	bicycle	7	7	3.90	3.92	1.85	2.40	.0047	.0027	5.00	4.89	.0098
150	ashtray01	87%	2	ashtray	brownie	7	7	3.20	3.29	2.30	1.75	.0043	.0033	4.97	4.82	.0055
151	binoculars01b	87%	1	binoculars	ammunition	10	10	3.59	3.61	3.45	3.00	.0065	.0057	5.00	4.88	.0099
152	baseball01a	87%	2	baseball	cinnamon	8	8	3.74	3.78	2.55	2.55	.0057	.0072	4.86	4.85	0
153	broom01	87%	1	broom	algae	5	5	3.56	3.44	1.15	1.60	.0045	.0026	4.89	4.93	.0051
154	balloon01b	87%	1	balloon	stomach	7	7	4.25	4.28	1.65	1.95	.0073	.0072	4.92	4.89	.0071
155	avocado01	87%	1	avocado	sparrow	7	7	3.25	3.38	2.55	1.80	.0031	.0046	4.89	4.85	.0025
156	sock01a	87%	2	sock	tuna	4	4	3.77	3.76	1.00	1.00	.0029	.0025	4.91	4.89	.0047
157	jeans01	88%	1	jeans	wagon	5	5	3.84	3.73	1.30	1.60	.0078	.0065	5.00	4.89	.0012
158	nose	88%	2	nose	mail	4	4	4.72	4.63	1.00	1.00	.0057	.0042	4.89	4.69	.0017
159	knee	88%	2	knee	soil	4	4	4.26	4.35	1.35	1.00	.0048	.0039	5.00	4.87	.0075
160	stool01	88%	2	stool	weeds	5	5	3.71	3.66	1.05	1.00	.0078	.0054	4.90	4.83	0
161	jeep	88%	1	jeep	wick	4	4	3.17	3.16	1.00	1.00	.0023	.0039	4.80	4.69	.0004
162	cannon	88%	2	cannon	throat	6	6	4.08	4.16	1.15	1.70	.0092	.0116	4.79	4.97	.0022
163	ostrich	88%	2	ostrich	shuttle	7	7	3.52	3.58	2.10	1.80	.0053	.0036	4.71	4.63	.0077
164	porcupine	88%	1	porcupine	lawnmower	9	9	3.06	3.11	3.25	3.45	.0064	.0052	5.00	4.97	.0023
165	arrow02	90%	2	arrow	jewel	5	5	3.78	3.76	1.00	1.75	.0059	.0035	4.97	4.96	.0006
166	tricycle	90%	2	tricycle	songbird	8	8	2.73	2.75	2.60	2.80	.0033	.0053	4.68	4.59	.0062
167	sponge01	90%	2	sponge	timber	6	6	4.12	4.05	1.45	1.40	.0068	.0075	5.00	4.90	.0002
168	celery	92%	1	celery	tattoo	6	6	3.66	3.78	1.90	1.85	.0082	.0067	4.80	4.71	.0039
169	violin	92%	1	violin	burger	6	6	3.82	3.90	1.75	1.15	.0081	.0065	4.96	4.93	.0014
170	iron01b	92%	1	iron	soup	4	4	4.52	4.41	1.00	1.00	.0078	.0086	4.59	4.72	.0060

	Image ID	%Agree	Set	Word		Length		Zipf		OLD20		BG		CNC		Cosine PPMI
				C	I	C	I	C	I	C	I	C	I	C	I	
171	lamp04a	92%	1	lamp	wool	4	4	4.09	4.11	1.00	1.00	.0030	.0036	4.97	4.86	.0093
172	scarf	92%	2	scarf	patio	5	5	3.76	3.73	1.05	1.35	.0043	.0058	4.97	4.89	.0026
173	microscope	92%	2	microscope	spacecraft	10	10	3.56	3.46	2.50	3.25	.0035	.0025	5.00	4.80	.0083
174	rice	92%	1	rice	bomb	4	4	4.42	4.49	1.00	1.00	.0050	.0031	4.86	4.84	.0075
175	rooster	93%	1	rooster	serpent	7	7	3.13	3.16	1.50	1.70	.0087	.0088	4.75	4.97	.0008
176	beaver	93%	1	beaver	shrine	6	6	3.50	3.53	1.00	1.55	.0098	.0088	4.68	4.47	.0097
177	trophy01	93%	2	trophy	jacket	6	6	4.37	4.29	1.90	1.40	.0025	.0032	4.89	4.86	.0032
178	cactus	93%	2	cactus	poodle	6	6	3.35	3.27	1.70	1.45	.0037	.0035	5.00	4.89	0
179	snowboard	95%	2	snowboard	amplifier	9	9	2.84	2.73	2.65	2.65	.0035	.0051	4.86	4.79	.0076
180	potato02b	95%	1	potato	ticket	6	6	4.44	4.51	1.60	1.35	.0071	.0048	4.85	4.70	.0086
181	apple07	95%	1	apple	penny	5	5	4.58	4.49	1.40	1.00	.0034	.0044	5.00	4.83	.0080
182	apron	95%	2	apron	lager	5	5	3.48	3.56	1.05	1.00	.0062	.0075	4.87	4.64	.0001
183	cigarette	95%	2	cigarette	porcelain	9	9	4.11	4.10	2.80	2.90	.0065	.0071	4.88	4.63	.0091
184	skunk	95%	1	skunk	quail	5	5	3.36	3.48	1.55	1.45	.0016	.0024	4.88	4.65	.0054
185	barnowl	95%	2	owl	jug	3	3	4.07	4.06	1.00	1.00	.0026	.0016	4.93	4.96	.0095
186	lipstick02a	95%	1	lipstick	cardigan	8	8	3.62	3.50	2.30	1.90	.0047	.0064	4.90	4.96	.0096
187	brick	95%	1	brick	robot	5	5	4.18	4.09	1.00	1.60	.0036	.0040	4.83	4.65	.0017
188	leaf02a	97%	2	leaf	pork	4	4	4.29	4.39	1.00	1.00	.0059	.0048	5.00	4.79	0
189	carrot01	97%	2	carrot	tissue	6	6	4.08	3.97	1.40	1.75	.0059	.0051	5.00	4.93	.0053
190	kite	98%	2	kite	cart	4	4	3.89	3.77	1.00	1.00	.0084	.0064	5.00	4.89	.0004
191	locker	98%	1	locker	manual	6	6	3.60	3.70	1.00	1.75	.0064	.0069	4.67	4.45	.0048
192	pumpkin	98%	2	pumpkin	trolley	7	7	3.79	3.82	1.70	1.70	.0052	.0055	4.90	4.73	.0025
193	zebra	98%	2	zebra	snail	5	5	3.69	3.69	1.80	1.45	.0016	.0026	4.86	4.93	.0062
194	kangaroo	98%	1	kangaroo	lemonade	8	8	3.62	3.57	2.75	2.70	.0077	.0055	4.86	4.83	.0058
195	squirrel	100%	1	squirrel	passport	8	8	3.94	4.01	2.10	2.25	.0045	.0045	4.89	5.00	.0051
196	mushroom01	100%	2	mushroom	carriage	8	8	3.87	3.98	2.60	1.90	.0040	.0044	4.83	4.86	.0003
197	pear01	100%	1	pear	lung	4	4	3.83	3.81	1.00	1.00	.0078	.0055	4.93	4.82	.0050
198	snowman	100%	1	snowman	pancake	7	7	3.52	3.48	1.90	2.05	.0060	.0059	4.64	4.86	.0054
199	onion	100%	1	onion	torch	5	5	4.28	4.21	1.70	1.30	.0086	.0073	4.86	4.76	.0013
200	toothbrush03b	100%	2	toothbrush	cheesecake	10	10	3.48	3.54	3.80	3.45	.0084	.0085	5.00	4.97	.0040

C.2 Details on the Shifted Log-Normal Bayesian Model Analysis of the Stimuli Validation RT Data

The shifted log-normal Bayesian model was fit to the RT data from the stimulus validation with *brms* (Bürkner, 2018), a high-level interface for STAN (STAN Development Team, 2021). This model estimated the plausibility of values for each parameter of the shifted log-normal distribution as a function of the maximal hierarchical structure justified by the experiment's design. The parameter of μ was modelled with an identity link function, while σ and δ were modelled with log link functions. The same predictors and random effects structure were used for each parameter as described for the EEG experiment, though with a key difference being that predictability was normalised between 12% and 100% rather between than 7% and 100%, due to different minima in the experiments' stimuli. The full formula, in *brms* syntax, was specified as:

```
rt ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id),
sigma ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id),
ndt ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id)
```

Prior distributions were specified to be broad enough as to be uninformative but constrained to be cover plausible values for response time distributions for a cognitive task (Figure C.9A). Fixed effects' slopes' prior distributions were drawn from $N(0, 2.5)$ and fixed effects' intercepts' prior distributions from $N(0, 7.5)$. The prior distributions for the standard deviations of random effects were specified as student's *t* distributions centred on zero, with 3 degrees of freedom and a scale parameter of 2. The model was fit with 5 Markov chains, each with 25,000 (17,500 warm-up and 7,500 sampling) iterations. The *adapt_delta* parameter was set to .99. The densities of the posterior distributions, relative to those of the priors, are shown in Figure C.9B.

In addition to the shifted log-normal Bayesian model, a Gamma family (identity link function) generalised linear mixed effects model (GLMM) was fit to the RT data from the stimulus validation using *lme4* (Bates et al., 2015). The same predictors and random effects structure were used as in the shifted log-normal distribution and as outlined in the power analysis for the EEG experiment. The GLMM estimated random intercepts and slopes (but no random correlations, to deal with non-convergence), using the maximal random effects structure justified by the experiment's design. Though less sensitive to changes in the distributions

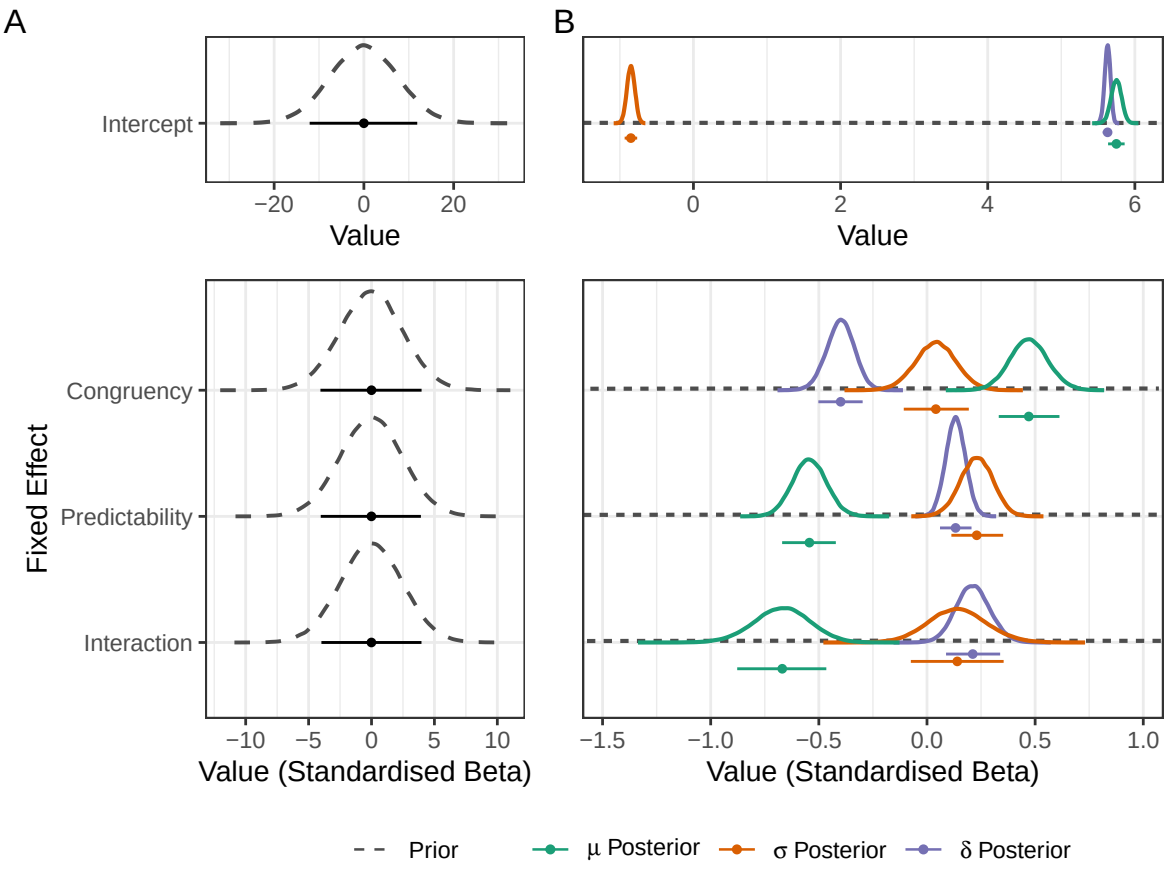


Figure C.9: Prior and posterior distributions for all fixed effects estimated in the Bayesian shifted log-normal model presented in the Stimulus Validation section. (A) Distributions of prior samples for all fixed effects. For a given fixed effect, priors were identical for all parameters of the shifted log-normal distribution so are concatenated here for simplicity. (B) Posterior distributions for all fixed effects, superimposed on the relative densities of the prior distributions (with limits of the x axis set to increase the visibility of the posterior distributions). For both panels, points below distributions' densities depict median posterior estimates, while the whiskers show the extents of 89% HDIs.

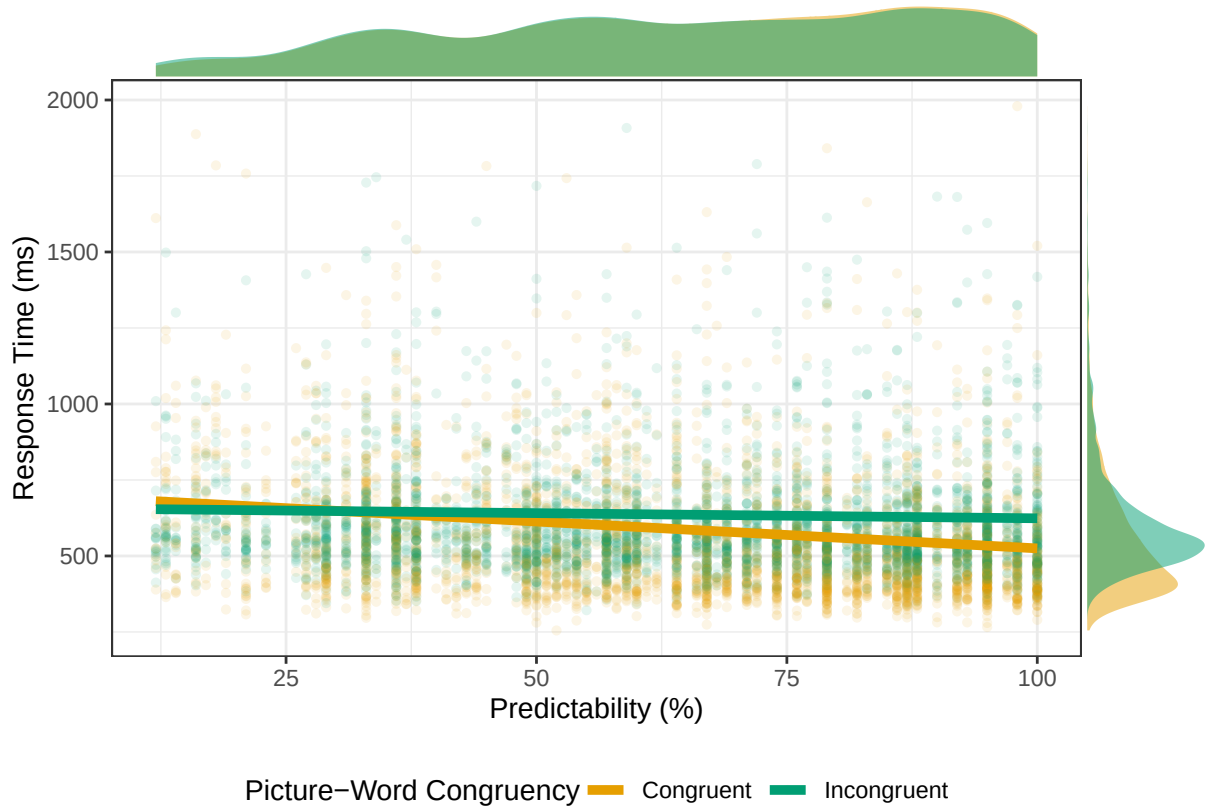


Figure C.10: Relationship between predictability and average response time. Points depict individual trial-level observations, while lines depict the linear relationships as estimated by the Gamma family GLMM. Marginal plots depict the density of each axis' variable, for congruent (*orange*) and incongruent (*green*) trials.

that do not reflect changes in central tendency, the results generally corroborated those of the shifted log-normal distribution. The model, and Chi-square model comparisons of effects, suggested there is a greater effect of predictability for congruent than incongruent trials (β=125.32, *SE*=5.12, $\chi^2(1)$=24.3, p<.001). Post-hoc tests (with Bonferroni-corrected p values reported as p_{bonf}) revealed that predictability is negatively related to response times for congruent trials (β=-156, *SE*=4.25, $\chi^2(1)$=32.51, p<.001, p_{bonf}<.001), whereas the effect of predictability for incongruent trials is notably smaller (β=-30.83, *SE*=12.45, $\chi^2(1)$=2.99, p=.084, p_{bonf}=.336). Conversely, the effect of congruency is large and significant, and in the predicted direction, at the highest level of predictability, with faster responses to congruent items (β=-99.75, *SE*=5.84, $\chi^2(1)$=34.99, p<.001, p_{bonf}<.001), but is small and non-significant, and in the opposite direction, at the lowest level of predictability (β=26.44, *SE*=4.25, $\chi^2(1)$=2.68, p=.101, p_{bonf}=.406). These linear relationships in the fixed effects are presented in Figure C.10.

C.3 Word Stimuli for Localiser Task

Table C.3: All word stimuli for the localiser task, and associated values on variables that were matched distribution-wise. False-font strings and phase-shuffled images are not presented here; false-font strings ware just the words in BACS2serif font, while a unique phase-shuffled image was generated for each trial. The columns are as follows: *Word* = words presented in the task; *Length* = number of characters; *Zipf* = Zipf frequency in SUBTLEX-UK; *PREV* = word prevalence values in Brysbaert et al. (2019); *OLD20* = OLD20 values in the LexOPS dataset; *BG* = mean character bigram probabilities in SUBTLEX-UK; *PoS* = dominant part of speech in SUBTLEX-UK; *CNC* = mean concreteness ratings in Brysbaert et al. (2014); *AoA* = mean age of acquisition ratings in Kuperman et al. (2012); *VAL*, *AROU*, and *DOM* = mean valence, arousal, and dominance ratings, respectively, from Warriner et al. (2013); *LDT RT* and *LDT Acc* = average response times (in ms) and accuracies in lexical decision, from the BLP. Rows are numbered for ease of reference.

