Atypically-Reading Adults: A Profile

An Exploratory, Longitudinal Study of Single Word Recognition Processes

This thesis is submitted for the degree of Doctor of Philosophy

Emma Louise Anouska Mills
Department of Psychology, Lancaster University

September, 2023

Abstract

Atypically-Reading Adults: An Exploratory, Longitudinal Study of Single Word Recognition Processes

Emma Mills

Approximately 16% of school leavers cannot read to a sufficient skill level so as to be called "functionally literate" (Castles et al., 2018; Leitch, 2006). This exploratory study explores the single word recognition processes of a group of atypically-reading adults in comparison with groups of younger and older readers.

In the main study, we assessed orthographic, phonological and semantic skills longitudinally. We estimated their influence plus that of psycholinguistic properties such as word-frequency, consistency and neighbourhood-size on single word recognition processes by way of reaction time and accuracy data from four experimental tasks (letter search, lexical decision, single word naming and sentence reading).

To support the estimation of our statistical models for the main study, we conducted a wide ranging meta-analysis of psycholinguistic predictor effects. We report the findings here and introduce the study as an accessible resource for use by the research community.

Linear-mixed-effects-models estimated that the rate of change in reading-related skills was either too small or too slow to detect within the time-frame or data. Adult-learners perform similarly to all comparison groups in response latencies across all tasks. They perform similarly to 11-12- and 16-17-year-old readers in the lexical decision and sentence reading accuracy measures. They are more accurate in letter search and less accurate in word naming accuracy measures.

Nonword reading skill, rather than word reading skill, is a reliable predictor in this sample. Word-frequency, age-of-acquisition, consistency and neighbourhood size show influence across tasks.

We interpret the results through the lexical quality hypothesis. The predictors that are influential across the models, and the similarity of adult-learners' performance to younger readers suggests that their orthographic, phonological and semantic knowledge is weakly correlated. Further, adult-learners may be using a dominant reading strategy that reflects sublexical processing, thereby impeding development of orthographic learning and knowledge over the longer term.

Contents

List of Tables

List of Figures

Acknowledgements

A great debt of thanks is owed to Dr. Rob Davies. Your generosity changes lives. Thank you. Thank you also to Dr. Anna Woollams for advice and guidance in the initial stages of the study. Nadine Wilson and Clare Race, whose administrative support was critical. Jonathan Barbrook, information specialist for the meta-analysis - thank you all. Mike Pacey, for support and advice when learning to use Lancaster's computer cluster - thank you. Dr. Andy Wharton, for hardware and software support, sounding board and supporting cups of tea and my partner, thank you.

More formally, many thanks to all the participants - community based adults, students and staff of the educational institutions, for their commitment and generous donation of their time.

This work was funded by the Economic and Social Research Council.

Declaration

The work submitted in this thesis is my own and has not been submitted towards the award of another degree or other qualifying work by myself or another person.
Sources of assistance are properly acknowledged and works appropriately referenced.

Name: Emma Mills

Signature:

Date: 28th March 2024

Introduction

Reading is a skill used in everyday life. As a critical component of literacy skills, poor reading skills are linked to reduced employment opportunities, lower health status and reduced social mobility (Wheater & Worth, 2014). The OECD estimates the number of adults without at least a C grade (or level 5 in the new grading structure) at GCSE English at 16.4%. Leitch (2006) categorised these adults as "functionally illiterate" and recommended that such individuals have the opportunity to engage in free GCSE courses to increase their literacy skills.

The approach, skills and content of the free GCSE course with which such adults engage is similar to that of 16-year-olds in their final year of secondary school who are preparing to take GCSE English for the first time. This may be appropriate if the assumption is correct that these adults are at the skill level of a final year GCSE candidate. The National Literacy Trust, however, estimates that approximately 14.6% of British adults have literacy levels equivalent to that of an 11-year-old learner (Morrisroe, 2014). This raises doubt about the precise level of literacy skills in this adult cohort. Are they more similar to either 16-year-old, final year GCSE candidates or to younger students in their literacy skills?

There is little research that speaks to this argument. Most reading research studies are conducted comparing skilled adult readers (Adelman et al., 2014; Andrews & Hersch, 2010; Yap et al., 2012) with readers with diagnosed reading disorders, such as dyslexia (Bruck, 1990; Murphy et al., 1988; Szeszulski & Manis, 1987). This cohort of adults would not fit into either of these participant groups. Given that these adults appear to be like younger readers, we may look to research findings from studies involving younger participants (Castles et al., 2018; Ehri & Wilce, 1983). However, the age range of such research may be *too* young, focusing largely on readers below 11-years-of-age. Research that explores reading between the ages of 11-16-years-of-age

is sparse. This represents a knowledge gap in reading research.

The adults of concern in the present study are neither skilled readers, as defined by their low achievement in terminal assessments, nor do they demonstrate a severity of low performance that may lead to a formal diagnosis of a reading disorder. Additionally, they are much older than younger, typically developing readers, with a great deal more natural language experience. The longer exposure to natural language could confer an advantage for this group of readers, however we do not know. This thesis aimed to investigate such questions. We explored single word reading performance for well-established experimental tasks such as lexical decision and word naming across a sample of younger and older typically and atypically developing readers. Our primary focus was centred upon a group of adults who were accessing free GCSE English classes within further education institutions. Using a longitudinal design, we tested participants on three separate occasions to better understand when, if any, change in performance measures occurs across the school year.

As well as lexical decision and single word naming, we included a letter search and a single sentence reading task. We aimed to better understand if adults show equivalent skill in recognising those critical sublexical parcels of information that letters represent and that are a precursor to strong reading skills. Having participants name words in sentence contexts helped us understand if wrapping single words in meaningful and non-meaningful contexts enhances word recognition efficiency, and whether these adults performed in similar ways to that observed across other participants.

We also collected measures of individual differences in word and nonword reading, phonological awareness, spelling and vocabulary knowledge and rapid automatic naming. To make inferences about different or similar skill profiles, we conducted simple statistical tests on the individual differences data, contributing new knowledge to this sparsely documented space in reading research.

These individual difference measures were coupled with psycholinguistic measures for the single word stimulus items in our four experimental tasks. Psycholinguistic predictors such as word frequency and consistency are known to play

important roles in single word reading for developing readers. These predictors are strongly implicated in theoretical and computational models of single word reading, such as the lexical quality hypothesis (Perfetti & Hart, 2002) and a series of parallel distributed processes models (Dilkina et al., 2008; Harm and Seidenberg, 2004, 1999; Plaut et al., 1996; Seidenberg and McClelland, 1989), which underpin the thesis. Building statistical models that use a range of psycholinguistic predictors will allow us to connect both to theoretical descriptions and explanations of single word reading processes and the expansive literature that documents observed effects in single word reading for different samples of readers.