	Word	Length	Zipf	PREV	OLD20	BG	PoS	CNC	AoA	VAL	AROU	DOM	LDT RT	LDT Acc
1	tracker	7	3.12	2.58	1.45	.0062	noun	3.89	9.61	4.87	4.59	5.00	583.21	.98
2	tablespoonful	13	1.97	1.40	5.30	.0045	adjective	4.24	7.58	-	-	-	-	-
3	curricular	10	2.50	1.61	2.85	.0042	adjective	2.77	10.10	-	-	-	-	-
4	sheathed	8	1.74	1.45	1.95	.0195	verb	3.04	-	-	-	-	699.78	.68
5	wasabi	6	2.74	1.74	2.00	.0041	noun	4.67	13.95	-	-	-	-	-
6	persecute	9	2.43	1.80	2.65	.0068	verb	2.53	10.06	3.11	5.11	4.09	-	-
7	enlarge	7	2.70	2.16	2.35	.0053	verb	3.17	8.26	5.33	3.87	5.89	568.70	.95
8	harvester	9	3.15	2.12	2.60	.0107	noun	4.21	9.53	-	-	-	-	-
9	campaign	8	4.90	2.44	2.20	.0027	noun	3.00	12.55	4.55	3.50	5.14	561.37	.98
10	menacingly	10	2.20	1.79	3.25	.0078	adverb	1.93	-	-	-	-	-	-
11	footwork	8	3.40	2.13	2.15	.0044	noun	3.32	10.63	5.74	3.96	5.58	680.59	.88
12	respective	10	3.20	2.10	2.65	.0067	adjective	1.79	10.78	5.90	3.76	6.42	-	-
13	layperson	9	1.65	1.35	2.85	.0068	noun	3.44	13.74	-	-	-	-	-
14	microcomputer	13	1.30	1.82	4.45	.0055	noun	4.55	13.89	-	-	-	-	-
15	flatterer	9	2.32	1.32	1.85	.0104	noun	2.89	12.44	-	-	-	-	-
16	chilled	7	3.63	2.35	1.75	.0074	verb	3.22	-	-	-	-	566.50	1.00
17	blackheads	10	1.93	2.07	2.35	.0065	noun	4.79	-	-	-	-	742.83	.97
18	fortunate	9	4.06	2.24	2.50	.0056	adjective	2.04	10.17	7.33	3.81	5.83	635.46	.95
19	screeching	10	2.81	2.24	2.55	.0092	verb	3.71	-	-	-	-	621.72	.93
20	chimp	5	3.42	2.23	1.35	.0048	noun	4.96	7.17	6.00	3.80	4.95	605.63	.88
21	payroll	7	3.10	2.43	2.40	.0042	noun	3.70	12.79	6.19	3.82	5.11	632.25	.97

	Word	Length	Zipf	PREV	OLD20	BG	PoS	CNC	AoA	VAL	AROU	DOM	LDT RT	LDT Acc
22	seer	4	2.48	1.26	1.00	.0110	noun	-	10.56	5.35	3.77	5.41	752.86	.53
23	coexist	7	2.00	1.99	2.45	.0053	verb	2.25	11.56	5.95	3.48	5.92	-	-
24	smelly	6	3.87	2.43	1.45	.0057	adjective	3.07	4.32	2.68	5.43	4.00	533.24	1.00
25	discouraging	12	2.69	2.33	3.25	.0084	verb	1.83	9.11	2.89	4.17	4.22	-	-
26	exotic	6	4.10	2.43	1.85	.0039	adjective	2.11	10.42	7.55	6.90	5.65	-	-
27	snow	4	4.79	2.33	1.00	.0040	noun	4.85	4.11	6.78	4.57	5.62	506.10	1.00
28	takeoff	7	2.84	1.92	2.45	.0035	noun	3.41	7.35	5.50	3.77	5.11	-	-
29	milkman	7	3.08	1.98	1.90	.0054	noun	4.61	6.37	5.75	2.73	5.54	626.19	1.00
30	intelligent	11	4.09	2.58	3.15	.0094	adjective	2.46	8.28	7.60	5.67	6.77	-	-
31	creak	5	2.69	1.40	1.30	.0078	verb	3.61	8.10	4.68	4.40	4.61	599.59	.85
32	punchy	6	3.03	1.51	1.55	.0024	adjective	2.21	13.18	4.78	4.32	3.96	657.00	.76
33	glutinous	9	2.09	1.53	2.70	.0089	adjective	2.62	14.32	-	-	-	-	-
34	monsieur	8	3.70	1.35	2.75	.0046	noun	3.54	10.12	5.50	3.30	5.89	-	-
35	sympathetic	11	3.70	2.58	3.50	.0105	adjective	1.77	9.39	6.67	3.29	6.30	-	-
36	neurotoxin	10	1.95	1.72	3.10	.0071	noun	3.12	13.58	-	-	-	-	-
37	singular	8	3.00	2.27	2.45	.0086	adjective	2.21	9.80	4.89	3.12	5.24	-	-
38	snip	4	3.64	2.00	1.00	.0012	noun	3.68	7.24	4.32	4.74	4.95	569.42	.95
39	bewildered	10	3.14	2.43	3.30	.0080	verb	1.80	11.63	4.32	4.57	4.42	-	-
40	devote	6	3.16	2.03	1.55	.0045	verb	2.00	9.58	5.53	4.05	7.05	600.51	.97
41	handily	7	2.30	1.62	1.90	.0101	adverb	2.08	-	-	-	-	-	-
42	orally	6	2.41	2.23	1.90	.0076	adverb	3.00	-	-	-	-	-	-
43	prerecorded	11	1.60	2.10	3.45	.0096	verb	2.58	10.22	-	-	-	-	-
44	yodel	5	3.11	1.49	1.55	.0054	name	4.20	8.16	6.10	3.33	5.90	703.75	.50
45	impertinently	13	1.17	1.38	3.90	.0082	adverb	-	-	-	-	-	-	-
46	vacation	8	3.36	2.58	1.85	.0063	noun	3.14	5.22	8.53	5.22	7.11	-	-
47	extravagance	12	2.86	2.20	3.85	.0038	noun	1.73	10.74	5.74	5.40	5.79	-	-
48	thud	4	3.01	2.26	1.00	.0139	noun	3.20	8.06	4.24	5.05	4.52	582.36	.83
49	forewarn	8	1.74	1.90	2.10	.0076	verb	2.20	11.16	-	-	-	703.91	.66
50	fatherhood	10	2.73	2.44	3.20	.0130	noun	2.76	8.50	6.77	4.57	5.61	-	-
51	correlate	9	2.20	2.04	2.60	.0083	verb	1.63	13.35	-	-	-	-	-
52	watercraft	10	1.54	1.61	2.90	.0056	noun	-	-	-	-	-	-	-
53	sunk	4	3.73	2.43	1.00	.0026	verb	3.46	-	-	-	-	611.78	.93

	Word	Length	Zipf	PREV	OLD20	BG	PoS	CNC	AoA	VAL	AROU	DOM	LDT RT	LDT Acc
54	flawlessness	12	1.39	1.58	3.30	.0042	noun	2.16	-	-	-	-	-	-
55	tranquilizer	12	1.47	2.02	2.30	.0054	noun	4.55	11.58	4.86	3.12	4.85	-	-
56	pituitary	9	2.47	1.52	3.70	.0065	adjective	3.33	13.06	4.79	4.40	4.91	-	-
57	courtside	9	1.81	2.00	2.85	.0059	noun	3.65	12.32	6.00	4.24	6.00	-	-
58	wicked	6	4.16	2.33	1.15	.0047	adjective	2.11	8.33	2.63	5.86	3.61	579.31	.93
59	regard	6	4.19	2.24	1.55	.0067	noun	1.79	10.20	5.70	3.39	6.38	545.31	.98
60	infidelity	10	2.71	2.33	3.55	.0072	noun	2.07	13.89	2.10	5.70	3.86	-	-
61	bumping	7	3.29	2.34	1.55	.0074	verb	4.00	-	-	-	-	660.94	.97
62	cannibal	8	2.60	2.31	2.45	.0058	adjective	3.82	9.11	2.90	6.10	3.20	-	-
63	texting	7	3.51	2.58	1.80	.0093	verb	4.23	-	-	-	-	-	-
64	apache	6	3.15	1.75	1.75	.0091	name	3.88	10.50	5.20	3.70	4.95	747.23	.68
65	generational	12	2.98	1.88	2.90	.0084	adjective	1.96	12.68	-	-	-	-	-
66	squint	6	2.79	2.33	1.75	.0075	noun	4.30	8.05	4.40	3.71	4.62	586.76	1.00
67	torture	7	4.00	2.43	1.80	.0089	verb	3.59	10.70	1.40	5.09	2.76	530.51	1.00
68	shattering	10	3.10	2.32	1.75	.0115	verb	3.43	8.00	3.67	5.00	4.63	-	-
69	freckled	8	1.30	2.43	1.90	.0061	adjective	3.86	6.58	-	-	-	645.19	.98
70	perversion	10	2.35	2.07	2.70	.0087	noun	2.04	13.11	3.55	5.48	3.85	-	-
71	shag	4	3.37	2.00	1.00	.0073	noun	3.15	10.53	5.38	4.95	4.86	546.18	.98
72	stifle	6	2.68	1.97	1.70	.0058	verb	2.59	10.26	-	-	-	659.20	.82
73	syllable	8	2.89	2.25	2.00	.0038	adjective	3.26	8.10	4.95	2.50	5.70	-	-
74	ionic	5	2.50	1.79	1.40	.0063	adjective	2.14	14.19	-	-	-	-	-
75	explicable	10	1.65	2.20	2.65	.0037	adjective	1.58	12.25	-	-	-	-	-
76	dashboard	9	3.06	2.33	2.65	.0038	noun	4.61	9.21	5.25	3.15	5.32	651.98	1.00
77	concessionary	13	2.78	1.37	3.25	.0064	adjective	2.15	14.43	-	-	-	-	-
78	retort	6	2.40	2.03	1.80	.0103	noun	2.75	11.50	-	-	-	628.15	.87
79	extent	6	4.40	2.34	1.70	.0063	noun	1.44	10.72	5.57	3.68	5.00	573.03	.97
80	mutual	6	3.72	2.14	1.85	.0038	adjective	2.21	8.90	6.48	3.50	6.45	598.86	.95
81	problematic	11	3.43	2.32	3.15	.0050	adjective	2.11	11.63	2.58	4.80	4.65	-	-
82	shiftless	9	1.30	1.62	2.40	.0047	adjective	2.27	12.12	-	-	-	693.12	.70
83	pleasantness	12	1.47	1.59	3.55	.0072	noun	2.00	8.44	-	-	-	-	-
84	nonpayment	10	1.17	1.71	3.60	.0064	noun	2.83	10.00	-	-	-	-	-
85	context	7	4.28	2.24	1.85	.0068	noun	2.17	10.00	5.00	3.18	5.60	597.95	.98

	Word	Length	Zipf	PREV	OLD20	BG	PoS	CNC	AoA	VAL	AROU	DOM	LDT RT	LDT Acc
86	shifting	8	3.73	2.34	1.65	.0088	verb	2.86	-	-	-	-	605.50	1.00
87	creamer	7	2.65	1.92	1.45	.0101	noun	4.66	8.72	5.47	2.81	6.09	738.38	.88
88	felicity	8	3.44	1.49	2.10	.0052	name	1.56	-	-	-	-	-	-
89	deferred	8	2.97	2.05	1.75	.0080	verb	2.00	-	-	-	-	666.58	.95
90	gyroscope	9	2.19	1.67	2.75	.0028	noun	4.25	12.69	-	-	-	-	-
91	recalculate	11	1.81	2.15	2.95	.0064	verb	2.93	11.53	-	-	-	-	-
92	frosty	6	3.51	2.35	1.80	.0046	adjective	3.90	6.33	6.15	4.61	5.00	607.38	.98
93	cohesiveness	12	1.60	1.85	3.85	.0088	noun	2.62	-	-	-	-	-	-
94	meld	4	2.19	1.34	1.00	.0060	verb	2.86	11.63	-	-	-	601.62	.34
95	awfulness	9	2.37	1.67	2.80	.0031	noun	2.20	9.67	-	-	-	-	-
96	rolled	6	4.16	2.25	1.45	.0069	verb	3.64	-	-	-	-	546.38	.97
97	orange	6	4.64	2.26	1.40	.0101	noun	4.66	3.26	6.81	4.04	5.58	519.53	.98
98	easily	6	4.69	2.43	1.75	.0061	adverb	1.80	-	-	-	-	-	-
99	reestablish	11	1.70	1.67	3.40	.0077	verb	2.54	10.33	6.14	4.00	6.18	-	-
100	lacquer	7	3.06	1.56	1.85	.0050	noun	4.28	13.19	4.95	3.30	5.00	699.11	.75

C.4 Power Analysis Random Effects Correlations

The power analysis in section 5.3 assumed that random effects correlations were all equal to zero. To examine whether this assumption impacted estimates of the statistical power associated with different sample sizes, I examined re-ran the simulations with random effects correlations all set to 0, .2, .4, .6, and .8. The results (Figure C.11) indicated that similar patterns in the sample size-power relationship should be expected across different random effect correlations.

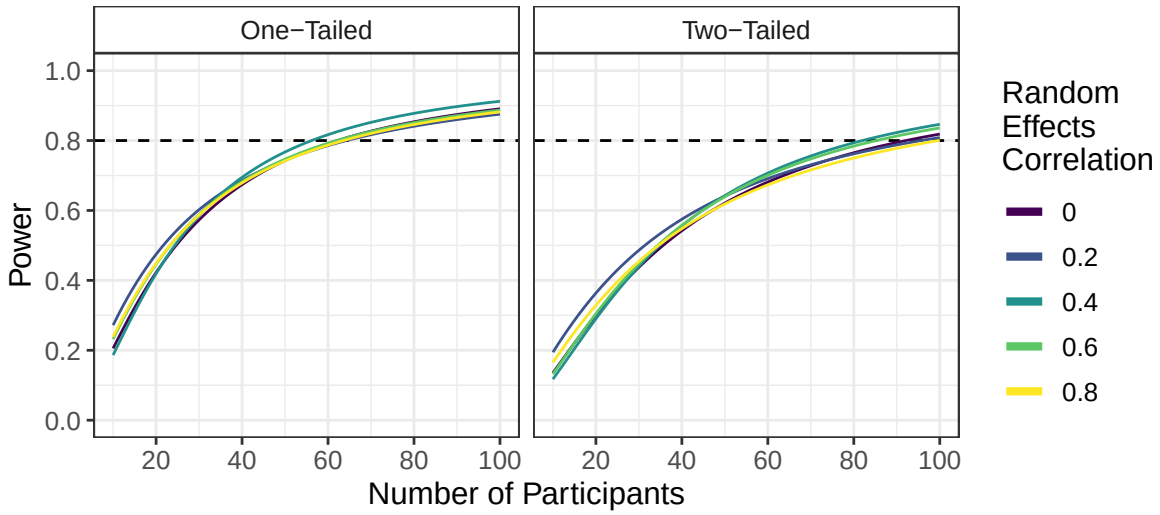


Figure C.11: Power curves when all random effect correlations are set to 0, .2, .4, .6, and .8. Each line depicts the predicted relationship between number of participants and power from a single loglinear binomial GLM. As in the original power analysis, results were simulated with N of 10 to 100 in steps of 5, though here with only 100 simulations at each step rather than 500. The overall relationship between the number of participants and the statistical power for finding the predicted interaction remains mostly unchanged across different random effects correlations. As in Figure 7, both one-tailed and two-tailed power are presented, though the *p* value used in the experiment is one-tailed. The dashed horizontal line highlights the 80% power target.

C.5 Task Instructions for the Localiser and Picture-Word Tasks

Instructions for the localiser and picture-word tasks, shown below, were presented multiple times: at the start of each task, after practice trials, and before the start of each block. The words AFFIRMATIVE and NEGATIVE below were replaced with the text "Left Control" or "Right Control" respectively, depending on which response group the participant was assigned to. In the practice trials, an additional line of text read, "For the practice trials, you will be given feedback on your accuracy for each trial.". For all other trials, this line instead read, "Unlike the practice trials, you will not be given feedback on your accuracy for each trial.".

The instructions for the localiser task were as follows:

In each trial, the following things will happen:
1) You will be shown a picture of a word, nonword, or noise image.
2) The image will turn green.
3) When the image turns green:
Press the AFFIRMATIVE key if the image is of a real word.
OR
Press the NEGATIVE key if it is not of a real word.

Once the image changes colour, try to respond as quickly and accurately as possible.

When you have read these instructions, press the space key to begin...

The instructions for the picture-word task were as follows:

In each trial, the following things will happen:
1) You will be shown a picture of an object for 2 seconds.
2) There will be a short delay.
3) You will be shown a word.
4) The word will turn green.
5) When the word turns green:
Press the AFFIRMATIVE key if the word describes the object you saw.
OR
Press the NEGATIVE key if it does not.

Once the word changes colour, try to respond as quickly and accurately as possible.

When you have read these instructions, press the space key to begin...

C.6 Picture-Word Planned Analysis Using the Word-Noise Maximal Electrode

In the planned analysis for the picture-word task, maximal electrodes were identified as those that showed the greatest sensitivity to the word vs. false font difference in the localiser task. For comparison, I also examined the pattern of results for electrodes that showed maximal sensitivity in the localiser task to the difference between words and phase-shuffled words. The results (Figure C.12) revealed a similar pattern of results. The model intercept, reflecting the average N1 amplitude at the lowest level of predictability, was estimated to be β=-4.432 μV (*SE*=.54), and the effect of predictability across both congruency conditions was estimated to be β=.62 μV (*SE*=.3). The predictability-congruency interaction (β=-1.19, *SE*=.5), where the effect of predictability, leading to less negative-going N1 amplitudes as predictability increases, was larger for picture-incongruent words than it was for picture-congruent words. A likelihood-ratio Chi-square model comparison revealed that this effect, although in the opposite direction to that hypothesised in the power analysis, had a two-tailed *p* value of .018 ($\chi^2(1)$=5.6). In decomposing the interaction, I report two-tailed *p* values from likelihood ratio Chi-square model comparisons, and report Bonferroni-corrected *p* values as p_{bonf}. For picture-incongruent words, the difference between the most and least predictable items was estimated to be β=1.2 μV (*SE*=.38, $\chi^2(1)$=9.04, *p*=.004, p_{bonf}=.008), while this difference for picture-congruent words was estimated to be β=-.01 (*SE*=.4, $\chi^2(1)$=.001, *p*=.971, p_{bonf}>.999).

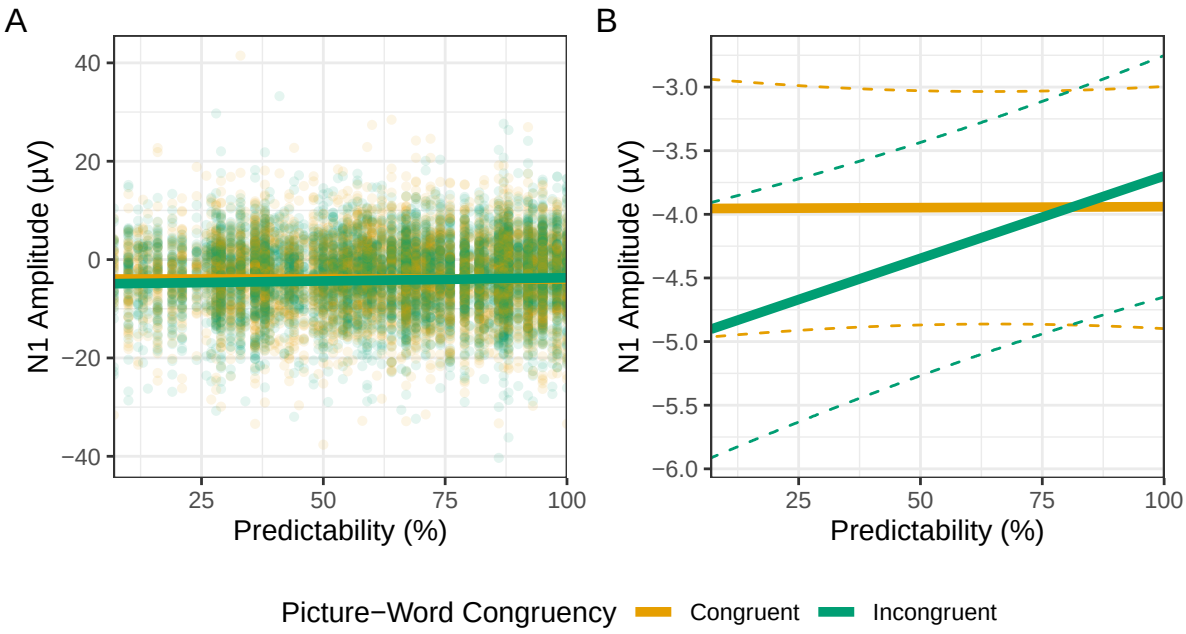


Figure C.12: Fixed effect predictions from an analysis of the picture-word task using electrodes that in the localiser task show maximal sensitivity to the difference between words and phase-shuffled words. (A) Model-derived fixed-effect predictions, visualised over results from all trials (individual points). (B) Fixed-effect predictions visualised alone for visibility, where dashed lines depict the bounds of 95% bootstrapped prediction intervals (estimated from 5,000 iterations), where bootstrapped predictions were generated using the *bootMer()* function of *lme4*. For feasibility, bootstrapped predictions were generated from a version of the model that lacked random slopes.

C.7 Details on the Shifted Log-Normal Bayesian Model Analysis of the EEG Experiment Picture-Word RT Data

A shifted log-normal Bayesian model was fit to the RT data from the picture-word task in the EEG experiment with brms (Bürkner, 2018). No trials were here excluded, such that there were 13,374 observations in total. The model was specified to be comparable to the stimulus validation analysis (Appendix C.2), estimating the plausibility of values for each parameter of the shifted log-normal distribution as a function of the maximal hierarchical structure justified by the experiment's design. The parameter of μ was modelled with an identity link function, while σ and δ were modelled with log link functions. The same predictors and random effects structure were used for each parameter as described for the EEG analysis and stimulus validation. The full formula, in *brms* syntax, was specified as:

```
rt ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id),
sigma ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id),
ndt ~ 1 + congruency * predictability +
        (1 + congruency * predictability | subject_id) +
        (1 + congruency | image_id) +
        (1 | word_id)
```

Prior distributions for both fixed and random effects were based on the posterior distributions of the stimulus validation analysis. However, prior distributions were not simply specified to be match posterior distributions exactly, as I anticipated that the 500 ms preview prior to response (see subsection 5.4.2), that was absent in the stimulus validation experiment, may vastly reduce the effect. Instead, I specified priors with more uncertainty than that observed in the stimulus validation posteriors, to reflect my expectation that results may change, although I did not know to what extent. Specifically, fixed and random effect prior distributions for the μ and σ parameters, and random effect priors for δ, were specified such that they were centred on the median estimate from the stimulus validation analysis, but with variance of the random effects multiplied to be ten times that observed in the stimulus validation posterior distributions. The fixed effect prior distributions for the δ parameter were specified to be more uninformative than this, as I expected this parameter to change the most. The prior distribution for the δ intercept was drawn from $\sim N(0, 7.5)$; while the fixed effect slopes' priors also had *SD*s of 7.5, but were centred on the posterior estimates from the stimulus validation analysis. Priors for all correlations of effects were kept as the *brms* default of a flat distribution between -1 and 1. The model was fit with 5 Markov chains, each with 10,000 (7,500 warm-up and 2,500 sampling)

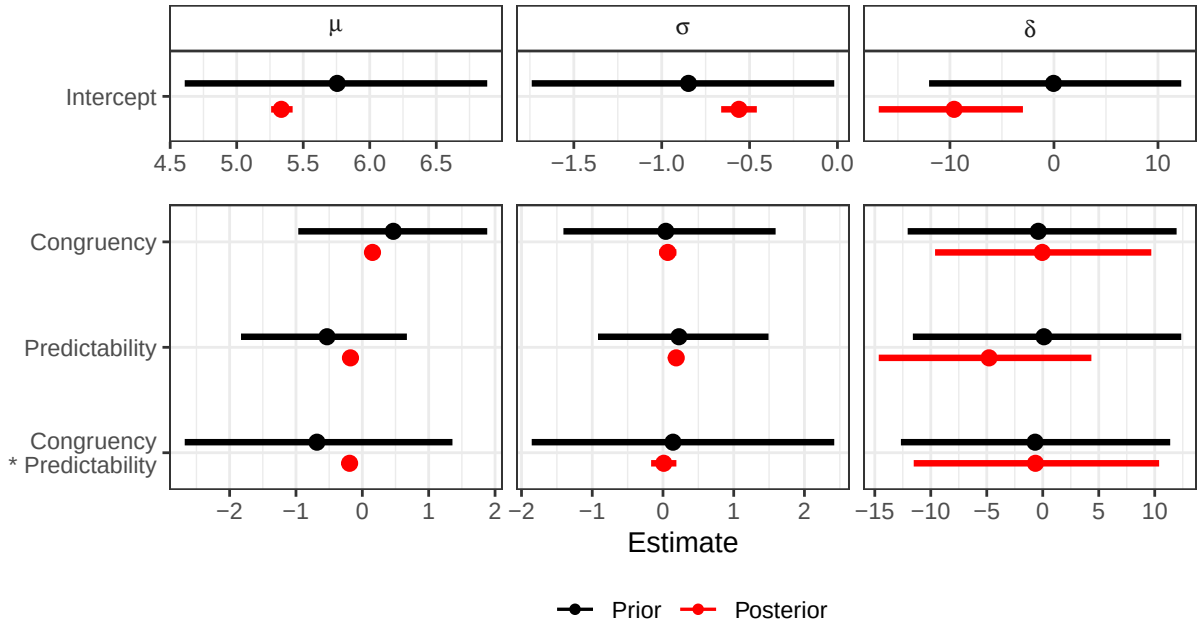


Figure C.13: Prior and posterior distributions for all fixed effects estimated in the Bayesian shifted log-normal model fit to describe RT data from the EEG experiment's picture-word task, for the μ, σ, and δ parameters. Points depict median estimates, while whiskers depict the extent of 89% HDIs, for prior (*black*) and posterior (*red*) distributions.

iterations. The *adapt_delta* parameter was set to .99, and the *max_treedepth* parameter was set to 10. Summaries of the fixed effect posterior distributions, relative to those of the priors, are shown in Figure C.13. Similar results are shown for all random effects in Figure C.14.

The impact of these estimates on RT distributions are described in the chapter's main text (section 5.5.2). One notable finding from the model's posterior distributions is that there is a high degree of uncertainty in the fixed effect posteriors for the δ parameter (Figure C.13). This is likely due to the large reduction in non-decision time observed in the EEG experiment but not in the stimuli validation experiment. Indeed, the posterior median for the δ intercept was equal to -10.07, which, since δ was modelled on a log scale, is equivalent to a non-decision time of 4.23e-5 ms (i.e., $e^{-10.07}$), indicating that shift in the RT distribution was so close to zero that the model struggled to describe it. As previously mentioned, this is likely due to participants being provided with 500 ms of preview for the stimulus, and only having to respond to a low-level change in stimulus colour that had consistent and predictable timing.

Posterior distributions for the *SD*s of random effects were more different from the specified priors than expected. One notable finding observed was that the *SD*s of random slopes for μ were estimated to be smaller in the EEG experiment than for the stimulus validation experiment, while *SD*s of random slopes for σ were estimated to be larger. For random intercepts, meanwhile, the opposite pattern was observed.

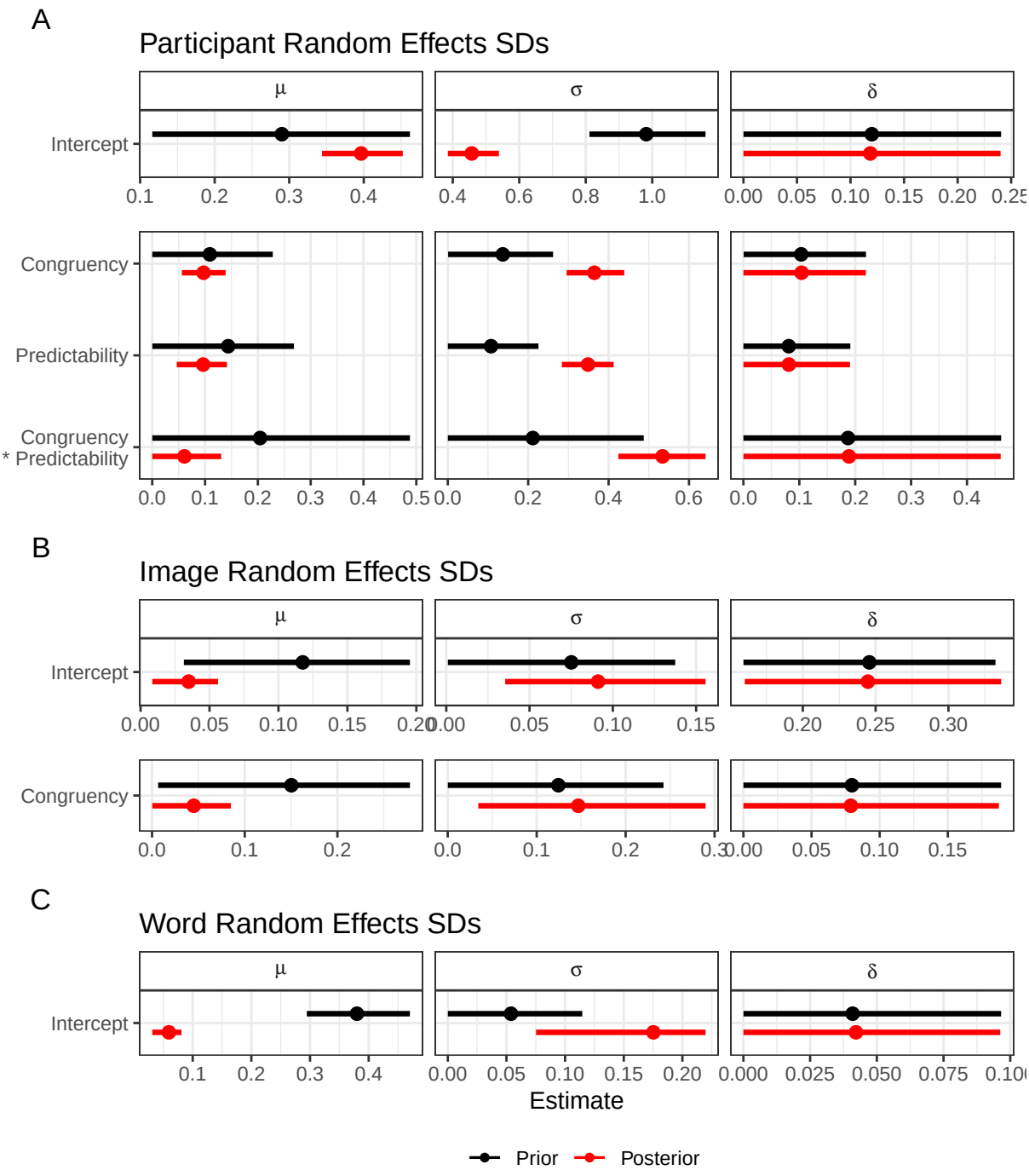


Figure C.14: Prior and posterior distributions for all random effects estimated in the Bayesian shifted log-normal model fit to describe RT data from the EEG experiment's picture-word task, for the μ, σ, and δ parameters. Results are separately for (A) participant, (B) image, and (C) word random effects. Points depict median estimates, while whiskers depict the extent of 89% HDIs, for prior (*black*) and posterior (*red*) distributions.

C.8 Sample-Level Analysis of Right-Hemispheric Occipitotemporal Effects in the Picture-Word Task

In addition to analysing sample-level effects on the left-lateralised occipitotemporal electrodes, I analysed data from a right-hemispheric occipitotemporal cluster (Figure C.15). Linear mixed effects models were fit to sample-level data (256 Hz), using the same model formula as that estimated for the left-hemispheric electrodes (section 5.5.2). Results revealed no clear predictability-congruency interaction before around 300 ms (Figure C.16).

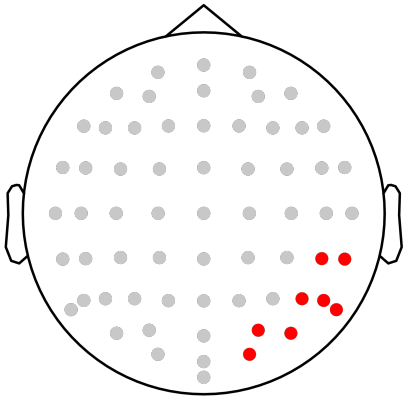


Figure C.15: Locations of right-hemispheric occipitotemporal electrodes. The selected electrodes were simply the right hemisphere homologues of the left-hemispheric occipitotemporal locations analysed elsewhere in Chapter 5.

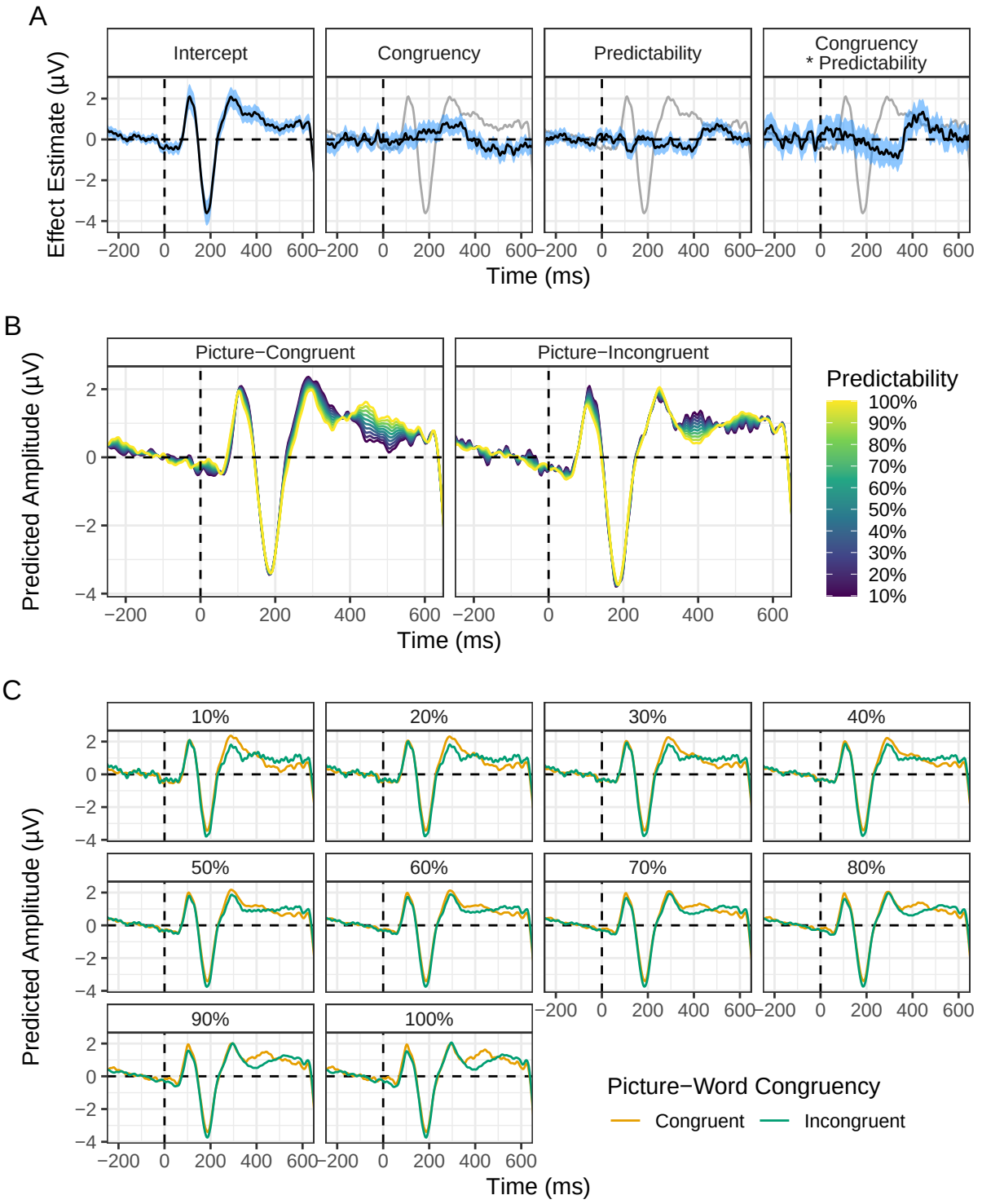


Figure C.16: Time-course of fixed effects from the sample-level analysis of the right-lateralised occipitotemporal region of interest. (A) Time-course of fixed effects estimates, with shaded regions depicting 95% confidence intervals. The model intercept (reflecting average amplitudes at the lowest level of predictability) is depicted as a grey line on each panel to provide a reference for the timing and magnitude of effects. (B) Fixed-effect predictions for picture-congruent and -incongruent words at levels of predictability from 10 to 100%, in steps of 10%. (C) Same data as (B), but split by predictability rather than congruency.

C.9 Sample-Level Analysis of Congruency * Predictability * Frequency Interaction

Much previous research on effects of predictions on early ERP components also examined whether there was an interaction with word frequency (e.g., Dambacher et al., 2012; Kretzschmar et al., 2015; Penolazzi et al., 2007; Sereno et al., 2003; Sereno et al., 2019), often arguing that, assuming effects of frequency index lexical access, an interaction between predictability and word frequency would provide evidence for top-down modulation of lexical access. For comparison to these results, I examined whether the interaction between congruency and predictability was frequency-dependent.