Given the sparsity of reading research, and specifically single word reading research, for our focus sample, we wanted to have a clear understanding of the magnitude of psycholinguistic predictors' influence on single word reading performance for well established groups of readers. Additionally, there is little research that considers child and adult participants at the same time. Therefore, we conducted a meta-analysis. Similar effect sizes across child and adult reading groups may indicate that the influence of a predictor is constant across age groups. Differences between groups of child or adult readers may suggest that we should look for interaction effects in statistical models in the longitudinal study. Performing a meta-analysis allowed us to estimate the presence and size of psycholinguistic predictors effects from multiple studies to represent our current state of knowledge for predictor effects on single word reading tasks.

The thesis comprises eight chapters. Chapter 1 introduces theoretical processes of single word reading, discussing the benchmark psycholinguistic effects of consistency, word-frequency and neighbourhood size through the theoretical lenses of the lexical quality hypothesis (Perfetti & Hart, 2002) and computational models of single word reading. Chapter 2 then reviews person-level measures of lexical quality, reviewing components of single word reading and explaining our choices of assessments to capture individual differences in the study participant sample. Chapter 3 locates accounts of individual differences in computational models and human behavioural data for single word reading tasks. In Chapter 4, we present a

meta-analysis, detailing our methods and summary findings for effects across five samples of readers and eight psycholinguistic effects.

From then on, the thesis reports the longitudinal study. Chapter 5 details the study methods. We present our findings in two parts. Chapter 6 presents descriptive statistics of the individual difference measures for the first data collection session. Here, we take time to describe how the adult learner group compares to the other participant groups. Chapter 7 presents our inferential test results for each of the four experimental tasks. In Chapter 8, we discuss our findings and implications for the study in light of extant theory and previous research.

1 Single Word Reading

1.1 Lexical Quality and Word Recognition

The lexical quality hypothesis (Perfetti and Hart, 2002) is a general account that describes how word knowledge may vary within and between individuals and how this variation may impact single word recognition for reading purposes. There are three components of orthographical, phonological and semantic information. A word has high lexical quality when each are simultaneously available for word recognition. A word of high lexical quality may be evidenced by a consistently accurate spelling (Andrews and Hersch, 2010; Castles et al., 2018; Perfetti, 2007), correct pronunciation and knowledge of the word's meaning. A word with low lexical quality may result from any one source of domain information being unavailable and be characterised by slower reaction time or variable pronunciation.

Given an assumption that a highly-skilled reader will have more words of high lexical quality in their vocabulary than a lower-skilled reader (Adelman et al., 2014), behavioural data from tasks associated with the orthographical, phonological and semantic components may describe how participants of different reading skill vary across the components. Supporting evidence comes from Perfetti and Hart (2002). They tested 445 university students for their spelling, auditory awareness, homophone choice, nonword reading, word reading, vocabulary and reading comprehension skills. The distribution of scores in the comprehension test was split to form groups of less-, average- and more-skilled readers with roughly one third of the sample in each group. For each group, Perfetti and Hart (2002) conducted a factor analysis of the assessment scores. A different factored solution was presented for each level of reading skill.

The more-skilled readers were best described by a two factor structure with orthography and phonology loading together onto one factor and semantics on a

second. Critically, orthography scores cross-loaded with the semantic factor while scores representing phonology did not. This suggests that orthography is the linking source of information for words of high lexical quality. The average-skilled readers also showed the same two factor solution as the more-skilled readers however orthography did not show the same cross-loading with semantics. The solution for the less-skilled readers showed three separate factors. Orthography, phonology and semantic variables loaded onto distinct factors. The phonological information cross-loaded with the orthographical factor but not the semantic factor.

Perfetti and Hart (2002) concluded that for more-skilled readers, high lexical quality is expressed by internally coherent factors that capture the three domains of orthography, phonology and semantics and that the factors will correlate strongly with each other (also supported by Yap et al., 2012). Critically, it is orthographic information as the linking variable. Readers of lower comprehension scores showed a less integrated structure and it is phonological information, rather than orthographical information that acts as a bridging source of information, and with orthographical information alone. A difference between readers of more and less skill appears to be the weighting between orthographical and phonological information.

This variation in coherence across skill levels has also been observed in studies using event-related potential (ERP) measurements (Hart and Perfetti, 2008; Yang et al., 2005). Yang et al. (2005) tested higher- and lower-skilled students on a sentence reading task and found that the time course for word processing and semantic retrieval was disparate in the lower-skilled students, producing two, later peaks of ERP signals with one synchronous and earlier ERP peak in the higher-skilled students. Breznitz and Misra (2003) also found differences in the synchronous processing of orthographic and phonological information in ERP signals during a word reading task. Higher-skilled students had equivalent amplitudes for orthographic and phonological information while lower-skilled students demonstrated asynchronicity between the two signals. This lack of coherence in the neurological signal correlated with slower reaction times in the behavioural data of the lower-skilled students.

Taken together, high lexical quality can be indexed by multiple sources of

information that activate strongly together. Lower lexical quality still shows the influence of the multiple sources however the correlation between the sources is weaker. Also apparent is that orthography links sources in more-skilled readers while phonology performs a linking role in less-skilled readers.

Attaining this strong coherence between the three components so that orthographical information activates both phonological and semantic information reflects the challenge of becoming a skilled reader. Years of explicit instruction and practice move a young learner from their pre-literacy language experience comprised of only phonological and semantic information (Chang and Monaghan, 2018; DfE, 2023; Harm and Seidenberg, 1999) to a state where orthographical information is sufficient for fast and accurate recognition of printed text and phonological and semantic information play a supporting role. The English language presents a greater challenge than most languages because words of similar spellings can have different pronunciations and similar pronunciations can have different spellings (Ziegler et al., 1997). Below, we describe the complexity of English language *spelling-sound* relationships.

1.2 English and Spelling-Sound Consistency

There are far fewer spelling-sound groupings to learn than there are possible mappings of an orthographic form to a meaning (Frost, 1998). English as a language is sufficiently regular that it is a much more efficient strategy to learn spelling-sound relationships and build words than to memorise whole words and attach them to referent meanings (Brysbaert, 2019; Taylor et al., 2017). Of critical importance, learning spelling-sound relationships begins the process of simultaneously learning the relative frequency of occurrence for a specific pattern. Over time, they are thought to become represented as a probabilistic distribution within a reader. Explicit knowledge of the spelling-sound relationships together with the implicit knowledge of their distribution within a language is referred to as orthographic knowledge (Nation, 2017; Zevin and Seidenberg, 2006). A person who has the orthographic knowledge of a

language possesses a skill that is generative in nature in that reading and writing of known and unknown words is possible (Mimeau et al., 2018; Nation and Castles, 2017; Share, 1999).

As a deep orthography, English has some complexity, however (Frost, 2012). The key phrase in the above description is "sufficiently regular". There is variability across identical clusters of letters for sound and clusters of sounds for letters. This poses a level of challenge for the learner as they map spelling-sound clusters for both specific word learning but also for learning at the level of orthographic knowledge.

Venezky (1970) described two categories of spelling-sound relationship: predictable and unpredictable. The predictable class contains patterns of letter clusters that, for the majority of the entries counted in the dictionary, followed an invariant pronunciation. A subset of predictable words produced variant pronunciations, however the quantity of pronunciations or frequency of their occurrence was sufficient to render them predictable. This invariance in pronunciation gave good conditions for learning by abstraction of the rules therein or by "transfer" of knowledge of smaller, sublexical sound patterns to unfamiliar words that contained the same pattern.

This was in contrast to the unpredictable class, which contained the remaining words that did not fit the predictable class. Unpredictable words needed paired associative learning strategies or rote learning of a whole word for successful pronunciation. In psycholinguistic research, the labels 'predictable' and 'unpredictable' are more often referred to as 'regular' and 'irregular' with a difference in reaction time and accuracy between the two classes of words producing the robust 'regularity effect', a dichotomous measure composed of regular and irregular words.

Psycholinguistic studies provided evidence of Venezky's (1970) two categories of predictable and unpredictable words, with findings of longer response latencies and higher error rates for "irregular" words compared to "regular" words (Baron and Strawson, 1976; Stanovich & Bauer, 1970 as cited in Glushko, 1979). Readers of acquired phonological dyslexia showed impaired reading of regular words with relatively good reading of irregular words while readers with acquired surface dyslexia

showed the opposite pattern (Patterson and Marcel, 1977; Plaut et al., 1996).

Teaching via the abstraction of rules for predictable words was supported by the regularisations errors for irregular word pronunciation by participants with acquired phonological dyslexia (Coltheart, 1996) and neurotypical skilled readers (Glushko, 1979). The errors across both groups of participants suggested that a cognitive process that supported learning via the abstraction of rules was feasible. Taken together, Venezky's (1970) theoretical suggestions and empirical data from behavioural studies with neurotypical and atypical participants suggested that word recognition occurs via multiple pathways.

This conclusion was challenged by Glushko (1979). In contrast to the stated hypothesis, the majority of pronunciation errors on nonwords made by skilled readers followed partial spelling-sound patterns from irregular words. Glushko (1979) concluded that there was sufficient overlap of spelling-sound information between irregular and regular words for irregular words to be read by analogy to regular words.

The access to sublexical information also suggested that whole word recognition was not necessary since parts of any type of word appeared available for partial processing of an unfamiliar word. This availability of information between regular and irregular words called the suggested structure of multiple pathways for word recognition into question. Since information between orthography and phonology was available for all words, a single pathway would suffice. By extension, rather than a dichotomous metric of regularity vs irregularity for two separate processes, a 'degree' of consistency was suggested that reflected a single process (Andrews, 1997; Glushko, 1979; Seidenberg and McClelland, 1989; Seidenberg et al., 1985).

The two accounts of Venezky (1970) and Glushko (1979) are instrumental in our current day understanding and practice of psycholinguistic research into single word reading processes. Not least because these verbal theories of structure for language and cognition have influenced the two dominant approaches to computational modelling of single word reading, dual route models (Coltheart et al., 2001) and parallel distributed process models (Plaut et al., 1996; Seidenberg and McClelland, 1989, see section 1.5) but also because they shape the construction of

measures for how we quantify the variability of spelling-sound information within and across words.

Different methods of calculating spelling-sound information have been used. The regularity variable is measured at a whole word level. While some researchers maintain a dichotomous split between regular and irregular words, others have grouped words into smaller categories that encompass regular-consistent, regular-inconsistent, exception, strange, and unique spelling-sound groupings to name but a few (Laxon et al., 1991; Seidenberg et al., 1984, 1985; Waters et al., 1984; Waters and Seidenberg, 1985). Findings suggest that regular words are named more accurately and faster than irregular words. Under the smaller grouping conditions, findings are less robust. A general trend is for exception and regular-consistent words to be named more accurately and faster than regular-inconsistent words (Seidenberg et al., 1984). The present work does not use a regularity metric so we do not discuss its construct any further.

Consistency captures sublexical parcels of spelling-sound information, while retaining the possibility of measurement at the whole word level. Ziegler and Goswami (2005) described these sublexical parcels as grain sizes and showed that consistency could be measured across a range of different grain sizes. Grain sizes can be single letters, letter clusters, word onsets, or rimes. A uniquely spelled and pronounced word can be its own grain size.

In a deep orthography, such as English, larger grain sizes demonstrate greater reliability and stability for pronunciation (Treiman et al., 1995). In monosyllabic words, the rime is one of the larger grain sizes and is of primary importance in the measurement of consistency. Jared et al. (1990) and Jared (1997) classified "friend" and "enemy" types of words. A word that shares a target word's rime spelling is either congruent (thus, a friend) or incongruent (thus, an enemy) with the target word's rime pronunciation. For instance, the rime -EAD, has two pronunciations as in MEAD or BREAD. The word BEAD is a friend of MEAD but an enemy of BREAD. The word HEAD is an enemy of MEAD but a friend of BREAD. Jared et al. (1990) and Jared (1997) counted the number of friends and enemies and also summed frequencies of the

friends and enemies (including the specific word) as predictors of consistency.

Adelman and Brown (2007) and Ziegler et al. (1997) constructed a probability ratio of consistency by dividing the number of friends by the sum of the number of friends and enemy type counts. This constrains the measurement to lie between 0 and 1. This consistency metric was a significant, negative predictor in the re-analysis of four large data sets for reaction time, over and above the count of friends and enemy types and log word-frequency.

It is intuitive that a word constructed of simple spelling-sound construction of an invariant pronunciation will be easier to master than a word constructed of complex spelling-sound patterns, for which there may be more than one pronunciation. Additionally, for words that share frequently occurring spelling-sound patterns, the learning of the pattern will be faster as a result of more frequent exposure to familiar words and "transfer" to novel words that contain the pattern (Venezky, 1970). The opportunity to reinforce sublexical patterns between orthographical and phonological information over multiple words is a contributing factor to the strength of the orthographical-phonological relationship that underpins a word's lexical quality.

To summarise, the learning of a novel word that overlaps with a familiar word for spelling-sound information will be facilitated relative to a word that does not overlap in spelling-sound information. Partial decoding of the shared spelling-sound information can contribute to the pronunciation of that part of the novel word. Simultaneously, the information within the sublexical parcel is strengthened, and at the next exposure to the familiar word, this stronger relationship should be reflected in faster recognition and accurate pronunciation. By implication, those words that are dissimilar to others in the corpus will take longer to recognise and parts of the word may be pronounced incorrectly for a longer time.

1.3 Phonological Recoding and Lexical Quality

We have described how the close relationship between orthographic and phonological information can lead to words of high lexical quality and also that when there is

sufficient overlap of spelling-sound information between words, they can facilitate the robust learning of one another. At a specific word level, reading a word strengthens the coherence between the orthographical, phonological and semantic information for the specific information pattern of that discrete word, contributing to orthographic learning of the specific word. Additionally, the distribution of the frequency of occurrence of the sublexical clusters of information within a word accrues with each exposure, building a representation for the orthographic knowledge of a language within an individual over time (Chen, 2008; Samara and Caravolas, 2014; Steacy et al., 2017a).