In the picture-word task stimuli, word frequency spanned a broad distribution, from 2.04 to 5.12 Zipf (from .11 to 132.5 occurrences per million), such that any clear interaction with word frequency could be expected to emerge for the given stimuli, assuming this interaction is linear. I fit sample-level linear mixed effects models to picture-word data, using the same model formula as that described for Section 5.5.2, but additionally estimating the effect of frequency, and interactions between frequency, congruency, and predictability. Here, frequency was parameterised as mean-centred Zipf, normalised such that there was a distance of 1 between the maximum and minimum, for comparability with the effect of predictability. Results (Figure C.17) revealed that while there may be a main effect of frequency, with more negative-going N1s observed as predictability increases, no clear interaction with congruency or predictability was found, although the broad confidence intervals and high variability in the baseline period for these interactions suggest a high degree of error in the estimates, likely due to a lack of statistical power.

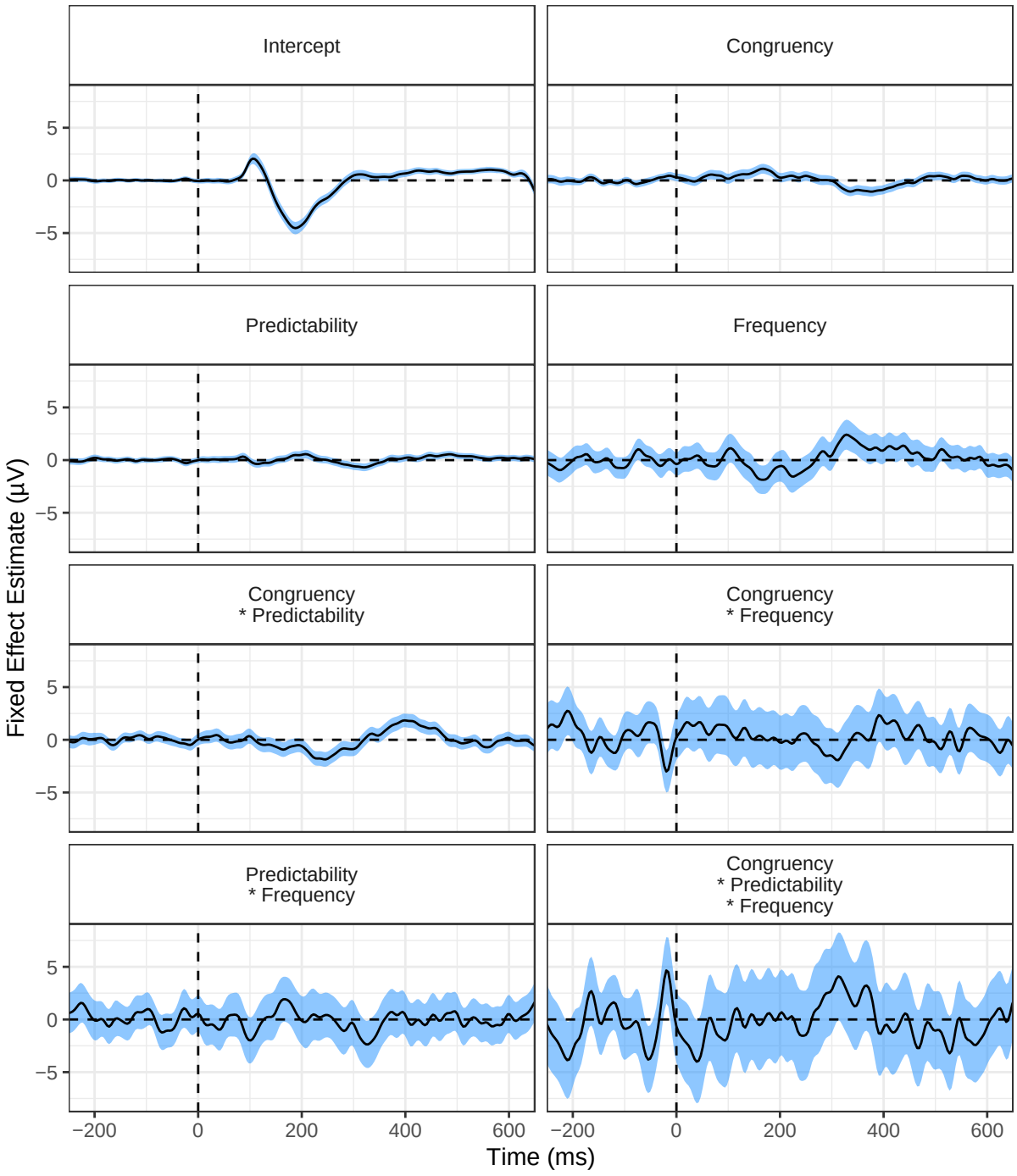


Figure C.17: Time-course of fixed effects from the sample-level analysis of the left-lateralised occipitotemporal region of interest, including interactions with word frequency. Shaded regions depict 95% confidence intervals for fixed effect estimates.

C.10 Details on the Behavioural Analysis of the Localiser Task

A logit-link binomial model of accuracy data was fit to data from the localiser task (Figure C.18A) with an informative prior for the model's logit intercept of $\sim N(5,1)$ (centred on average accuracy of .993), reflecting the expectation that accuracy overall would be very high. Weakly informative priors were defined for fixed effect slopes ($\sim N(0,5)$) and for the *SD*s of random effect distributions ($\sim t(5,0,1)$). Prior distributions for correlations within the model were specified to be flat. The model was fit with 5 chains, each with 10,000 iterations (7,500 warmup, 2,500 sampling). The *adapt_delta* parameter was set to .99, and the *max_tree_depth* was set to 10. In *brms* syntax, the model estimated coefficients from the following formula:

```
correct ~ 1 + false_font + noise +
  (1 + false_font + noise | participant_id) +
  (1 + false_font + noise | match_set) +
  (1 | item_id)
```

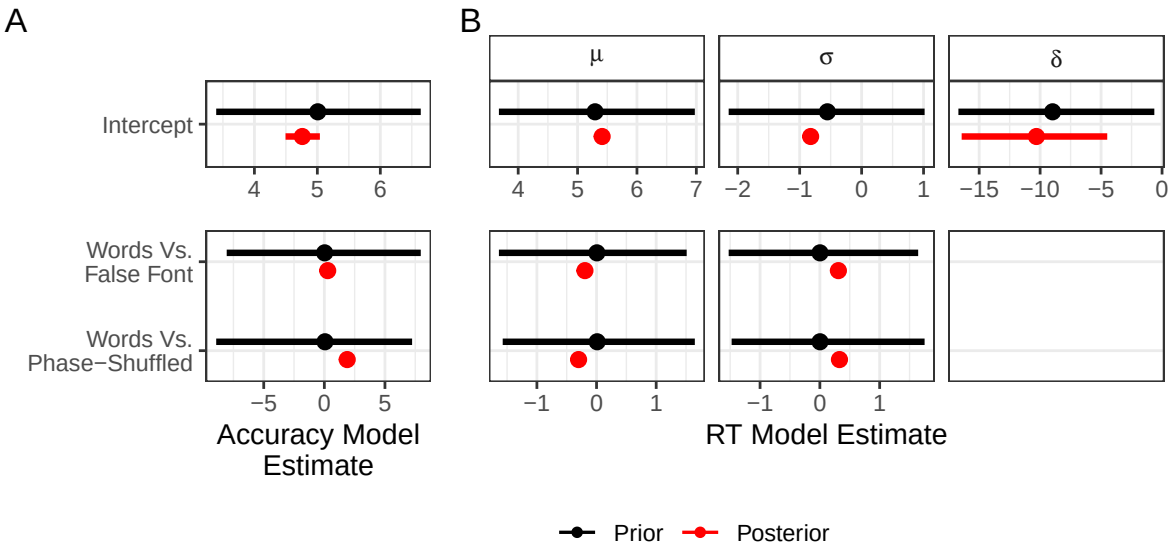


Figure C.18: Prior and posterior distributions for all fixed effects estimated by the (A) logit-link Binomial model to describe accuracies and (B) Bayesian shifted log-normal model fit to describe RT data from the localiser task. Estimates in (A) are in logit units. Estimates in (B) are depicted for each shifted log-normal parameter separately. In both panels, points depict median estimates, while whiskers depict the extent of 89% HDIs, for prior (*black*) and posterior (*red*) distributions.

RT data were modelled with a shifted log-normal model (Figure C.18B). The parameter of μ was modelled with an identity link function, while σ and δ were modelled with log link functions. The maximal random effects structure was estimated for the distributional parameters μ and σ, whereas the δ parameter was modelled with a global intercept only. This decision was

taken based on persistent divergent transitions in the Hamiltonian Monte Carlo sampler used to explore the model's parameter space. These divergences were caused by the extremely low shift (non-decision time) in the RT data from the EEG experiment, which approached 0 ($-\infty$ on a log scale). This problem is also the likely cause of the high uncertainty for the δ intercept, and also for the parameter coefficients in the analysis of the RT data from the picture-word task (Appendix C.7). Priors for fixed-effect intercepts were specified to be centred on posterior averages from the picture-word study RT analysis, though with additional uncertainty specified in the distributions to reflect the expectation that RT distributions would differ somewhat from the picture-word task. Specifically, the intercept for μ was specified as $\sim N(5.3,1)$, σ as $\sim N(-.56,1)$, and δ as $\sim N(-9,5)$. Priors for fixed effect slopes were specified as $\sim N(0,1)$. Prior distributions for the *SD*s of random effects were drawn from Student's *t* distributions centred on 0, with 5 degrees of freedom and a σ parameter of 1. Prior distributions for all correlations were flat. As with the model of accuracies, the RT model was fit with 5 chains, each with 10,000 iterations (7,500 warmup, 2,500 sampling). The *adapt_delta* parameter was set to .9, and the *max_tree_depth* was set to 10. In *brms* syntax, the model estimated coefficients from the following formula:

```
bf(
  rt ~ 1 + false_font + noise +
    (1 + false_font + noise | participant_id) +
    (1 + false_font + noise | match_set) +
    (1 | item_id),
  sigma ~ 1 + false_font + noise +
    (1 + false_font + noise | participant_id) +
    (1 + false_font + noise | match_set) +
    (1 | item_id),
  ndt ~ 1
)
```

D Chapter 6 Appendices

D.1 Separating Analysis of Simpson et al. (2013) Results by Letter Case

Rumelhart-Siple characters (Rumelhart & Siple, 1974) have only one case, resembling typical upper-case characters more than they resemble lower-case characters. To examine whether the quality of models predicting character similarity ratings (collected by Simpson et al., 2013), using the bit-wise approach outlined Rumelhart-Siple characters, differs between lower- and upper-case characters, I reran the analysis described in Section 6.2.1 for lower- and upper-case character pairs separately (Figure D.19. For comparison, I also fit separate versions of all other model variants to upper- and lower-case pairs.

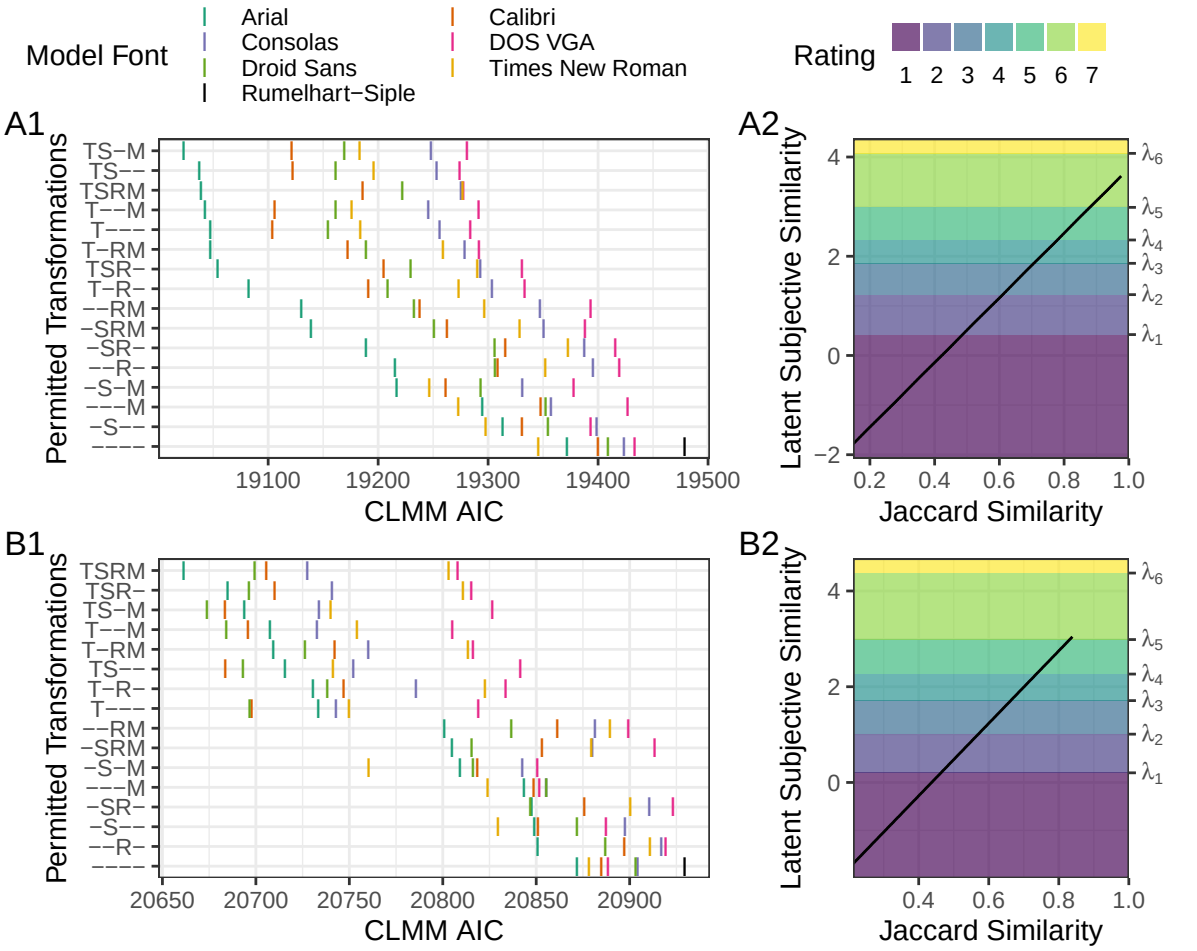


Figure D.19: Results from the analysis of the relationship between calculated Jaccard similarity and subjective ratings of character similarity collected by Simpson et al. (2013), where upper- and lower-case character pairs were analysed separately. (A1) AIC values for all models using Jaccard similarities to predict *lower-case* similarity ratings. (A2) Estimated relationship between Jaccard similarity and character similarity ratings for the *lower-case* model derived from Arial font, permitting all geometric transformations. (B1) AIC values for all models using Jaccard similarities to predict *upper-case* similarity ratings. (B2) Estimated relationship between Jaccard similarity and character similarity ratings for the best performing *upper-case*model derived from Arial font, permitting all geometric transformations. In all panels, layout and styling matches that used in Section 6.3.1. As the models were fit to separate data, AICs are only comparable within, and not between, panels A1 and B1.

Results revealed that, as expected, all pixel-based measures outperformed the bit-wise approach implemented for Rumelhart-Siple characters, for both upper- and lower-case character pairs. Two unexpected findings were also observed. First, the superiority of Arial-derived Jaccard similarity values was observed less clearly for upper-case characters than it was for lower-case characters, with models using Calibri- and Droid Sans-derived Jaccard similarities outperforming Arial in some cases (although the best performing model overall was that fit using Arial-derived values; Figure D.19A1). Second, for lower-case characters, the optimal model was not that permitting all geometric transformations, but rather that permitting all transformations except rotation (Figure D.19B1). However, it is also notable that the seven best models for lower-case character pairs, all fit using Arial-derived Jaccard similarities, performed very similarly. A key difference between upper- and lower-case characters that may contribute to this finding is that both Jaccard similarities and ratings for upper-case characters show less spread in their distributions for upper-case characters.

Finally, the effect of Jaccard similarity on character similarity ratings was estimated to be very similar across lower- (Figure D.19A2) and upper-case characters (Figure D.19B2).

D.2 Predicting BLP Behaviour from OLD20 and SCOLD20

In addition to examining model performance of models predicting lexical decision RTs and accuracies from the English Lexicon Project (ELP; Balota et al., 2007) as a function of OLD20 and SCOLD20, I also examined how well these metrics predict lexical decision behaviour for words in the *British* Lexicon Project (BLP; Keuleers et al., 2012). Trials were excluded from both the RT and accuracy analyses the BLP if responded to faster than 3000 ms (a lower upper bound than for ELP data, due to overall faster responses in the BLP). Trials were additionally excluded from the accuracy analysis if responded to incorrectly. Finally, trials were only included if the presented word was a member of the pool of ELP words from which OLD20 and SCOLD20 metrics were calculated. Following these exclusions, there were 722,039 trials in the accuracy analysis, and in the RT analysis, 614,915.

Models were fit exactly as described for the ELP analysis (see section 6.4.3), with logit-link GLMMs fit to describe accuracies, and linear mixed effects models predicting inverse RTs to summarise changes in RT distributions. Results (Figure D.20) showed a similar pattern of results as that observed for the ELP, with the OLD20 model outperforming all SCOLD20 models, for both accuracy and RT data.

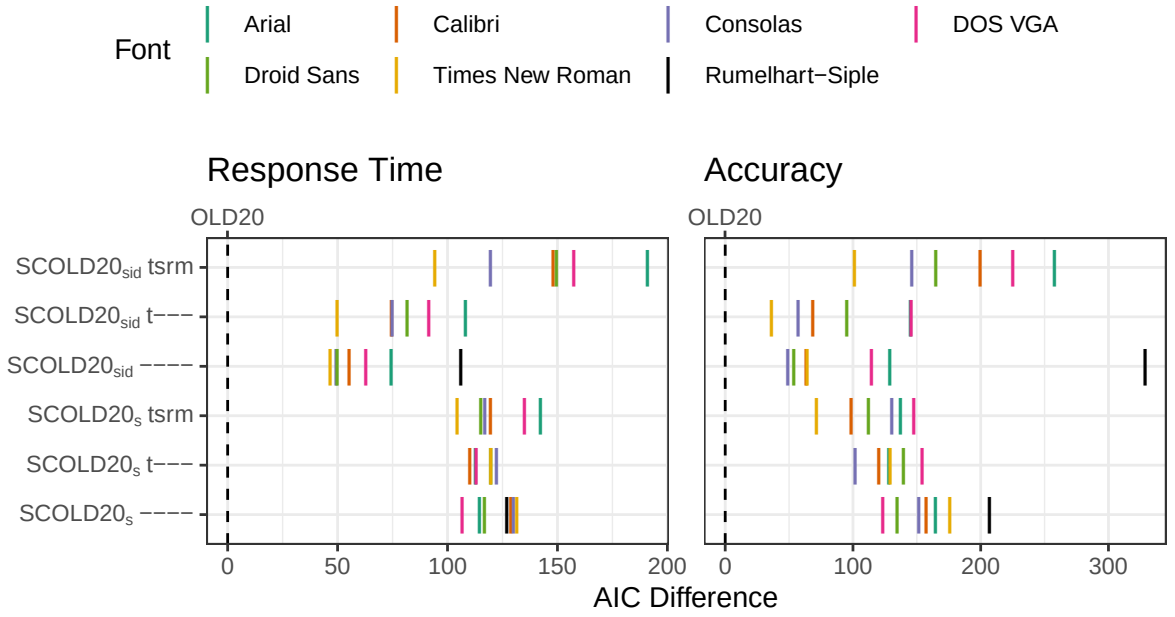


Figure D.20: AIC differences between models predicting BLP lexical decision behaviour using OLD20, and using all calculated SCOLD20 variants. AIC difference values reflect the difference between AICs from each SCOLD20 model and the OLD20 model, where a positive AIC difference indicates superior performance in the OLD20 model.