The crucial aspect of the above is exposure to the orthographic form of a word. It is the orthographic information that *loads* with phonological information and *links* with semantic information in the lexical quality hypothesis (Perfetti and Hart, 2002). It is the greater predictive relationship shared between orthographical and phonological information than that between orthographical and semantic information that forms the basis of the stronger coupling. The mechanism by which orthographical and phonological information can become strongly correlated is called phonological recoding (Share, 1995).

On seeing the orthographic form for a word, phonological recoding states that both an orthographic and phonological code is processed for the word, providing the conditions for strong coupling of orthographical and phonological information (Frost, 1998; Share, 1995). Critically, the recoding mechanism acts in one direction. *Orthographic* recoding does not occur when a word is *heard*. Only a phonological code is processed for the acoustic form of a word and the conditions for strong coupling between the orthographic and phonological information are not present.

Nelson et al. (2005) found that for one visual exposure to a word, higher-skilled readers accessed both an orthographic and a phonological code. In a systematic review, Colenbrander et al. (2019) found that seeing the word at the time of learning facilitates memory for both the pronunciation and spelling of the word. They concluded that efficient visual decoding has twice the learning opportunity for the skilled reader, contributing to the growth of orthographic learning and knowledge

simultaneously.

This is critical to an understanding of how orthographic knowledge and lexical quality may vary within and between people, dependent upon amount and type of exposure to word forms. It suggests that spoken language exposure holds less information for orthographic learning and knowledge development than written language exposure due to a processing of only one source of information at each exposure to words that are heard but not also seen.

This has implications for the transition of words from low lexical quality to high lexical quality. For a word of lower lexical quality, if the information sources are not present to the same extent or not simultaneously available, the opportunity to strengthen an association and bring about strong learning is constrained. The strength of any potential reinforcement may be weaker where the orthographical and phonological information are not simultaneously available (Breznitz and Misra, 2003). Consequently, orthographic learning is less efficient within the episode of an exposure. This necessitates a greater number of exposures of the word to attain a high quality representation, slowing overall development for the number of known words. By implication, if the source of information is a weaker signal by definition, i.e. word forms are heard and not seen, the rate of development may be further slowed.

Furthermore, the incremental contribution to orthographic knowledge per exposure for the distribution of spelling-sound patterns may be smaller or weaker. Since orthographic knowledge works across words, this reduces the influence on all words that share the same structure. When a reader has many words of low lexical quality, the greater number of exposures needed to attain higher quality representations provides greater opportunities for errors. Backman et al. (1984a) found that poorer readers were more variable in their application of orthographical knowledge.

So far, we have seen that the lexical quality hypothesis and the properties of English orthography predict that, for the average reader, words that are experienced more often, are spelled as they sound and are similar in spelling to lots of other words will be easier to read than words that are experienced less often, may have more than

one plausible pronunciation or are less similar in spelling to other words. Experimental findings reflect these predictions. Reaction times and accuracy rates for items that are manipulated for these conditions robustly demonstrate these effects for typical readers (Backman et al., 1984a; Beech and Harding, 1984; Bruck, 1988; Laxon et al., 1991; Lovett, 1987; Olson et al., 1985; Seidenberg et al., 1984; Siegel and Ryan, 1988; Treiman and Hirsh-Pasek, 1985).

1.4 Benchmark Psycholinguistic Effects

It is clear that the rate of development for a word's lexical quality will vary as a function of consistency, the frequency with which a word occurs and the availability of other words with which a target word shares partial spelling-sound structures - i.e., its neighbourhood. These benchmark effects and experimental findings underlie verbal theories of the structure of the reading system. Below we describe three psycholinguistic effects as observed in lexical decision and word naming studies in more detail.

1.4.1 Consistency

As described earlier, words that have only one spelling pattern associated with one pronunciation evoke faster and more accurate responses in lexical decision and word naming tasks than words that have either more than one pronunciation associated with one spelling pattern or one pronunciation generates more than one spelling pattern (Glushko, 1979; Ziegler et al., 1997).

The consistency effect tends to be more robust for word naming task outcomes than lexical decision. This is thought to be due to the phonological requirements of the pronunciation of the item (Waters and Seidenberg, 1985). Effects for consistency wax and wane in lexical decision as a function of the type of items that are included. For instance, Seidenberg et al. (1984) found that a consistency effect only appeared when "strange" words were added to the item sample and when

the lists of different levels of consistency were mixed. Furthermore, the strange words were responded to faster and more accurately than the words that have more than one spelling or pronunciation.

In the individual differences literature, consistency effects tend to be smaller for more-skilled than less-skilled readers (Bruck, 1990; Mahé et al., 2018; Mason, 1978; Romani et al., 2008; Strain and Herdman, 1999), however there are some studies that find equivalent sizes of effects for groups (Ben-Dror et al., 1991; Parrila et al., 2007). Even in strong readers, consistency effects may remain for words of low frequency, less consistent spelling-sound pronunciations and novel words.

Given that all the above studies contrast skilled readers with readers with dyslexia, a contributing factor to the difference is likely to be the persistent difficulties with phonological processing and knowledge that readers with dyslexia experience (Bruck, 1990; Castles and Coltheart, 1993).

1.4.2 Word-frequency

The word-frequency effect is one of the strongest and most robust effects in word recognition studies. Irrespective of the frequency rating measure that is used (Baayen et al., 1995; Kucera and Francis, 1967; Thorndike, 1944; Van Heuven et al., 2014), words rated as occurring more frequently are named or decided upon faster and more accurately than words that occur with lower frequency. Word-frequency effects tend to be larger for lexical decision reaction times than word naming reaction times (Balota et al., 2004).

Word-frequency effects tend to be larger for younger developing readers than older developed readers (Davies et al., 2017; Zoccolotti et al., 2009). Frequency effects for individuals with a history of dyslexia tend to be larger than those for readers without dyslexia (Barber, 2009; Barca et al., 2006; Bruck, 1990; Dujardin et al., 2011; Kuperman, 2013; Suarez-Coalla and Cuetos, 2015). We address this further in the meta-analysis (Chapter 4).

Observations of a stable word-frequency effect over the adult lifespan are

inconsistent across studies. Some studies demonstrate stability over groups of different ages (Allen et al., 1993, 2011; Cohen-Shikora and Balota, 2016; Tainturier et al., 1989) while others see change (Balota and Ferraro, 1993; Balota et al., 2004; Spieler and Balota, 2000).

Frequency often interacts with other psycholinguistic variables. These interactions may arise because the effect of the second variable is most often observed in only lower frequency word items for typically-reading adults while for less-skilled or younger readers, the interaction effects may occur at both levels of word-frequency (Allen et al., 2011; Balota and Ferraro, 1993; Bruck, 1990; Hino and Lupker, 1996; Jared et al., 1990; Lichacz et al., 1999; Seidenberg et al., 1984; Waters et al., 1984).

1.4.3 Neighbourhood-size

Words are described as belonging to the same neighbourhood when they share the same letters across letter positions but for one. Orthographic (ON) and phonological neighbourhood (PN) measures describe the number of new, real words that can be made by changing one letter (orthographic neighbour; e.g. word -> work) or one sound (phonological neighbour: e.g. cat -> caught) of the base word while maintaining the position of the other letters of the base word (Coltheart et al., 1977). Each of these effects describe either a slowing / speeding of response or significant change in accuracy rates as a function of a word's neighbourhood size (N-size).

N-size effects are observed in both word naming and lexical decision tasks (Andrews, 1997). In word naming, a word from a larger neighbourhood is likely to be named faster and more accurately than a word from a smaller neighbourhood. In lexical decision, the effect can present as both facilitatory and inhibitory. In samples of typically-reading adults, Andrews (1997) and Balota et al. (2004) observed facilitatory effects but for words that had a neighbour with a higher frequency value, where they observed an inhibitory effect. Marinus and de Jong (2010) observed this same relationship between relative frequencies of neighbours in child readers.

Effects of N-size tend to attenuate under increasing reading skill in both child

and adult samples (Coltheart et al., 1977; Davies et al., 2017; Dunabeitia and Vidal-Abarca, 2008; Keating, 1987; Laxon et al., 1988). Greater reading skill is equated with a greater quantity of words with high lexical quality. Words of high lexical quality need only orthographical information for recognition; they are less likely to need support from their neighbourhood. In this sense, N-size is a phonological information measure and works at a sublexical level.

N-size often interacts with word-frequency, particularly on low frequency words. Where a word from a large neighbourhood is of low frequency, the orthographic knowledge accrued from its neighbours helps bring about word recognition. This effect is observed in adults (Balota et al., 2004) and children (Marinelli et al., 2013). N-size will also often show interaction effects with length, with longer words showing N-size effects rather than shorter words, especially for those of low frequency (Davies et al., 2007).

Variant measures of N-size have been constructed. Phonographic neighbours (Adelman and Brown, 2007) describe neighbours that are made by a single change of a consonant letter that maintains the vowel sound of the base word. Phonographic neighbours are named faster than orthographic neighbours.

Yarkoni et al. (2008) defined an orthographic Levenshtein distance metric (OLD). In OLD, the target word is compared to each word in a corpus database and counts made of the number of insertions, deletions, and substitutions to change the target word to its comparator. OLD20 is the mean number of changes taken over the first 20 of the base word's closest orthographic neighbours. As a variant of ON, the OLD metric correlates strongly with ON and allows for an item sample that contains a much wider range of length of words.

1.5 Accounts of Word Recognition from Computational Models

While empirical studies find effects of key psycholinguistic variables, converging lines of evidence also arise from simulations with computational models. Two models for single word recognition dominate the literature, the Dual Route Cascaded model (DRC, Coltheart et al., 2001; Pritchard et al., 2018) and parallel distributed process models (PDP) built using connectionist network principles (Dilkina et al., 2008; Harm and Seidenberg, 2004, 1999; Hoffman et al., 2018; Plaut et al., 1996; Seidenberg and McClelland, 1989). Both models are able to simulate frequency, consistency and N-size effects and perform word naming and lexical decision tasks. They differ from each other in theoretical assumptions and architecture, however.

1.5.1 The Dual Route Cascaded Model of Visual Word Recognition

The DRC (Coltheart et al., 2001) simulates mature skilled reading. The model has two independent pathways to word recognition: a lexical route processes whole, irregular or regular familiar words and a non-lexical route that processes novel words that are regular in spelling-sound patterns. Both routes are fed by a visual input layer and feed into a phoneme recognition, output layer.

Knowledge in the DRC is stored in two forms, whole word representations and letters. The lexical route contains an orthographic lexicon, in which there is a whole word node for every word of the CELEX corpus (Baayen et al., 1995). The non-lexical route has nodes for each letter of the alphabet that represent a complete knowledge of the grapheme to phoneme correspondence (GPC) rules (Adelman et al., 2014) for the host language. Servicing both routes is a phonological lexicon containing pronunciation nodes that are matched for every node in the orthographic lexicon. Both routes feed into the phoneme recognition layer where single phonemes of the

host language are represented.

The two routes use different methods of processing orthographical input. The lexical route *addresses* the lexicon for the whole word representation by processing each letter in the item in parallel. Every node in the orthographic lexicon that contains a letter in the same position as the item is activated by cascaded signals from the lexical route. All nodes that do not share the letter at that position are inhibited. The single orthographic lexicon node with the highest activation feeds forward activation to the matching node in the phonological lexicon. Subsequently, activation passes to the phoneme system, where the phonemic representation of the word is assembled.

In the non-lexical route unfamiliar words are *assembled* by serial processing from left to right. Letters are processed individually, and single, two and three letter GPC clusters are searched for applicable rules. Upon attributing a rule, the corresponding node in the phoneme system is given excitatory activation. This serial search process continues until all letters in the item are processed.

On presentation of a letter string, both routes process the available information. In the case of irregular / exception words, a whole word from the lexical route and a regularised pronunciation from the non-lexical route are fed forward to the phoneme layer and will conflict with each other. The additional time taken by the model to resolve this conflict produces the regularity effect where irregular words take longer and are more likely to be mispronounced than regular words. In this way, the architecture of the DRC recovers a regularity effect rather than a consistency effect.

The DRC replicates the word-frequency effect by weighting the activation rates of the nodes in the orthographic lexicon by a scaled frequency parameter. All CELEX frequency values are scaled such that the word with the lowest frequency has a value of -1 and the highest frequency word has a value of 0; all other frequency values fall in between. Nodes with values closer to zero (i.e. of higher frequency) are activated at a faster rate than nodes with values closer to -1 (i.e. of lower frequency), producing a word-frequency effect. The same level of activation for an orthographic node is fed forward to the matching node in the phonological lexicon, ensuring the

frequency effect is maintained. There is no frequency effect for the non-lexical route as the unit of analysis is the letter level rather than the word level.

N-size effects are returned by the DRC. Items receive support from activated nodes in the orthographic lexicon that share the same letter in the same position. Nonwords that share a spelling-sound structure with words from high neighbourhoods receive greater activation than nonwords that share spelling-sound structures with words of low neighbourhoods, resulting in faster recognition. Researchers relaxed the parameters for lateral inhibition of nodes in the orthographic lexicon so that word nodes could support word letter string recognition in the lexical route to produce an N-size effect, without which the returned N-size effect is muted (Seidenberg and Plaut, 2006).

1.5.2 Parallel Distributed Process Models

The architecture of a PDP model consists of three layers. Each layer contains units specialised for processing either orthographic, phonological or semantic information of the item. Units are connected to each other within a layer and layers are connected to each other with bidirectional flow of information. A layer of hidden units mediates transfer of information between layers. Each unit bears a weight that determines the unit's activation rate. At the beginning of training, the weights on the connections are set to random values.