D.3 Effect of OLD20 on ERP data from Chapter 4

The effect of OLD20 in data from chapter 4 was estimated to compare the relative predictive utilities of OLD20 and SCOLD20 metrics over time. However, before examining the relative predictive abilities of models fit to ERPs with the SCOLD20 metrics, I first examined the pattern of observed effects. Following previous findings of interactions between orthographic similarity and lexical status (Baeck et al., 2015), and anticipating that task effects may interact with orthographic variables like OLD20, I estimated the original model formula from Chapter 4, as well as both the main effect of OLD20, and all possible interactions between OLD20 and other variables in the fixed effects structure. Because of the large number of models that needed to be fit, no random slopes were estimated, the only random effects estimated were per-participant, per-item, and per-match-set random intercepts. The specific model formula was specified, in *lme4* (Bates et al., 2015) syntax, as:

```
amplitude ~ 1 + (category_relevant + pseudoword + nonword) * task * measure +
  (1 | participant) +
  (1 | match_set) +
  (1 | item)
```

Here, *measure* reflected the OLD20 (or in later models, SCOLD20) values, scaled by standard deviation, while all variables except *measure* match those described in Chapter 4. Models were estimated for three distinct clusters of electrodes, in left occipitotemporal, centroparietal, and right occipitotemporal regions (Figure D.21A).

Notable model findings include effects of OLD20 in the left and right hemispheric N1 components elicited by words (Figure D.21B). Smaller OLD20 values (larger orthographic neighbourhoods) elicited less negative-going left hemispheric N1 components (Figure D.22A). Over the right hemisphere, however, although effects were less pronounced, model estimates suggest that smaller OLD20 values elicited less negative amplitudes at the component's peak, smaller OLD20 values also more negative amplitudes during the component's onset (Figure D.22C). The models also revealed possible interactions between OLD20, stimulus, and task - especially in the ERP observed centroparietally. However, considering the post-hoc nature of this analysis, and the high standard errors observed for interaction estimates (reflected in their broad confidence intervals; Figure D.21E), care should be taken not to over-interpret these exploratory results.

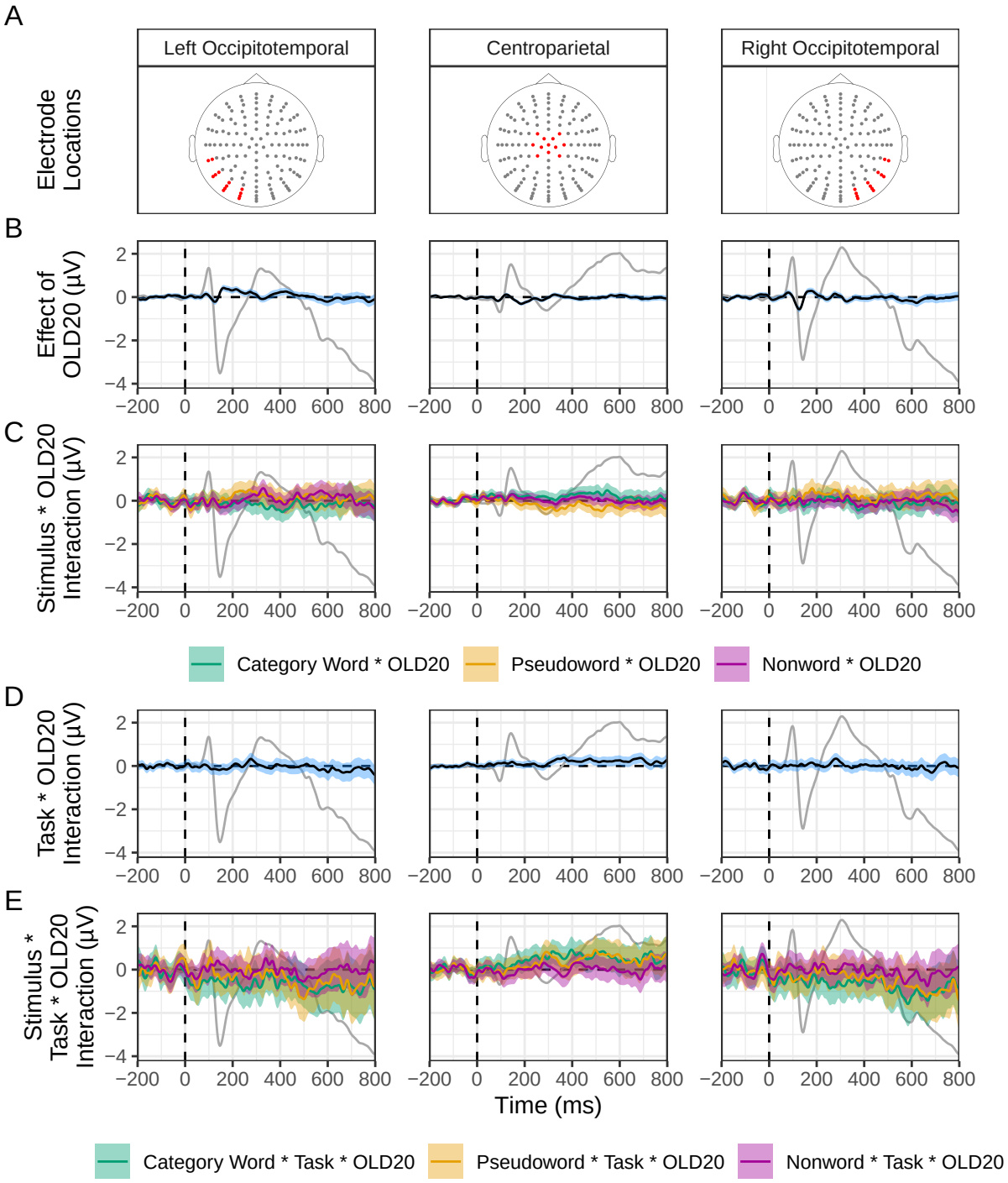


Figure D.21: Model estimates from the exploratory OLD20 analysis of data from Chapter 4. (A) The locations of electrodes included in each region of interest. Subsequent panels match the order of regions in (A), depicting the following model estimates (lines) and 95% confidence intervals (shaded regions): (B) the effect of OLD20, (C) interactions between stimulus and OLD20, (D) the interaction between task and OLD20, and (E) three-way interactions between task, stimulus, and OLD20. In all panels from (B) to (E), the grey line depicts the model intercept (i.e., predicted amplitude of all category-irrelevant words across both tasks), to provide reference for the timing and size of effects.

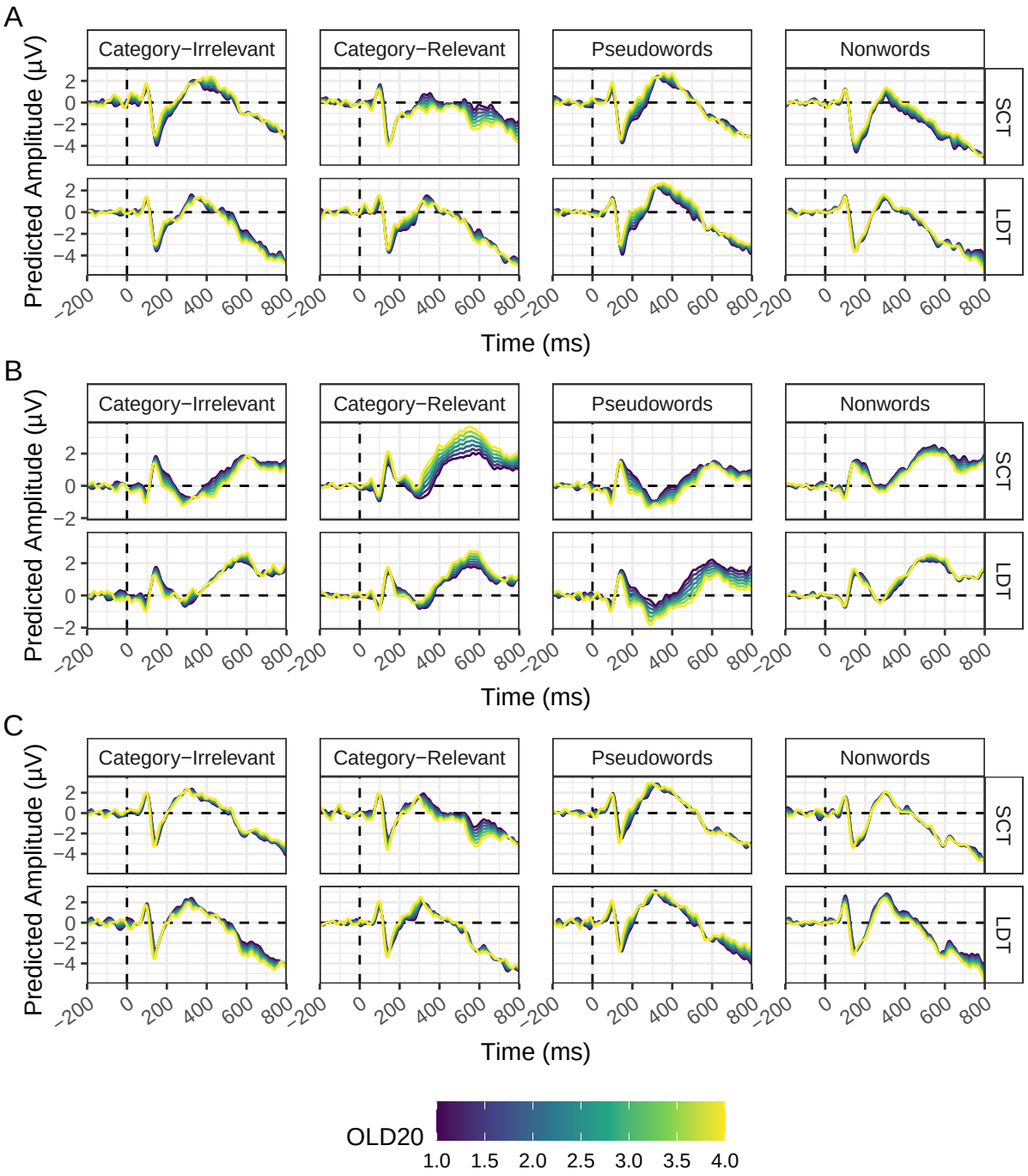


Figure D.22: Fixed effect predictions from the exploratory OLD20 analysis of data from Chapter 4. Results depict the predicted ERPs for all factorial cells in Chapter 4, where the colours of lines vary to show the effect of OLD20. Results are shown separately for (A) left occipitotemporal, (B) centroparietal, and (C) right occipitotemporal regions of interest.

References

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. *Brain and Language*, *122*(1), 42–54. https://doi.org/10.1016/j.bandl.2012.04.014

Adelman, J. S. (2011). Letters in Time and Retinotopic Space. *Psychological Review*, *118*(4), 570–582. https://doi.org/10.1037/a0024811

Allison, T., Mccarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex*, *4*(5), 544–554. https://doi.org/10.1093/cercor/4.5.544

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *4*(4), 439–461. https://doi.org/10.3758/BF03214334

Appelbaum, L. G., Liotti, M., Perez, R., Fox, S. P., & Woldorff, M. G. (2009). The temporal dynamics of implicit processing of non-letter, letter, and word-forms in the human visual cortex. *Frontiers in Human Neuroscience*, *3*, 1–11. https://doi.org/10.3389/neuro.09.056.2009

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and frequency: a group study using MEG. *NeuroReport*, *14*(8), 1183–1187. https://doi.org/10.1097/00001756-200306110-00016

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., & Malach, R. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *Journal of Neurophysiology*, *87*(6), 3102–3116. https://doi.org/10.1152/jn.2002.87.6.3102

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

Baeck, A., Kravitz, D., Baker, C., & Op de Beeck, H. P. (2015). Influence of lexical status and orthographic similarity on the multi-voxel response of the visual word form area. *NeuroImage*, *111*, 321–328. https://doi.org/10.1016/j.neuroimage.2015.01.060

Bai, J., Shi, J., Jiang, Y., He, S., & Weng, X. (2011). Chinese and Korean characters engage the same visual word form area in proficient early Chinese-Korean bilinguals. *PLoS ONE*, *6*(7), 1–8. https://doi.org/10.1371/journal.pone.0022765

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barlow, H. B. (1997). The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1141–1147. https://doi.org/10.1098/rstb.1997.0097

Barr, D. J., Levy, J., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(Supplement), 1–63. https://doi.org/10.1016/j.jml.2012.11.001.Random

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

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beech, J. R., & Mayall, K. A. (2005). The word shape hypothesis re-examined: Evidence for an external feature advantage in visual word recognition. *Journal of Research in Reading*, *28*(3), 302–319. https://doi.org/10.1111/j.1467-9817.2005.00271.x

Behrmann, M., & Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, *17*(5), 210–219. https://doi.org/10.1016/j.tics.2013.03.007

Belfi, A. M., & Kacirek, K. (2021). The famous melodies stimulus set. *Behavior Research Methods*, *53*, 34–48. https://doi.org/10.3758/s13428-020-01411-6

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260. https://doi.org/10.1162/089892999563373

Besner, D., & Smith, M. C. (1992). Models of visual word recognition: When obscuring the stimulus yields a clearer view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 468–482. https://doi.org/10.1037/0278-7393.18.3.468

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. https://doi.org/10.3758/s13428-012-0195-z

Bianchi, B., Bengolea Monzón, G., Ferrer, L., Fernández Slezak, D., Shalom, D. E., & Kamienkowski, J. E. (2020). Human and computer estimations of Predictability of words in written language. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-61353-z

Blau, V. C., Maurer, U., Tottenham, N., & McCandliss, B. D. (2007). The face-specific N170 component is modulated by emotional facial expression. *Behavioral and Brain Functions*, *3*, 1–13. https://doi.org/10.1186/1744-9081-3-7

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bouhali, F., de Schotten, M. T., Pinel, P., Poupon, C., Mangin, J. F., Dehaene, S., & Cohen, L. (2014). Anatomical connections of the visual word form area. *Journal of Neuroscience*, *34*(46), 15402–15414. https://doi.org/10.1523/JNEUROSCI.4918-13.2014

Brainerd, C. J., Chang, M., Bialer, D. M., & Toglia, M. P. (2021). Semantic ambiguity and memory. *Journal of Memory and Language*, *121*(March), 104286. https://doi.org/10.1016/j.jml.2021.104286

Brem, S., Bucher, K., Halder, P., Summers, P., Dietrich, T., Martin, E., & Brandeis, D. (2006). Evidence for developmental changes in the visual word processing network beyond adolescence. *NeuroImage*, *29*(3), 822–837. https://doi.org/10.1016/j.neuroimage.2005.09.023

Brem, S., Halder, P., Bucher, K., Summers, P., Martin, E., & Brandeis, D. (2009). Tuning of the visual word processing system: Distinct developmental ERP and fMRI effects. *Human Brain Mapping*, *30*(6), 1833–1844. https://doi.org/10.1002/hbm.20751

Brem, S., Hunkeler, E., Mächler, M., Kronschnabel, J., Karipidis, I. I., Pleisch, G., & Brandeis, D. (2018). Increasing expertise to a novel script modulates the visual N1 ERP in healthy adults. *International Journal of Behavioral Development*, *42*(3), 333–341. https://doi.org/10.1177/0165025417727871

Brennan, C., Cao, F., Pedroarena-Leal, N., Mcnorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems. *Human Brain Mapping*, *34*(12), 3354–3368. https://doi.org/10.1002/hbm.22147

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS ONE*, *9*(9), 1–10. https://doi.org/10.1371/journal.pone.0106953

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Brown, G. D., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, *15*(3), 208–216. https://doi.org/10.3758/BF03197718

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Languagea*, *109*, 1–30. https://doi.org/10.1016/j.jml.2019.104047

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–1523. https://doi.org/10.3758/s13428-016-0811-4

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

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., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1–20. https://doi.org/10.5334/joc.10

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1–11. https://doi.org/10.3389/fpsyg.2016.01116

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Büchel, C., Price, C., & Friston, K. (1998). A multimodal language region in the ventral visual pathway. *Nature*, *394*, 274–277. https://doi.org/10.1038/28389

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., & Rosen, B. R. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14878–14883. https://doi.org/10.1073/pnas.93.25.14878

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Bürkner, P.-C. (2020). Bayesian item response modeling in R with brms and Stan. *arXiv e-prints*, arXiv:1905.09501. https://doi.org/10.48550/arXiv.1905.09501Focustolearnmore

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in Psychology : A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(4), 857–871. https://doi.org/10.1037/0278-7393.23.4.857

Centanni, T. M., Norton, E. S., Park, A., Beach, S. D., Halverson, K., Ozernov-Palchik, O., Gaab, N., & Gabrieli, J. D. (2018). Early development of letter specialization in left fusiform is associated with better word reading and smaller fusiform face area. *Developmental Science*, *21*(5), 1–10. https://doi.org/10.1111/desc.12658

Chang, C. Y., Hsu, S. H., Pion-Tonachini, L., & Jung, T. P. (2020). Evaluation of Artifact Subspace Reconstruction for automatic artifact components removal in multi-channel EEG recordings. *IEEE Transactions on Biomedical Engineering*, *67*(4), 1114–1121. https://doi.org/10.1109/TBME.2019.2930186

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2018). shiny: Web application framework for R. https://cran.r-project.org/package=shiny

Chen, L., Wassermann, D., Abrams, D. A., Kochalka, J., Gallardo-Diez, G., & Menon, V. (2019). The visual word form area (VWFA) is part of both language and attention circuitry. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-13634-z

Chen, Y., Davis, M. H., Pulvermüller, F., & Hauk, O. (2013). Task modulation of brain responses in visual word recognition as studied using EEG/MEG and fMRI. *Frontiers in Human Neuroscience*, *7*, 1–14. https://doi.org/10.3389/fnhum.2013.00376

Chen, Y., Davis, M. H., Pulvermüller, F., & Hauk, O. (2015). Early visual word processing is flexible: Evidence from spatiotemporal brain dynamics. *Journal of Cognitive Neuroscience*, *27*(9), 1738–1751. https://doi.org/10.1162/jocn_a_00815

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007). Reading spoken words: Orthographic effects in auditory priming. *Cognition*, *102*(3), 341–360. https://doi.org/10.1016/j.cognition.2006.01.001

Christensen, R. H. B. (2020). ordinal: Regression Models for Ordinal Data
R package version 2020.8-22. https://CRAN.R-project.org/package=ordinal.

Chwilla, D. J., Brown, C. M., & Hagoort, P. (1995). The N400 as a function of the level of processing. *Psychophysiology*, *32*(3), 274–285. https://doi.org/10.1111/j.1469-8986.1995.tb02956.x

Clark, A. (2020). Pillow (PIL Fork) documentation. https://pillow.readthedocs.io/en/stable/reference/

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

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods*, *36*(3), 371–383. https://doi.org/10.3758/BF03195584

Cohen, L., & Dehaene, S. (2004). Specialization within the ventral stream: The case for the visual word form area. *NeuroImage*, *22*(1), 466–476. https://doi.org/10.1016/j.neuroimage.2003.12.049

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*, 291–307.

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain*, *125*(5), 1054–1069. https://doi.org/10.1093/brain/awf094

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*(4), 497–505. https://doi.org/10.1080/14640748108400805

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance vi* (pp. 535–555). Lawrence Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037//0033-295x.108.1.204

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

Culham, J. C., Cavina-Pratesi, C., & Singhal, A. (2006). The role of parietal cortex in visuomotor control: What have we learned from neuroimaging? *Neuropsychologia*, *44*(13), 2668–2684. https://doi.org/10.1016/j.neuropsychologia.2005.11.003

Cutler, A., Treiman, R., & van Ooijen, B. (2010). Strategic deployment of orthographic knowledge in phoneme detection. *Language and Speech*, *53*(3), 307–320. https://doi.org/10.1177/0023830910371445

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852–1870. https://doi.org/10.1016/j.neuropsychologia.2012.04.011

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*(3), 713–758. https://doi.org/10.1037/a0019738

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). Measuring the associative structure of English: The "Small World of Words" norms for word association. *Behavior Resarch Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

DeBruine, L. (2020). faux: Simulation for factorial designs. https://doi.org/10.5281/zenodo.2669586

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J. B., Bihan, D. L., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*(5), 307–313. https://doi.org/10.1111/j.0956-7976.2004.00674.x

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, *4*(7), 752–8. https://doi.org/10.1038/89551

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Dehaene, S., Cohen, L., Morais, J., & Kolinsky, R. (2015). Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, *16*(4), 234–244. https://doi.org/10.1038/nrn3924

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341. https://doi.org/10.1016/j.tics.2005.05.004

Dehaene, S., & Dehaene-Lambertz, G. (2016). Is the brain prewired for letters? *Nature Neuroscience*, *19*(9), 1192–1193. https://doi.org/10.1038/nn.4369

Dehaene, S., Le Clec'H, G., Poline, J. B., Le Bihan, D., & Cohen, L. (2002). The visual word form area: A prelexical representation of visual words in the fusiform gyrus. *NeuroReport*, *13*(3), 321–325. https://doi.org/10.1097/00001756-200203040-00015

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, *330*(6009), 1359–1364. https://doi.org/10.1126/science.1194140

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS Biology*, *16*(3), 1–34. https://doi.org/10.1371/journal.pbio.2004103

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634). https://doi.org/10.1098/rstb.2012.0394

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Devlin, J. T., Jamison, H. L., Gonnerman, L. M., & Matthews, P. M. (2006). The role of the posterior fusiform gyrus in reading. *Journal of Cognitive Neuroscience*, *18*(6), 911–922. https://doi.org/10.1162/jocn.2006.18.6.911

Dikker, S., & Pylkkanen, L. (2011). Before the N400: Effects of lexical-semantic violations in visual cortex. *Brain and Language*, *118*(1-2), 23–28. https://doi.org/10.1016/j.bandl.2011.02.006

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, *110*(3), 293–321. https://doi.org/10.1016/j.cognition.2008.09.008

Dou, H., Liang, L., Ma, J., Lu, J., Zhang, W., & Li, Y. (2021). Irrelevant task suppresses the N170 of automatic attention allocation to fearful faces. *Scientific Reports*, *11*(1), 1–10. https://doi.org/10.1038/s41598-021-91237-9

Dundas, E. M., Plaut, D. C., & Behrmann, M. (2014). An ERP investigation of the co-development of hemispheric lateralization of face and word recognition. *Neuropsychologia*, *61*(1), 315–323. https://doi.org/10.1016/j.neuropsychologia.2014.05.006

Dunn-Rankin, P., Leton, D. A., & Shelton, V. F. (1968). Congruency factors related to visual confusion of English letters. *Perceptual and Motor Skills*, *26*(2), 659–666. https://doi.org/10.2466/pms.1968.26.2.659

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

Dzigiel-Fivet, G., Plewko, J., Szczerbiński, M., Marchewka, A., Szwed, M., & Jednoróg, K. (2021). Neural network for Braille reading and the speech-reading convergence in the blind: Similarities and differences to visual reading. *NeuroImage*, *231*, 1–11. https://doi.org/10.1016/j.neuroimage.2021.117851

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2020). *{GNU Octave} version 6.1.0 manual: a high-level interactive language for numerical computations*. https://www.gnu.org/software/octave/doc/v6.1.0/

Eberhard-Moscicka, A. K., Jost, L. B., Fehlbaum, L. V., Pfenninger, S. E., & Maurer, U. (2016). Temporal dynamics of early visual word processing - Early versus late N1 sensitivity in children and adults. *Neuropsychologia*, *91*, 509–518. https://doi.org/10.1016/j.neuropsychologia.2016.09.014

Eisenhauer, S., Gagl, B., & Fiebach, C. J. (2022). Predictive pre-activation of orthographic and lexical-semantic representations facilitates visual word recognition. *Psychophysiology*, *59*(3), 1–26. https://doi.org/10.1111/psyp.13970

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*(2), 181–192.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, *50*(3), 1116–1124. https://doi.org/10.3758/s13428-017-0930-6

eSpeak version 1.48.15. (2015). http://espeak.sourceforge.net/

Fan, C., Chen, S., Zhang, L., Qi, Z., Jin, Y., Wang, Q., Luo, Y., Li, H., & Luo, W. (2015). N170 changes reflect competition between faces and identifiable characters during early visual processing. *NeuroImage*, *110*, 32–38. https://doi.org/10.1016/j.neuroimage.2015.01.047

Feng, X., Monzalvo, K., Dehaene, S., & Dehaene-lambertz, G. (2022). Evolution of reading and face circuits during the first three years of reading acquisition. *Neuroimage*, *in press*. https://doi.org/10.1016/j.neuroimage.2022.119394

Feredoes, E., Heinen, K., Weiskopf, N., Ruff, C., & Driver, J. (2011). Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. *PNAS*, *108*(42), 17510–17515. https://doi.org/10.1073/pnas.1106439108

Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., & Binder, J. R. (2016). Heteromodal cortical areas encode sensory-motor features of word meaning. *Journal of Neuroscience*, *36*(38), 9763–9769. https://doi.org/10.1523/JNEUROSCI.4095-15.2016

Ferrand, L., & Grainger, J. (1994). Effects of orthography are independent of phonology in masked form priming. *The Quarterly Journal of Experimental Psychology Section A*, *47*(2), 365–382. https://doi.org/10.1080/14640749408401116

Fischer, J., & Whitney, D. (2009). Attention narrows position tuning of population responses in V1. *Current Biology*, *19*(16), 1356–1361. https://doi.org/10.1016/j.cub.2009.06.059

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S., & Tamez, E. R. (2017). Decoding levels of representation in reading: A representational similarity approach. *Cortex*, *90*, 88–102. https://doi.org/10.1016/j.cortex.2017.02.017

Fodor, J. (1983). Input Systems as Modules. In *The modularity of mind* (pp. 47–101). MIT Press.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. Cooper & E. Walker (Eds.), *Sentence processing: Psycholinguistic essays presented to merrill garrett* (pp. 27–85). Erlbaum.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. https://doi.org/10.1016/S1364-6613(00)01839-8

Friederici, A. D., & Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax-semantic interface. *Brain Research*, *1146*(1), 50–58. https://doi.org/10.1016/j.brainres.2006.08.038

Frisson, S., Bélanger, N. N., & Rayner, K. (2014). Phonological and orthographic overlap effects in fast and masked priming. *Quarterly Journal of Experimental Psychology*, *67*(9), 1742–1767. https://doi.org/10.1080/17470218.2013.869614

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gagl, B., Richlan, F., Ludersdorfer, P., Sassenhagen, J., Eisenhauer, S., Gregorova, K., & Fiebach, C. J. (2022). The lexical categorization model: A computational model of left ventral occipito-temporal cortex activation in visual word recognition. *PLoS Computational Biology*, *18*(6), e1009995. https://doi.org/10.1371/journal.pcbi.1009995

Gagl, B., Sassenhagen, J., Haan, S., Gregorova, K., Richlan, F., & Fiebach, C. J. (2020). An orthographic prediction error as the basis for efficient visual word recognition. *NeuroImage*, *214*(August 2019), 116727. https://doi.org/10.1016/j.neuroimage.2020.116727

Gentner, D., & Asmuth, J. (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298–1307. https://doi.org/10.1080/23273798.2017.1410560

Gervais, M. J., Harvey, L. O., & Roberts, J. O. (1984). Identification confusions among letters of the alphabet. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 655–666. https://doi.org/10.1037/0096-1523.10.5.655

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. *Cerebral cortex (New York, N.Y. : 1991)*, *32*(7), 1405–1418. https://doi.org/10.1093/cercor/bhab287

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363. https://doi.org/10.1038/nrn3476

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L. C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in Psychology*, *7*, 1–10. https://doi.org/10.3389/fpsyg.2016.00976

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*(3), 1–52. https://doi.org/10.1037/a0012667

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face like stimuli by newborn infants. *Pediatrics*, *56*(4), 544–549. https://doi.org/10.1542/peds.56.4.544

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *14*(5), 505–509. https://doi.org/10.1111/1467-9280.03452

Grainger, J., & Jacobs, A. M. (1996). Orthographic Processing in Visual Word Recognition: A Multiple Read-Out Model. *Psychological Review*, *103*(3), 518–565. https://doi.org/10.1037/0033-295X.103.3.518

Grainger, J., & van Heuven, W. J. (2004). Modelling letter position coding in printed word perception. In P. Bonin (Ed.), *Mental lexicon: "some words to talk about words"* (pp. 1–23). Nova Science Publishers.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 1–13. https://doi.org/10.3389/fnins.2013.00267

Grefkes, C., & Fink, G. R. (2005). The functional organization of the intraparietal sulcus in humans and monkeys. *Journal of Anatomy*, *207*(1), 3–17. https://doi.org/10.1111/j.1469-7580.2005.00426.x

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, *27*, 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Grühn, D., & Scheibe, S. (2008). Age-related differences in valence and arousal ratings of pictures from the International Affective Picture System (LAPS): Do ratings become more extreme with age? *Behavior Research Methods*, *40*(2), 512–521. https://doi.org/10.3758/BRM.40.2.512

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Gutiérrez-Sigut, E., Marcet, A., & Perea, M. (2019). Tracking the time course of letter visual-similarity effects during word recognition: A masked priming ERP investigation. *Cognitive, Affective and Behavioral Neuroscience*, *19*(4), 966–984. https://doi.org/10.3758/s13415-019-00696-1

Gwilliams, L., Lewis, G. A., & Marantz, A. (2016). Functional characterisation of letter-specific responses in time, space and current polarity using magnetoencephalography. *NeuroImage*, *132*, 320–333. https://doi.org/10.1016/j.neuroimage.2016.02.057

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis. *Journal of Cognitive Neuroscience*, *11*(2), 194–205. https://doi.org/10.1162/089892999563328

Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, *13*(3), 339–356. https://doi.org/10.1016/S0926-6410(01)00127-6

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(46), 1–12. https://doi.org/10.1073/pnas.2104779118

Hannagan, T., & Grainger, J. (2013). The lazy visual word form area: Computational insights into location-sensitivity. *PLoS Computational Biology*, *9*(10), 1–12. https://doi.org/10.1371/journal.pcbi.1003250

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hauk, O., Coutout, C., Holden, A., & Chen, Y. (2012). The time-course of single-word reading: Evidence from fast behavioral and brain responses. *NeuroImage*, *60*(2), 1462–1477. https://doi.org/10.1016/j.neuroimage.2012.01.061

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103. https://doi.org/10.1016/j.clinph.2003.12.020

Hauk, O. (2016). Only time will tell - why temporal information is essential for our neuroscientific understanding of semantics. *Psychonomic Bulletin and Review*, *23*(4), 1072–1079. https://doi.org/10.3758/s13423-015-0873-9

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340–347. https://doi.org/10.1037//0033-2909.109.2.340

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & de Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13996-4

Helenius, P., Tarkiainen, A., Cornelissen, P., Hansen, P. C., & Salmelin, R. (1999). Dissociation of normal feature analysis and deficient processing of letter-strings in dyslexic adults. *Cerebral Cortex*, *9*(5), 476–483. https://doi.org/10.1093/cercor/9.5.476

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience*, *14*(6), 938–950. https://doi.org/10.1162/089892902760191153

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin and Review*, *23*(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, *50*(1), 115–133. https://doi.org/10.3758/s13428-017-1009-0

Hsieh, P. J., Vul, E., & Kanwisher, N. (2010). Recognition alters the spatial pattern of fMRI activation in early retinotopic cortex. *Journal of Neurophysiology*, *103*(3), 1501–1507. https://doi.org/10.1152/jn.00812.2009

Hsu, C. H., Lee, C. Y., & Marantz, A. (2011). Effects of visual complexity and sublexical information in the occipitotemporal cortex in the reading of Chinese phonograms: A single-trial analysis with MEG. *Brain and Language*, *117*(1), 1–11. https://doi.org/10.1016/j.bandl.2010.10.002

Huang, X., Wong, W. L., Tse, C.-Y., Sommer, W., Dimigen, O., & Maurer, U. (2022). Is there magnocellular facilitation of early neural processes underlying visual word recognition? Evidence from masked repetition priming with ERPs. *Neuropsychologia*, *170*(April), 108230. https://doi.org/10.1016/j.neuropsychologia.2022.108230

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *31*(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492. https://doi.org/10.1162/neco.1997.9.7.1483

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*, 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jastorff, J., & Orban, G. A. (2009). Human functional magnetic resonance imaging reveals separation and integration of shape and motion cues in biological motion processing. *Journal of Neuroscience*, *29*(22), 7315–7329. https://doi.org/10.1523/JNEUROSCI.4870-08.2009

Johns, B., Mewhort, D. J. K., & Jones, M. N. (2017). Small worlds and big data: Examining the simplification assumption in cognitive modeling. In *Big data in cognitive science.* (pp. 227–245). Routledge/Taylor & Francis Group.

Jonas, K. G., & Markon, K. E. (2019). Modeling response style using vignettes and person-specific item response theory. *Applied Psychological Measurement*, *43*(1), 3–17. https://doi.org/10.1177/0146621618798663

Joyce, C., & Rossion, B. (2005). The face-sensitive N170 and VPP components manifest the same brain processes: The effect of reference electrode site. *Clinical Neurophysiology*, *116*(11), 2613–2631. https://doi.org/10.1016/j.clinph.2005.07.005

Kanske, P., & Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, *42*(4), 987–991. https://doi.org/10.3758/BRM.42.4.987

Katz, L., & Frost, R. (1992). The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis. *Advances in Psychology*, *94*(100), 67–84. https://doi.org/10.1016/S0166-4115(08)62789-2

Kay, K. N., & Yeatman, J. D. (2017). Bottom-up and top-down computations in word- and face-selective cortex. *eLife*, *6*, 1–29. https://doi.org/10.7554/eLife.22341

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. https://cran.r-project.org/package=vwr

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Khanna, M. M., & Cortese, M. J. (2021). How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. *Memory*, *29*(5), 622–636. https://doi.org/10.1080/09658211.2021.1924789

Kherif, F., Josse, G., & Price, C. J. (2011). Automatic top-down processing explains common left occipito-temporal responses to visual words and objects. *Cerebral Cortex*, *21*, 103–114. https://doi.org/10.1093/cercor/bhq063

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112. https://doi.org/10.1162/jocn_a_00148

Kim, A. E., & Gilley, P. M. (2013). Neural mechanisms of rapid sensitivity to syntactic anomaly. *Frontiers in Psychology*, *4*, 1–15. https://doi.org/10.3389/fpsyg.2013.00045

Kim, H. (2021). A k-mismatch string matching for generalized edit distance using diagonal skipping method. *PLOS ONE*, *16*(5), 1–18. https://doi.org/10.1371/journal.pone.0251047

Kim, S. G., Richter, W., & UÇ§urbil, K. (1997). Limitations of temporal resolution in functional MRI. *Magnetic Resonance in Medicine*, *37*(4), 631–636. https://doi.org/10.1002/mrm.1910370427

Kinoshita, S., Robidoux, S., Guilbert, D., & Norris, D. (2015). Context-dependent similarity effects in letter recognition. *Psychonomic Bulletin and Review*, *22*(5), 1458–1464. https://doi.org/10.3758/s13423-015-0826-3

Kinoshita, S., Robidoux, S., Mills, L., & Norris, D. (2014). Visual similarity effects on masked priming. *Memory and Cognition*, *42*(5), 821–833. https://doi.org/10.3758/s13421-013-0388-4

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). "What's new in Psychtoolbox-3?". *Perception 36 ECVP Abstract Supplement*. https://doi.org/10.1068/v070821

Kok, P., & De Lange, F. P. (2014). Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. *Current Biology*, *24*(13), 1531–1535. https://doi.org/10.1016/j.cub.2014.05.042

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83–91.

Kondrak, G., & Dorr, B. (2006). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29–42. https://doi.org/10.1016/j.artmed.2005.07.005

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*(1), 14–34. https://doi.org/10.1037/a0021446

Kovesi, P. (2003). Phase congruency detects corners and edges. In C. Sun, H. Talbot, S. Ourselin, & T. Adriaansen (Eds.), *Digital image computing: Techniques and applications: Proceedings of the viith biennial australian pattern recognition society conference - dicta 2003* (pp. 309–318). Csiro Publishing.