There are no whole word representations as in the DRC, instead the family of PDP models distributes grapheme and phoneme knowledge across the units within orthographic and phonological layers (Plaut et al., 1996). Each inputted letter string is consequently represented by several units within each of the three layers, representing the orthographic, phonological and semantic information of the target word.

All sources of information are processed identically via spreading activation (Dilkina et al., 2008; Harm and Seidenberg, 2004, 1999; Plaut et al., 1996; Seidenberg and McClelland, 1989). Spreading activation means that information from the input

is accrued over time at the units within layers. Shared information between units promotes excitatory activation and units that do not share information with the input receive inhibitory activation. The cluster of units with the strongest levels of activation becomes the model's representation of the input.

The pattern of activated units is "learned" due to small additions made to weights on each successfully activated unit at each exposure. This moves the resting state of the unit closer to the threshold value. Activated once more, the unit will be quicker to reach threshold with the iterative changes to weights inducing faster responses over time.

The location of learning at the unit level, rather than the cluster of units for an input, supports both learning for the specific input and also the mapping of a statistical distribution of the spelling-sound patterns that is the orthographic knowledge of the model language (Seidenberg and McClelland, 1989).

The word-frequency effect is recovered by the PDP model as a function of the training cycle and the resultant adaptations made on the weights over time. Word items within a training set are exposed to the model at rates that reflect their frequency values from a chosen corpus (Plaut, 1997; Plaut et al., 1996; Seidenberg and McClelland, 1989). It follows that those words of higher frequency will have the greater amount of exposure and the adaptation on the weights will be greater, contributing to faster activation for those letter patterns.

The variation in observed frequency effect sizes is explained by the adaptive function of the weights. The propensity for adaptation on the next exposure is a fraction of the amount of change in the preceding exposure. Changes accrue in smaller amounts over repeated exposures as the weights move toward the threshold limit. These smaller changes reflect the small effects observed for words of high frequency over time. Larger effects for lower frequency words are explained by the change to the weights per exposure being relatively large due to a lower exposure rate and thus fewer episodes of adaptation to the weights' from their initial starting values.

The same logic applies for consistency effects. Consistent words of invariant pronunciation activate identical units and incur changes on the weights with every

exposure. In contrast, words of variant pronunciation share their spelling pattern with multiple plausible pronunciations. The activation for each exposure is diverted to only one of the possible phoneme units where the weight is subsequently adapted. Consequently, the range of units associated with possible pronunciations are activated less frequently and the respective weights change their values at a slower rate. The pattern of activation for each pronunciation is slower to stabilise, producing a difference in the recognition and accuracy rates for words of differing consistency values.

N-size effects arise due to the distributed nature of the knowledge within the PDP architecture. A large neighbourhood indicates that a word shares a portion of its spelling-sound structure with many other words. The greater number of words that contain the same pattern of letters induces greater frequency of activation for the shared unit or collection of units than words that share letters with only a small number of other words. Additionally, the historical record of exposure is stored at the weight on a unit. Units may be resting at a higher value because they are a high frequency sublexical unit that is a also part of a large neighbourhood. Collectively, a large neighbourhood set of words contributes to a faster adaptation of the weight on the unit for the shared letters, facilitating recognition of words from large neighbourhoods relative to words from small neighbourhoods.

In summary, the DRC and the PDP contribute contrasting processes to emulate the same psycholinguistic effects. Apart from the obvious structural and processing differences, two further critical differences between the models are the representation of knowledge and the location of the word-frequency effect.

In the DRC model, knowledge is represented as whole words and single letters while in the PDP model, it is in graphemes and phonemes of varying sizes. The patterns of unit activation are representations of whole words, generated dynamically upon presentation of a letter-string. Conceptually, the PDP model's structure allows for representation of whole words without the need for a separate orthographic lexicon as in the DRC, or a separate route to manage exception words. Exception words and familiar words are represented in the same way as consistent words and unfamiliar

words, with all units able to contribute to recognition and the pattern trace of a word being stored across several units in their respective weights.

The location of the frequency effect is fixed at the individual word node in the orthographic lexicon of the DRC model, and passed to the phonological lexicon as needed. In the PDP model, the frequency effect resides in the weights of the units in each layer. Units contribute at different rates which gives the range of effects sizes that are possible for the frequency effect. Each of these instances are managed by the adaptive function on the weights which produces the change required as a function of the rate of exposure to different items.

In the above, we have described how efficient reading may depend upon an individual having many word representations of high lexical quality, but that the English orthography makes the attainment of high lexical quality for a word quite challenging. The properties of the English orthography predict certain psycholinguistic effects that help us understand what makes a word harder or easier to attain high lexical quality. The primary effects of word-frequency, consistency and N-size are robustly observed in the experimental literature with human participants and have been recovered in simulated data by the DRC and the series of PDP models. In the next chapter, we turn to a discussion of how orthography, phonology and semantics is represented in the typical reader and how effects may vary across people.

2 Person-Level Measures of Lexical Quality

In the previous chapter we considered how three components of phonological, orthographic and semantic information contribute simultaneously to successful word recognition for items of high lexical quality. We described verbal and computational accounts of word recognition processes and also benchmark psycholinguistic effects that are predicted by the lexical quality hypothesis and observed in empirical studies.

This chapter will discuss the three components at a person level, outlining established measures and how variation in measures across people impact single word recognition. At this point, we introduce the focus sample of the present work: adult-learners. We briefly define adult-learners before reviewing the literature of individual differences in phonological, orthographical and semantic knowledge and skills, taking some time to locate adult-learners within that literature.

2.1 Adult-Learners

The OECD estimates that 20% of school-leavers fail to attain a skill level in reading that equips them for daily literacy tasks (Castles et al., 2018). In the 2022-2023 school year, approximately 132,000 adults (age 19+ years) enrolled in free further education (FE) courses to gain their English or Maths GCSE[1]. This number documents individuals enrolled on FE courses. As such they represent only a proportion of adults with low-literacy in the UK. When we discuss adult-learners, it is this population to which we refer. Adults who, despite sufficient opportunity and in

[1]Data retrieved from table Adult Basic Skills participation and achievements by subject and level (Aug – Apr)/English/Level 2, https://explore-education-statistics.service.gov.uk/find-statistics/further-education-and-skills#dataBlock-30e42381-4841-4f96-a3e9-9da5e35df7d9-tables, 29 July 2023.

the absence of formal diagnoses of reading difficulties, have not yet mastered a level of reading skill sufficient to be regarded as functionally literate[2].

There are reasons to assume that adult-learners differ in their skills' levels from other populations upon which word recognition research is based. We might estimate that adult-learners are approximately three years delayed in their reading skills compared to undergraduate student readers, since GCSEs are typically taken a full two years before enrolment on a higher education course. We may assume that adult-learners read at a higher skill level than children. Adult-learners are likely to have completed compulsory schooling, and so have more years of formal educational experience than most children in word recognition research. Additionally, we may assume that skills would continue to develop by mere exposure to print and language over time.