Krafnick, A. J., Tan, L. H., Flowers, D. L., Luetje, M. M., Napoliello, E. M., Siok, W. T., Perfetti, C., & Eden, G. F. (2016). Chinese Character and English Word processing in children's ventral occipitotemporal cortex: FMRI evidence for script invariance. *NeuroImage*, *133*, 302–312. https://doi.org/10.1016/j.neuroimage.2016.03.021

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, *41*(6), 1648–1662. https://doi.org/10.1037/xlm0000128

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kronbichler, M., Hutzler, F., Wimmer, H., Mair, A., Staffen, W., & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *NeuroImage*, *21*(3), 946–953. https://doi.org/10.1016/j.neuroimage.2003.10.021

Kuennapas, T., & Janson, A. J. (1969). Multidimensional similarity of letters. *Perceptual and motor skills*, *28*(1), 3–12. https://doi.org/10.2466/pms.1969.28.1.3

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kurita-Tashima, S., Tobimatsu, S., Nakayama-Hiromatsu, M., & Kato, M. (1991). Effect of check size on the pattern reversal visual evoked potential. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *68*(3), 219–222. https://doi.org/10.1016/0168-5597(87)90029-3

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kuwahata, H., Adachi, I., Fujita, K., Tomonaga, M., & Matsuzawa, T. (2004). Development of schematic face preference in macaque monkeys. *Behavioural Processes*, *66*(1), 17–21. https://doi.org/10.1016/j.beproc.2003.11.002

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language*, *98*(1), 74–88. https://doi.org/10.1016/j.bandl.2006.02.003

Lee, C. Y., Liu, Y. N., & Tsai, J. L. (2012). The time course of contextual effects on visual word recognition. *Frontiers in Psychology*, *3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00285

Lewis, G., Solomyak, O., & Marantz, A. (2011). The neural basis of obligatory decomposition of suffixed words. *Brain and Language*, *118*(3), 118–127. https://doi.org/10.1016/j.bandl.2011.04.004

Li, J., Osher, D. E., Hansen, H. A., & Saygin, Z. M. (2020). Innate connectivity patterns drive the development of the visual word form area. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-75015-7

Li, X., Harbottle, G., Zhang, J., & Wang, C. (2003). The earliest writing? Sign use in the seventh millennium BC at Jiahu, Henan Province, China. *Antiquity*, *77*(295), 31–44. https://doi.org/10.1017/S0003598X00061329

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*(August), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lien, M. C., Allen, P. A., & Ruthruff, E. (2021). Multiple routes to word recognition: evidence from event-related potentials. *Psychological Research*, *85*(1), 151–180. https://doi.org/10.1007/s00426-019-01256-5

Ling, S., Lee, A. C. H., Armstrong, B. C., & Nestor, A. (2019). How are visual words represented? Insights from EEG-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, *40*(17), 5056–5068. https://doi.org/10.1002/hbm.24757

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*(1171), 1–16. https://doi.org/10.3389/fpsyg.2015.01171

Locke, L. L. (1912). The ancient Quipu, a Peruvian knot record. *American Anthropologist*, *14*(2), 325–332.

López-Barroso, D., Thiebaut de Schotten, M., Morais, J., Kolinsky, R., Braga, L. W., Guerreiro-Tauil, A., Dehaene, S., & Cohen, L. (2020). Impact of literacy on the functional connectivity of vision and language related networks. *NeuroImage*, *213*, 1–12. https://doi.org/10.1016/j.neuroimage.2020.116722

Lopukhina, A., Konstantin, L., & Laurinavichyute, A. (2021). Morphosyntactic but not lexical corpus-based probabilities can substitute for cloze probabilities in reading experiments. *PLOS ONE*, *16*(1), 1–26. https://doi.org/10.1371/journal.pone.0246133

Lu, C., Li, H., Fu, R., Qu, J., Yue, Q., & Mei, L. (2021). Neural representation in visual word form area during word reading. *Neuroscience*, *452*, 49–62. https://doi.org/10.1016/j.neuroscience.2020.10.040

Luke, S. G., & Christianson, K. (2015). Predicting inflectional morphology from context. *Language, Cognition and Neuroscience*, *30*(6), 735–748. https://doi.org/10.1080/23273798.2015.1009918

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, *50*, 826–833. https://doi.org/10.3758/s13428-017-0908-4

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Madec, S., Le Goff, K., Anton, J.-L., Longcamp, M., Velay, J.-L., Nazarian, B., Roth, M., Courrieu, P., Grainger, J., & Rey, A. (2016). Brain correlates of phonological recoding of visual symbols. *NeuroImage*, *132*, 359–372. https://doi.org/10.1016/j.neuroimage.2016.02.010

Madec, S., Le Goff, K., Riès, S. K., Legou, T., Rousselet, G., Courrieu, P., Alario, F. X., Grainger, J., & Rey, A. (2016). The time course of visual influences in letter recognition. *Cognitive, Affective and Behavioral Neuroscience*, *16*(3), 406–414. https://doi.org/10.3758/s13415-015-0400-5

Maier, M., & Abdel Rahman, R. (2019). No matter how: Top-down effects of verbal and semantic category knowledge on early visual perception. *Cognitive, Affective and Behavioral Neuroscience*, *19*(4), 859–876. https://doi.org/10.3758/s13415-018-00679-8

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004

Marcet, A., & Perea, M. (2018). Can I order a burger at rnacdonalds.com? Visual similarity effects of multi-letter combinations at the early stages of word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(5), 699–706. https://doi.org/10.1037/xlm0000477

Martin, L., Durisko, C., Moore, M. W., Coutanche, M. N., Chen, D., & Fiez, J. A. (2019). The VWFA is the home of orthographic learning when houses are used as letters. *eNeuro*, *6*(1), 1–13. https://doi.org/10.1523/ENEURO.0425-17.2019

Martín-Loeches, M., Nigbur, R., Casado, P., Hohlfeld, A., & Sommer, W. (2006). Semantics prevalence over syntax during sentence processing: A brain potential study of noun-adjective agreement in Spanish. *Brain Research*, *1093*(1), 178–189. https://doi.org/10.1016/j.brainres.2006.03.094

Masson, M. E., & MacLeod, C. M. (2002). Covert operations: Orthographic recoding as a basis for repetition priming in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5), 858–871. https://doi.org/10.1037/0278-7393.28.5.858

Matar, S., Pylkkänen, L., & Marantz, A. (2019). Left occipital and right frontal involvement in syntactic category prediction: MEG evidence from Standard Arabic. *Neuropsychologia*, *135*. https://doi.org/10.1016/j.neuropsychologia.2019.107230

MATLAB. (2020). *MATLAB Version 9.9.0 (R2020b)*. The MathWorks Inc.

Mattavelli, G., Rosanova, M., Casali, A. G., Papagno, C., & Lauro, L. J. R. (2013). Top-down interference and cortical responsiveness in face processing: A TMS-EEG study. *NeuroImage*, *76*, 24–32. https://doi.org/10.1016/j.neuroimage.2013.03.020

Maurer, U., Brandeis, D., & McCandliss, B. D. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. *Behavioral and Brain Functions*, *1*, 1–12. https://doi.org/10.1186/1744-9081-1-13

Maurer, U., Brem, S., Bucher, K., & Brandeis, D. (2005). Emerging neurophysiological specialization for letter strings. *Journal of Cognitive Neuroscience*, *17*(10), 1532–1552. https://doi.org/10.1162/089892905774597218

Maurer, U., Rossion, B., & McCandliss, B. D. (2008). Category specificity in early perception: Face and word N170 responses differ in both lateralization and habituation properties. *Frontiers in Human Neuroscience*, *2*, 1–7. https://doi.org/10.3389/neuro.09.018.2008

Maurer, U., Zevin, J. D., & McCandliss, B. D. (2008). Left-lateralized N170 effects of visual expertise in reading: Evidence from Japanese syllabic and logographic scripts. *Journal of Cognitive Neuroscience*, *20*(10), 1878–1891. https://doi.org/10.1162/jocn.2008.20125

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*(7), 293–299. https://doi.org/10.1016/S1364-6613(03)00134-7

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An acccount of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

McManus, J. N., Li, W., & Gilbert, C. D. (2011). Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9739–9746. https://doi.org/10.1073/pnas.1105855108

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930. https://doi.org/10.1016/j.cortex.2011.02.019

Molko, N., Cohen, L., Mangin, J. F., Chochon, F., Lehéricy, S., Le Bihan, D., & Dehaene, S. (2002). Visualizing the neural bases of a disconnection syndrome with diffusion tensor imaging. *Journal of Cognitive Neuroscience*, *14*(4), 629–636. https://doi.org/10.1162/08989290260045864

Momma, S., & Phillips, C. (2018). The relationship between parsing and generation. *Annual Review of Linguistics*, *4*, 233–254. https://doi.org/10.1146/annurev-linguistics-011817-045719

Moulton, E., Bouhali, F., Monzalvo, K., Poupon, C., Zhang, H., Dehaene, S., Dehaene-Lambertz, G., & Dubois, J. (2019). Connectivity between the visual word form area and the parietal lobe improves after the first year of reading instruction: a longitudinal MRI study in children. *Brain Structure and Function*, *224*(4), 1519–1536. https://doi.org/10.1007/s00429-019-01855-3

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, *139*(1), 19–37. https://doi.org/10.1016/j.actpsy.2011.09.014

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Muneaux, M., & Ziegler, J. C. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, *19*(5), 641–660. https://doi.org/10.1080/01690960444000052

Neath, I., & Surprenant, A. M. (2020). Concreteness and disagreement: Comment on Pollock (2018). *Memory and Cognition*, *48*, 683–690. https://doi.org/10.3758/s13421-019-00992-8

Neisser, U. (1967). *Cognitive Psychology*. Prentice-Hall. https://doi.org/10.4324/9781315736174

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*(2), 151–165. https://doi.org/10.1162/jocn.1991.3.2.151

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

Nobre, A. C., Allison, T., & McCarthy, G. (1994). Word recognition in the human inferior temporal lobe. *372*, 260–263. https://doi.org/10.1038/372260a0

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524. https://doi.org/10.1016/j.tics.2013.08.003

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review*, *119*(3), 517–545. https://doi.org/10.1037/a0028450

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. g. l. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Parker, A. J., Egan, C., Grant, J. H., Harte, S., Hudson, B. T., & Woodhead, Z. V. (2021). The role of orthographic neighbourhood effects in lateralized lexical decision: a replication study and meta-analysis. *PeerJ*, *9*, 1–25. https://doi.org/10.7717/peerj.11266

Parr, L. A., Hecht, E., Barks, S. K., Preuss, T. M., & Votaw, J. R. (2009). Face processing in the chimpanzee brain. *Current Biology*, *19*(1), 50–53. https://doi.org/10.1016/j.cub.2008.11.048

Parviainen, T., Helenius, P., Poskiparta, E., Niemi, P., & Salmelin, R. (2006). Cortical sequence of word perception in beginning readers. *Journal of Neuroscience*, *26*(22), 6052–6061. https://doi.org/10.1523/JNEUROSCI.0673-06.2006

Pastore, M. (2018). Overlapping: a R package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, *3*(32), 1023. https://doi.org/10.21105/joss.01023

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, *10*(1089), 1–8. https://doi.org/10.3389/fpsyg.2019.01089

Pattamadilok, C., Morais, J., Colin, C., & Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language*, *137*, 103–111. https://doi.org/10.1016/j.bandl.2014.08.005

Pattamadilok, C., Perre, L., & Ziegler, J. C. (2011). Beyond rhyme or reason: ERPs reveal task-specific activation of orthography on spoken language. *Brain and Language*, *116*(3), 116–124. https://doi.org/10.1016/j.bandl.2010.12.002

Pattamadilok, C., Planton, S., & Bonnard, M. (2019). Spoken language coding neurons in the Visual Word Form Area: Evidence from a TMS adaptation paradigm. *NeuroImage*, *186*, 278–285. https://doi.org/10.1016/j.neuroimage.2018.11.014

Pecher, D., De Rooij, J., & Zeelenberg, R. (2009). Does a pear growl? Interference from semantic properties of orthographic neighbors. *Memory and Cognition*, *37*(5), 541–546. https://doi.org/10.3758/MC.37.5.541

Pecher, D., Wagenmakers, E. J., & Zeelenberg, R. (2005). Enemies and friends in the neighborhood: Orthographic similarity effects in semantic categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(1), 121–128. https://doi.org/10.1037/0278-7393.31.1.121

Pegado, F., Comerlato, E., Ventura, F., Jobert, A., Nakamura, K., Buiatti, M., Ventura, P., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., Braga, L. W., Cohen, L., & Dehaene, S. (2014). Timing the impact of literacy on visual processing. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(49), E5233–E5242. https://doi.org/10.1073/pnas.1417347111

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, *74*(3), 374–388. https://doi.org/10.1016/j.biopsycho.2006.09.008

Perna, A., Tosetti, M., Montanaro, D., & Morrone, M. C. (2008). BOLD response to spatial phase congruency in human brain. *Journal of Vision*, *8*(10), 1–15. https://doi.org/10.1167/8.10.15

Perre, L., Bertrand, D., & Ziegler, J. C. (2011). Literacy affects spoken language in a non-linguistic task: An ERP study. *Frontiers in Psychology*, *2*, 1–8. https://doi.org/10.3389/fpsyg.2011.00274

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain Research*, *1188*(1), 132–138. https://doi.org/10.1016/j.brainres.2007.10.084

Perrin, F., Pernier, J., & Bertrand, O. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical Neurophysiology*, *72*, 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science*, *21*(3), 1–8. https://doi.org/10.1111/desc.12572

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, *331*(6157), 585–589. https://doi.org/10.1038/331585a0

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., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body-object interaction ratings for more than 9,000 English words. *Behavior Research Methods*, *51*(2), 453–466. https://doi.org/10.3758/s13428-018-1171-z

Pfeuffer, J., McCullough, J. C., Van De Moortele, P. F., Ugurbil, K., & Hu, X. (2003). Spatial dependence of the nonlinear BOLD response at short stimulus duration. *NeuroImage*, *18*(4), 990–1000. https://doi.org/10.1016/S1053-8119(03)00035-1

Phillips, J. A., Humphreys, G. W., Noppeney, U., & Price, C. J. (2002). The neural substrates of action retrieval: An examination of semantic and visual routes to action. *Visual Cognition*, *9*(4-5), 662–685. https://doi.org/10.1080/13506280143000610

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS* (J. Chambers, W. Eddy, W. Härdle, S. Sheather, & L. Tierney, Eds.). Springer-Verlag New York. https://doi.org/10.1007/b98882

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Planton, S., Chanoine, V., Sein, J., Anton, J. L., Nazarian, B., Pallier, C., & Pattamadilok, C. (2019). Top-down activation of the visuo-orthographic system during spoken sentence processing. *NeuroImage*, *202*(June). https://doi.org/10.1016/j.neuroimage.2019.116135

Pleisch, G., Karipidis, I. I., Brem, A., Röthlisberger, M., Roth, A., Brandeis, D., Walitza, S., & Brem, S. (2019). Simultaneous EEG and fMRI reveals stronger sensitivity to orthographic strings in the left occipito-temporal cortex of typical versus poor beginning readers. *Developmental Cognitive Neuroscience*, *40*, 1–13. https://doi.org/10.1016/j.dcn.2019.100717

Podgorny, P., & Garner, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics*, *26*(1), 37–52. https://doi.org/10.3758/BF03199860

Polk, T. A., & Farah, M. (2002). Functional MRI evidence for an abstract, not perceptual, word-form area. *Journal of Experimental Psychology: General*, *131*(1), 65–72. https://doi.org/10.1037/0096-3445.131.1.65

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, *50*(3), 1198–1216. https://doi.org/10.3758/s13428-017-0938-y

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, *174*(March), 340–351. https://doi.org/10.1016/j.neuroimage.2018.03.041

Posner, M. I., Sandson, J., Dhawan, M., & Shulman, G. L. (1989). Is word recognition automatic? A cognitive-anatomical approach. *Journal of Cognitive Neuroscience*, *1*(1), 50–60. https://doi.org/10.1162/jocn.1989.1.1.50

Posner, M. I., & Keele, S. W. (1968). On the Genesis of Abstract Ideas. *Journal of Experimental Psychology*, *77*(3), 354–363. https://doi.org/10.1037/h0025953

Posse, S., Ackley, E., Mutihac, R., Rick, J., Shane, M., Murray-Krezan, C., Zaitsev, M., & Speck, O. (2012). Enhancement of temporal resolution and BOLD sensitivity in real-time fMRI using multi-slab echo-volumar imaging. *NeuroImage*, *61*(1), 115–130. https://doi.org/10.1016/j.neuroimage.2012.02.059

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26–46. https://doi.org/10.1.1.443.7693

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Price, C. J., & Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage*, *19*(3), 473–481. https://doi.org/10.1016/S1053-8119(03)00084-3

Price, C. J., & Devlin, J. T. (2011). The Interactive Account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*(6), 246–253. https://doi.org/10.1016/j.tics.2011.04.001

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Constable, R. T., Skudlarski, P., Fulbright, R. K., Bronen, R. A., Shankweiler, D. P., Katz, L., Fletcher, J. M., & Gore, J. C. (1996). Cerebral organization of component processes in reading. *Brain*, *119*(4), 1221–1238. https://doi.org/10.1093/brain/119.4.1221

Purcell, J. J., Shea, J., & Rapp, B. (2014). Beyond the visual word form area: The orthography-semantics interface in spelling and reading. *Cognitive Neuropsychology*, *31*(5-6), 482–510. https://doi.org/10.1080/02643294.2014.909399

Qu, J., Pang, Y., Liu, X., Cao, Y., Huang, C., & Mei, L. (2022). Task modulates the orthographic and phonological representations in the bilateral ventral Occipitotemporal cortex. *Brain Imaging and Behavior*, 1–13. https://doi.org/10.1007/s11682-022-00641-w

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

Rahimi, S., Farahibozorg, S. R., Jackson, R., & Hauk, O. (2022). Task modulation of spatiotemporal dynamics in semantic brain networks: An EEG/MEG study. *NeuroImage*, *246*, 118768. https://doi.org/10.1016/j.neuroimage.2021.118768

Ramsey, R., & Ward, R. (2020). Challenges and opportunities for top-down modulation research in cognitive psychology. *Acta Psychologica*, *209*(June), 103118. https://doi.org/10.1016/j.actpsy.2020.103118

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rauschecker, A. M., Bowen, R. F., Parvizi, J., & Wandell, B. A. (2012). Position sensitivity in the visual word form area. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24). https://doi.org/10.1073/pnas.1121304109

Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding. *Frontiers in Psychology*, *4*(276), 1–8. https://doi.org/10.3389/fpsyg.2013.00276

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Rayner, K., Schotter, E. R., Masson, M. E., Potter, M. C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest, Supplement*, *17*(1), 4–34. https://doi.org/10.1177/1529100615623267

Reich, L., Szwed, M., Cohen, L., & Amedi, A. (2011). A ventral visual stream reading center independent of visual experience. *Current Biology*, *21*(5), 363–368. https://doi.org/10.1016/j.cub.2011.01.040

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. https://doi.org/10.1017/S0140525X03000104

Rodd, J. M. (2004). When do leotards get their spots? Semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin and Review*, *11*(3), 434–439. https://doi.org/10.3758/BF03196591

Rodrigues, A. P., Rebola, J., Pereira, M., Van Asselen, M., & Castelo-Branco, M. (2019). Neural responses of the anterior ventral occipitotemporal cortex in developmental dyslexia: Beyond the visual word form area. *Investigative Ophthalmology and Visual Science*, *60*(4), 1063–1068. https://doi.org/10.1167/iovs.18-26325

Rose, M., Schmid, C., Winzen, A., Sommer, T., & Büchel, C. (2005). The functional and temporal characteristics of top-down modulation in visual selection. *Cerebral Cortex*, *15*(9), 1290–1298. https://doi.org/10.1093/cercor/bhi012

Rossion, B., Joyce, C. A., Cottrell, G. W., & Tarr, M. J. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. *NeuroImage*, *20*(3), 1609–1624. https://doi.org/10.1016/j.neuroimage.2003.07.010

Rossion, B., & Lochy, A. (2022). Is human face recognition lateralized to the right hemisphere due to neural competition with left-lateralized visual word recognition? A critical review. *Brain Structure and Function*, *227*(2), 599–629. https://doi.org/10.1007/s00429-021-02370-0

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. https://doi.org/10.3758/BF03257252

Rousselet, G. A. (2012). Does filtering preclude us from studying ERP time-courses? *Frontiers in Psychology*, *3*, 1–9. https://doi.org/10.3389/fpsyg.2012.00131

Rousselet, G. A., Gaspar, C. M., Wieczorek, K. P., & Pernet, C. R. (2011). Modeling single-trial ERP reveals modulation of bottom-up face visual processing by top-down task constraints (in some subjects). *Frontiers in Psychology*, *2*(137), 1–19. https://doi.org/10.3389/fpsyg.2011.00137

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance vi* (pp. 573–606). Lawrence Erlbaum.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 133–159). The MIT Press.