The adult-learner population is of interest for two reasons. First, the curricular activities offered to this group of learners are very similar to those offered to their younger 15-year-old peers who are studying for their English GCSE. This represents an implicit assumption that the adult-learners are merely older and do not differ in skills or reading processes, irrespective of their prior experience and low exam attainment. We can test this assumption by aligning adult-learner reading skills with a population of younger readers' skills.

Second, most theories and models of reading are built around research findings based on samples of mature, skilled readers (undergraduate students mainly), younger learners or those with atypical reading behaviours. Recently, Wild et al. (2022) demonstrated conclusively that findings from reading studies involving undergraduate student readers are not representative of the reading behaviour in the wider, general population. If non-representative samples form the basis of cognitive models, the generalisability of effects and predictions from the models are unstable (Yarkoni, 2022).

We do not know how the adult-learner sample would be qualitatively or

[2]It is important to note at the outset that this population of adult-learners is not the same population described as illiterate in Morais et al. (1986), Morais et al. (1979), and Bertelson et al. (1989).

quantitatively different from skilled adult readers. A focus on adult-learners and how they may vary in using orthographical, phonological or semantic information may contribute to the literature, informing theory or model development that is based upon behaviours as observed in a more inclusive and representative sample of the general population.

In the following, we use the components of the lexical quality hypothesis (section 1.1) to structure our discussion of individual differences. In each part of phonology, orthography and semantics, we will address the individual differences research and detail the relevant accounts of person-level variation that have been observed in the literature. We first describe how a component skill is believed to develop, describe findings for typical readers and finally review research conducted with adult-learner participants.

2.2 Phonology

Phonological skill is a critical component of successful reading development. There are two major strands of research. The first strand, phonological awareness, describes an individual's knowledge of how to manipulate sounds of a language. Manipulating units of sounds is a precursor to mapping those sounds onto letters or letter clusters and discovering the alphabetic principle of the English language (Castles et al., 2018; Castles and Coltheart, 2004; Perfetti, 2011). The awareness and ability to manipulate units of sounds of a language is the basis of the self-teaching hypothesis and so is integral to a learner's independence in reading skill development (Share, 2004, 2021). Tasks such as phoneme blending, deletion, isolation or segmentation may be used to measure such skill.

The second strand of research is phonological knowledge. Phonological knowledge measures the extent to which the letters or letter clusters of an orthography are correctly associated with the sounds of a spoken language in the individual. Tests of nonword decoding are often used as measures of phonological knowledge. Since letter strings are printed, the tests involve an element of

orthographic processing, however, the use of unfamiliar nonwords that are novel to the participant places a larger emphasis on phonological knowledge as they are unlikely to be reading from memory.

2.2.1 Phonological Awareness

Phonological awareness skill is positively correlated with early word reading skills (Hulslander et al., 2010; Melby-Lervag et al., 2012; Mellard et al., 2010; Ricketts et al., 2011; Shanahan and Lonigan, 2010; Stanovich, 1986) and to later spelling development in school aged learners (Hulslander et al., 2010; Shanahan and Lonigan, 2010; Vellutino et al., 2007) and adult-learners (Braze et al., 2007; Fracasso et al., 2016; Nilssen-Nergård and Hulme, 2014; Parrila et al., 2007; Scarborough, 1998). Poor phonological awareness (and knowledge) is long established as an underlying and persistent component of developmental phonological dyslexia (Castles and Coltheart, 1993; Kwok and Ellis, 2014; Melby-Lervag et al., 2012; Snowling and Melby-Lervag, 2016).

Phonological awareness (PA) is reported as generally low in adult-learners (Bakhtiari et al., 2015; Jimenez et al., 2008; MacArthur et al., 2010; Nanda et al., 2010; Scarborough, 1998). In a meta-analysis of several components for reading skill, Tighe and Schatschneider (2016) recovered a moderate correlation of $r = .34$ between PA and reading comprehension. It was one of the weakest relationships compared to correlations for other reading-related skills with reading comprehension.

Adult-learners tend to demonstrate weaker PA skills than their peers and also developing readers. Greenberg et al. (1997) reported their sample of 72 adult-learners as having phonological processing skills equivalent to 11-year-old learners. Thompkins and Binder (2003) found that adult-learners with an average of nine years education performed worse on a phoneme recognition task than younger learners of 1.5 years of education.

Variation in PA skills may discriminate distinct subgroups within the adult-learner population (Braze et al., 2007; Mellard and Patterson, 2008; Mellard

et al., 2016, 2012b, 2010). Mellard et al. (2012b) tested a sample of 335 16-25-year-old, low-literacy adults and identified four subgroups according to the number of words read per minute and the number of word errors. While the four subgroups differed significantly from each other on PA scores, the pattern of relative strengths across the tests for PA was the same with the weakest being the application of phonological skills for the learning of new words.

Similarly, Bone et al. (2002) constructed two groups of adult-learners, one with and one without discrepancies between predicted and actual reading achievement scores. Both groups showed deficits on phonological awareness tasks compared to a control sample of typically-reading college students. Narrative accounts report that adult-learners reporting a history of reading difficulties are observed to be approximately two years lower than adult-learners with no reported history of reading difficulties on PA skills (Hock, 2012; Mellard and Patterson, 2008).

2.2.2 Phonological Knowledge

Studies show that phonological knowledge (PK) is reliant upon the properties of word-frequency and consistency. Treiman et al. (1990) found that for young and skilled mature readers, nonword reading accuracy was better for nonwords that contained high frequency rather than low frequency spelling-sound letter patterns. Brown and Deavers (1999) conducted a similar experiment but manipulated regularity rather than frequency across the nonwords. Both studies found that words with less frequently occurring spelling-sound patterns were more likely to be read by applying grapheme-phoneme correspondence rules.

This reliance on frequency is observed in readers of low PK. In Treiman et al. (1990), third grade, poor readers were better at reading nonwords with high frequency patterns of letters than beginning first grade readers, however they were worse than the first graders on the nonwords of low frequency construction. In the same study, typically-reading students produced more real word substitutions on the high frequency patterned or regular nonwords than the younger readers, suggestive of a

strategy of reading by analogy.

Taken together, the Treiman et al. (1990) and Brown and Deavers (1999) studies may suggest that readers of all ages and experience are more likely to read high frequency patterns, and consistently spelled nonwords more accurately or by analogy than nonwords of low frequency patterns or inconsistent spelling. Less-experienced and less-skilled readers revert to explicit, serial decoding to approximate a pronunciation for an unfamiliar letter string (Brown and Deavers, 1999). A reading by analogy strategy may not be as available to younger readers by dint of vocabulary knowledge, compared to that of older student readers. Further, readers of greater experience (and greater vocabulary) may switch between different strategies of reading as the context demands.

Nonword reading has been tested in students with and without dyslexia. Kwok and Ellis (2014) found no difference in accuracy levels between the two groups. Students with dyslexia were slower to read the nonwords than students without dyslexia. As item length increased, both groups showed a length effect but dyslexic students showed a larger length effect than students without dyslexia. Two conclusions may be drawn from this: serial decoding was present for both sets of readers for initial exposures to the nonwords, evidenced by the presence of length effects. Also, phonological knowledge in mature readers with dyslexia may be complete however the efficiency of its application remains slower compared to readers without dyslexia.