Rumelhart, D. E., & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review*, *81*(2), 99–118. https://doi.org/10.1037/h0036117

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, *42*(1), 82–95. https://doi.org/10.3758/BRM.42.1.82

Salverda, A. P., & Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(5), 1108–1117. https://doi.org/10.1037/a0019901

Sampson, G. (2016). Writing systems: Methods for recording language. In K. Allan (Ed.), *The routledge handbook of linguistics* (pp. 47–61). Routledge.

Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D., & Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, *19*(9), 1250–1255. https://doi.org/10.1038/nn.4354

Schacht, A., Sommer, W., Shmuilovich, O., Martíenz, P. C., & Martín-Loeches, M. (2014). Differential task effects on N400 and P600 elicited by semantic and syntactic violations. *PLOS ONE*, *9*(3), 1–7. https://doi.org/10.1371/journal.pone.0091226

Schoenmakers, S., Barth, M., Heskes, T., & van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, *83*, 951–961. https://doi.org/10.1016/j.neuroimage.2013.07.043

Schuster, S., Hawelka, S., Richlan, F., Ludersdorfer, P., & Hutzler, F. (2015). Eyes on words: A fixation-related fMRI study of the left occipito-temporal cortex during self-paced silent reading of words and pseudowords. *Scientific Reports*, *5*(July), 1–11. https://doi.org/10.1038/srep12686

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Scott, G. G., O'Donnell, P. J., Leuthold, H., & Sereno, S. C. (2009). Early emotion word processing: Evidence from event-related potentials. *Biological Psychology*, *80*(1), 95–104. https://doi.org/10.1016/j.biopsycho.2008.03.010

Segalowitz, S. J., & Zheng, X. (2009). An ERP study of category priming: Evidence of early lexical semantic access. *Biological Psychology*, *80*(1), 122–129. https://doi.org/10.1016/j.biopsycho.2008.04.009

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E. J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*(5), 1953–1967. https://doi.org/10.3758/s13428-019-01231-3

Sereno, S. C., Brewer, C. C., & O'Donnell, P. J. (2003). Context effects in word recognition: Evidence for early interactive processing. *Psychological Science*, *14*(4), 328–333. https://doi.org/10.1111/1467-9280.14471

Sereno, S. C., Hand, C. J., Shahid, A., Mackenzie, I. G., & Leuthold, H. (2019). Early EEG correlates of word frequency and contextual predictability in reading. *Language, Cognition and Neuroscience*, *35*(5), 625–640. https://doi.org/10.1080/23273798.2019.1580753

Sereno, S. C., & Rayner, K. (2000). The when and where of reading in the brain. *Brain and Cognition*, *42*(1), 78–81. https://doi.org/10.1006/brcg.1999.1167

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport*, *9*(10), 2195–2200. https://doi.org/10.1097/00001756-199807130-00009

Sereno, S. C., Scott, G. G., Yao, B., Thaden, E. J., & O'Donnell, P. J. (2015). Emotion word processing: Does mood make a difference? *Frontiers in Psychology*, *6*, 1–13. https://doi.org/10.3389/fpsyg.2015.01191

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), 1–23. https://doi.org/10.1371/journal.pcbi.1006633

Siero, J. C., Petridou, N., Hoogduin, H., Luijten, P. R., & Ramsey, N. F. (2011). Cortical depth-dependent temporal dynamics of the BOLD response in the human brain. *Journal of Cerebral Blood Flow and Metabolism*, *31*(10), 1999–2008. https://doi.org/10.1038/jcbfm.2011.57

Siew, C. S. (2018). The orthographic similarity structure of English words: Insights from network science. *Applied Network Science*, *3*(1), 1–18. https://doi.org/10.1007/s41109-018-0068-1

Simon, G., Petit, L., Bernard, C., & Rebaï, M. (2007). N170 ERPs could represent a logographic processing strategy in visual word recognition. *Behavioral and Brain Functions*, *3*, 1–11. https://doi.org/10.1186/1744-9081-3-21

Simpson, I. C., Mousikou, P., Montoya, J. M., & Defior, S. (2013). A letter visual-similarity matrix for Latin-based alphabets. *Behavior Research Methods*, *45*(2), 431–439. https://doi.org/10.3758/s13428-012-0271-4

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281. https://doi.org/10.1214/aoms/1177730256

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042–2057. https://doi.org/10.1162/jocn.2009.21296

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598–605. https://doi.org/10.3758/BF03193891

STAN Development Team. (2021). Stan Modeling Language Users Guide and Reference Manual, 2.28. https://mc-stan.org

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

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1280–1293. https://doi.org/10.1037/a0016896

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135–162. https://doi.org/10.1016/j.bandl.2011.07.001

Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with sounds: Sensory substitution selectively activates the visual word form area in the blind. *Neuron*, *76*(3), 640–652. https://doi.org/10.1016/j.neuron.2012.08.026

Strijkers, K., Bertrand, D., & Grainger, J. (2015). Seeing the same words differently: The time course of automaticity and top-down intention in reading. *Journal of Cognitive Neuroscience*, *27*(8), 1542–1551. https://doi.org/10.1162/jocn_a_00797

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(1), 394–398. https://doi.org/10.1073/pnas.0706079105

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods*, *50*(6), 2606–2629. https://doi.org/10.3758/s13428-018-1038-3

Taha, H., Ibrahim, R., & Khateb, A. (2013). How does arabic orthographic connectivity modulate brain activity during visual word recognition: An ERP study. *Brain Topography*, *26*(2), 292–302. https://doi.org/10.1007/s10548-012-0241-2

Takamiya, N., Maekawa, T., Yamasaki, T., Ogata, K., Yamada, E., Tanaka, M., & Tobimatsu, S. (2020). Different hemispheric specialization for face/word recognition: A high-density ERP study with hemifield visual stimulation. *Brain and Behavior*, *10*(6), 1–17. https://doi.org/10.1002/brb3.1649

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96–116. https://doi.org/10.1111/j.1467-9582.2010.01176.x

Tanenhaus, M. K., Flanigan, H. P., & Seidenberg, M. S. (1980). Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition*, *8*(6), 513–520. https://doi.org/10.3758/BF03213770

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Tarkiainen, A., Helenius, P., Hansen, P. C., Cornelissen, P. L., & Salmelin, R. (1999). Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, *122*(11), 2119–2131. https://doi.org/10.1093/brain/122.11.2119

Taylor, J. S., Davis, M. H., & Rastle, K. (2019). Mapping visual symbols onto spoken language along the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(36), 17723–17728. https://doi.org/10.1073/pnas.1818575116

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*, 2372–2382. https://doi.org/10.3758/s13428-020-01389-1

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

The British National Corpus, version 3 (BNC XML Edition). (2007). http://www.natcorp.ox.ac.uk/

Thesen, T., McDonald, C. R., Carlson, C., Doyle, W., Cash, S., Sherfey, J., Felsovalyi, O., Girard, H., Barr, W., Devinsky, O., Kuzniecky, R., & Halgren, E. (2012). Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nature Communications*, *3*, 1–8. https://doi.org/10.1038/ncomms2220

Tinker, M. A. (1928). The relative legibility of the letters, the digits, and of certain mathematical signs. *Journal of General Psychology*, *1*(3-4), 472–496. https://doi.org/10.1080/00221309.1928.9918022

Tobimatsu, S., Kurita-Tashima, S., Nakayama-Hiromatsu, M., Akazawa, K., & Kato, M. (1993). Age-related changes in pattern visual evoked potentials: Differential effects of luminance, contrast and check size. *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials*, *88*(1), 12–19. https://doi.org/10.1016/0168-5597(93)90023-I

Tootell, R. B., Hadjikhani, N. K., Vanduffel, W., Liu, A. K., Mendola, J. D., Sereno, M. I., & Dale, A. M. (1998). Functional analysis of primary visual cortex (V1) in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 811–817. https://doi.org/10.1073/pnas.95.3.811

Turkeltaub, P. E., Goldberg, E. M., Postman-Caucheteux, W. A., Palovcak, M., Quinn, C., Cantor, C., & Coslett, H. B. (2014). Alexia due to ischemic stroke of the visual word form area. *Neurocase*, *20*(2), 230–235. https://doi.org/10.1080/13554794.2013.770873

Uttal, W. R. (1970). Masking of alphabetic character recognition by ultrahigh-density dynamic visual noise. *Perception & Psychophysics*, *7*(1), 19–22. https://doi.org/10.3758/BF03210125

Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 892–903. https://doi.org/10.1037/0096-1523.22.4.892

Van der Stigchel, S., Belopolsky, A. V., Peters, J. C., Wijnen, J. G., Meeter, M., & Theeuwes, J. (2009). The limits of top-down control of visual attention. *Acta Psychologica*, *132*(3), 201–212. https://doi.org/10.1016/j.actpsy.2009.07.001

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace. https://doi.org/10.5555/1593511

van Casteren, M., & Davis, M. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, *39*(4), 973–978.

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

van Paridon, J., Ostarek, M., Arunkumar, M., & Huettig, F. (2021). Does neuronal recycling result in destructive competition? The influence of learning to read on the recognition of faces. *Psychological Science*, *32*(3), 459–465. https://doi.org/10.1177/0956797620971652

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology*, *2*, 1–6. https://doi.org/10.3389/fpsyg.2011.00365

Varga, V., Tóth, D., & Csépe, V. (2020). Orthographic-Phonological mapping and the emergence of visual expertise for print: A developmental event-related potential study. *Child Development*, *91*(1), e1–e13. https://doi.org/10.1111/cdev.13159

Veale, J. F. (2014). Edinburgh Handedness Inventory - Short Form: A revised version based on confirmatory factor analysis. *Laterality*, *19*(2), 164–177. https://doi.org/10.1080/1357650X.2013.783045

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PLoS ONE*, *11*(1), 1–15. https://doi.org/10.1371/journal.pone.0147924

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels Artificial Character Sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods*, *49*(6), 2093–2112. https://doi.org/10.3758/s13428-016-0844-8

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156. https://doi.org/10.1016/j.neuron.2007.05.031

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Kibernetika*, *4*, 52–57. https://doi.org/10.1007/BF01074755

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R., Jones, E., Kern, R., Larson, E., … Vázquez-

Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vogel, A. C., Miezin, F. M., Petersen, S. E., & Schlaggar, B. L. (2012). The putative visual word form area is functionally connected to the dorsal attention network. *Cerebral Cortex*, *22*(3), 537–549. https://doi.org/10.1093/cercor/bhr100

Vogel, A. C., Petersen, S. E., & Schlaggar, B. L. (2014). The VWFA: It's not just for words anymore. *Frontiers in Human Neuroscience*, *8*, 1–10. https://doi.org/10.3389/fnhum.2014.00088

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, *21*(1), 168–173. https://doi.org/10.1145/321796.321811

Walker, C. B. F. (1987). *Reading the Past: Cuneiform*. British Museum Press.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Wang, F., & Maurer, U. (2017). Top-down modulation of early print-tuned neural activity in reading. *Neuropsychologia*, *102*, 29–38. https://doi.org/10.1016/j.neuropsychologia.2017.05.028

Wang, F., & Maurer, U. (2020). Interaction of top-down category-level expectation and bottom-up sensory input in early stages of visual-orthographic processing. *Neuropsychologia*, *137*, 107299. https://doi.org/10.1016/j.neuropsychologia.2019.107299

Wang, J., Deng, Y., & Booth, J. R. (2019). Automatic semantic influence on early visual word recognition in the ventral occipito-temporal cortex. *Neuropsychologia*, *133*, 107188. https://doi.org/10.1016/j.neuropsychologia.2019.107188

Wang, S., Planton, S., Chanoine, V., Sein, J., & Anton, J.-l. (2022). Graph theoretical analysis reveals the adaptive role of the left ventral occipito-temporal cortex in the brain networks during speech processing. *bioRxiv Neuroscience*, 1–46. https://doi.org/10.1101/2022.02.03.478936

Wang, X., Xu, Y., Wang, Y., Zeng, Y., Zhang, J., Ling, Z., & Bi, Y. (2018). Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-21062-0

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Weide, R. (2014). The Carnegie Mellon pronouncing dictionary version 0.7b. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Whaley, M. L., Kadipasaoglu, C. M., Cox, S. J., & Tandon, N. (2016). Modulation of orthographic decoding by frontal cortex. *Journal of Neuroscience*, *36*(4), 1173–1184. https://doi.org/10.1523/JNEUROSCI.2985-15.2016

White, A. L., Palmer, J., Boynton, G. M., & Yeatman, J. D. (2019). Parallel spatial channels converge at a bottleneck in anterior word-selective cortex. *Proceedings of the National*

Academy of Sciences of the United States of America, *116*(20), 10087–10096. https://doi.org/10.1073/pnas.1822137116

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The seriol model and selective literature review. *Psychonomic Bulletin and Review*, *8*(2), 221–243. https://doi.org/10.3758/BF03196158

Whitney, C., Ross, P., Zhou, Z., & Strother. (2019). A novel hypothesis for the original functionality of the Visual Word Form Area: Processing shape sequences. *psyArxiv*, 1–24. https://doi.org/10.31234/osf.io/g3n2m

Wicke, J. D., Donchin, E., & Lindsley, D. B. (1964). Visual evoked potentials as a function of flash luminance and duration. *Science*, *146*(3640), 83–85. https://doi.org/10.1126/science.146.3640.83

Wieser, M. J., & Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, *3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00471

Wilcox, R. R., & Muska, J. (1999). Measuring effect size: A non-parametric analogue of $\omega 2$. *British Journal of Mathematical and Statistical Psychology*, *52*(1), 93–110. https://doi.org/10.1348/000711099158982

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Bosch, A. V. D. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*, 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wilson, S. M., Rising, K., Stib, M. T., Rapcsak, S. Z., & Beeson, P. M. (2013). Dysfunctional visual word form processing in progressive alexia. *Brain*, *136*(4), 1260–1273. https://doi.org/10.1093/brain/awt034

Woodhead, Z. V., Barnes, G. R., Penny, W., Moran, R., Teki, S., Price, C. J., & Leff, A. P. (2014). Reading front to back: MEG evidence for early feedback effects during word recognition. *Cerebral Cortex*, *24*(3), 817–825. https://doi.org/10.1093/cercor/bhs365

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398. https://doi.org/10.1038/s41562-020-00982-w

Xiang, D., Dien, J., & Bolger, D. J. (2019). Testing models of the visual word form area using combined ERP and fMRI using the special properties of Chinese characters. *bioRxiv*. https://doi.org/10.1101/841817

Yao, B., Keitel, A., Bruce, G., Scott, G. G., O'Donnell, P. J., & Sereno, S. C. (2018). Differential emotional processing in concrete and abstract words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(7), 1064–1074. https://doi.org/10.1037/xlm0000464

Yao, P., Staub, A., & Li, X. (2022). Predictability eliminates neighborhood effects during Chinese sentence reading. *Psychonomic Bulletin and Review*, *29*(1), 243–252. https://doi.org/10.3758/s13423-021-01966-1

Yao, Z., & Wang, Z. (2013). The effects of the concreteness of differently valenced words on affective priming. *Acta Psychologica*, *143*(3), 269–276. https://doi.org/10.1016/j.actpsy.2013.04.008

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

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

Yeatman, J. D., Rauschecker, A. M., & Wandell, B. A. (2013). Anatomy of the visual word form area: Adjacent cortical circuits and long-range white matter connections. *Brain and Language*, *125*(2), 146–155. https://doi.org/10.1016/j.bandl.2012.04.010

Zanto, T. P., Rubens, M. T., Thangavel, A., & Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience*, *14*(5), 656–663. https://doi.org/10.1038/nn.2773

Zhao, J., Kipp, K., Gaspar, C., Maurer, U., Weng, X., Mecklinger, A., & Li, S. (2014). Fine neural tuning for orthographic properties of words emerges early in children reading alphabetic script. *Journal of Cognitive Neuroscience*, *26*(11), 2431–2442. https://doi.org/10.1162/jocn_a_00660

Zhao, J., Maurer, U., He, S., & Weng, X. (2019). Development of neural specialization for print: Evidence for predictive coding in visual word recognition. *PLoS Biology*, *17*(10), 1–17. https://doi.org/10.1371/journal.pbio.3000474

Zhao, L., Chen, C., Shao, L., Wang, Y., Xiao, X., Chen, C., Yang, J., Zevin, J., & Xue, G. (2017). Orthographic and phonological representations in the fusiform cortex. *Cerebral Cortex*, *27*, 5197–5210. https://doi.org/10.1093/cercor/bhw300

Zhou, W., Wang, X., Xia, Z., Bi, Y., Li, P., & Shu, H. (2016). Neural mechanisms of dorsal and ventral visual regions during text reading. *Frontiers in Psychology*, *7*(1399), 1–10. https://doi.org/10.3389/fpsyg.2016.01399

Zhou, Z., Vilis, T., & Strother, L. (2019). Functionally Separable Font-invariant and Font-sensitive Neural Populations in Occipitotemporal Cortex. *Journal of Cognitive Neuroscience*, *31*(7), 1018–1029. https://doi.org/10.1162/jocn_a_01408

Zhou, Z., Whitney, C., & Strother, L. (2019). Embedded word priming elicits enhanced fMRI responses in the visual word form area. *PLoS ONE*, *14*(1), 1–18. https://doi.org/10.1371/journal.pone.0208318

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793. https://doi.org/10.1016/S0749-596X(03)00006-8

Zou, L., Desroches, A. S., Liu, Y., Xia, Z., & Shu, H. (2012). Orthographic facilitation in Chinese spoken word recognition: An ERP study. *Brain and Language*, *123*(3), 164–173. https://doi.org/10.1016/j.bandl.2012.09.006

Zweig, E., & Pylkkänen, L. (2009). A visual M170 effect of morphological complexity. *Language and Cognitive Processes*, *24*(3), 412–439. https://doi.org/10.1080/01690960802180420