Adult-learners may appear similar to younger, typically developing readers in nonword reading performance. Greenberg et al. (1997), Mellard et al. (2010), Nanda et al. (2010) and Sabatini et al. (2010) found that adult-learners had lower nonword reading skills than 4th grade readers (10-11 years). Tighe and Schatschneider (2016) found a similar strength of relationship between nonword reading and reading comprehension between adult-learners and that reported by the National Early Literacy Panel (Shanahan and Lonigan, 2010) for younger readers (adult $r = .42$; younger reader $r = .44$). Although similar in measured strength, this does not equate with similar reading behaviour or influences.

Binder et al. (2011) tested adult-learners on phonological knowledge tasks.

While typically-reading children generally reach ceiling level of performance at the end of third grade, adult-learners did not reach ceiling levels of performance on the same task until eighth grade level. This suggests that an adult-learner accrues phonological knowledge at a much slower rate than that of typical readers. In support of this, Mellard et al. (2012b) found that the 'fast and accurate group', the best of their four constructed groups in their sample of adult-learners, who showed near average PA and PK scores, contained the older members of the group (mean age of 20.3 years).

Evidence of PA and PK knowledge in the adult-learner population clearly demonstrates a slower acquisition and more variable application of skills when completing tasks compared to typical readers.

2.3 Orthography

The lexical quality hypothesis predicts that for words of high lexical quality, orthographic information alone is sufficient for efficient and accurate word recognition. For words of lower lexical quality, orthographic information is necessary but not sufficient for accurate word recognition (Perfetti and Hart, 2002). Phonological and semantic information may contribute more to word recognition processes for words of low lexical quality.

Paired-associative methods for learning of letter-sound relationships forms the bedrock of early reading pedagogy in the UK education system (DfE, 2023). After a short time, the quantity of knowledge learned approximates an "initial set" by which a developing reader can initiate the self-teaching process via phonological recoding through practice (Share, 1995, p. 156). In this way, over several years, learners are systematically exposed to the spelling-sound patterns of the English orthography to develop fluent orthographic knowledge and learning.

In the absence of fluent orthographic knowledge, a compensatory strategy for learning of novel words may be to continue with paired-associative learning. Unfamiliar words are then recognised from memory as whole words (Castles et al., 2018). In the long term, this strategy is too costly given the estimated vocabulary of a

skilled adult reader (approx 40,000 words - Brysbaert, 2019). Additionally, whole word reading is believed to reduce the opportunity for learning of sublexical spelling-sound patterns, thereby truncating the development of orthographic knowledge.

An alternate strategy is reading by analogy. Parts of known words are recognised in unknown, novel words and applied to the recognition process (Glushko, 1979). This strategy is also costly though not as detrimental to further development of orthographic knowledge as memorising whole words. The transfer of partial information from known to novel words still provides an opportunity for activation of sublexical components of a word. Consequently, growth of orthographic learning and knowledge can continue, albeit more slowly.

The success of a reading by analogy strategy may depend upon the size of an individual's vocabulary (Brown and Deavers, 1999). Readers with lower receptive vocabularies may have fewer examples by which to match parts of words for decoding. Yurovsky et al. (2014) showed that greater exposure of word types and diverse reading contexts is needed for reading by analogy to result in robust orthographic learning. Consequently, reading by analogy may not be as useful a strategy for adult-learners.

2.3.1 Word Reading

Variation in orthographic learning is often indexed by scores on tests of word reading. Strong skill in word reading is represented by fast and accurate responses to single words. Multiple studies suggest that orthographic learning of adult-learners is truncated around the grade 5 level (age 11 years), while being relatively stronger than their phonological skills (Greenberg et al., 1997; Mellard et al., 2010; Tighe and Schatschneider, 2016). Greenberg et al. (1997) demonstrated that adult-learners' orthographic skills were stronger than grade 3 readers but weaker than grade 5 readers. Mellard et al. (2012a) concluded that word reading skill has "complete dominance" status over six other predictors, including vocabulary, rapid letter naming and nonword reading.

Rather than a strong correlation between orthographical and phonological

information as observed in skilled readers (Perfetti and Hart, 2002), Greenberg et al. (1997) observed weak correlations between orthographic and phonological task measurements in adult-learners. Weak orthographical-phonological correlations suggest that the orthographical and phonological information have not yet integrated. They concluded that rather than reading words, adult-learners see words, retrieving familiar words from memory rather than applying phonological decoding strategies. This is problematic as seeing a word may not induce phonological recoding to the same extent as it may for a typical reader who is reading the same word. The capacity to develop orthographic knowledge and learning development within the exposure may be reduced. Under these conditions, a single reading episode is unlikely to render the same benefits for an adult-learner compared to a skilled reader. Orthographic knowledge development is slowed and a word will take longer to attain high lexical quality.

2.3.2 Spelling Skill

As a measure of precision, spelling dictation tasks measure both orthographic learning (each test item) and orthographic knowledge (the sublexical letter clusters across all items, Protopapas et al., 2017; Ricketts et al., 2009, 2011). Consistently high spelling scores can therefore reflect many word representations of high lexical quality within a person (Andrews et al., 2020; Perfetti and Hart, 2002) and also a robust representation of orthographical knowledge.

Adelman et al. (2014) suggested that spelling could be a key source of variance between people who are otherwise competent readers since spelling skill in the general population is highly variable. Andrews and Lo (2012) reported spelling scores between the ranges of 4 – 20 words in a spelling assessment of 20 items in a sample of 97 student readers. Masterson et al. (2007) showed an equivalent spread of scores in a sample of 40 students and an accuracy range of 9 – 30 words in an assessment of 30 items.

Low spelling skill in the context of high levels of reading skill has been

hypothesised to arise from partial processing of words. Essentially, good readers may recognise a word without fully processing the sublexical components. This partial processing results in information for spelling being acquired less effectively and over time, this results in sub-optimal spelling knowledge (Andrews et al., 2020; Masterson et al., 2007).

In contrast, explanations of low spelling skill amongst less-skilled readers have been suggested as lower vocabulary scores or less reading practice. Lower vocabulary scores presumably are a proxy for fewer word representations with the same letter patterns, consequently a lower exposure to spelling-sound patterns. Less-skilled spellers are more error prone across consistently and inconsistently spelled words, and show a wider variety of spelling errors than more-skilled spellers (Masterson et al., 2007). Poor spellers are also more likely to incorrectly classify words as nonwords and nonwords as words in lexical decision tasks.

Martin-Chang et al. (2014) measured standard spelling in the traditional correct / incorrect sense but also the variability of a person's spelling errors over repeated assessments. They found those words that were incorrectly but consistently (mis)spelled by a participant were named faster than those inaccurately spelled words that varied in the type of spelling errors. The faster reaction times for the consistently-incorrectly spelled word suggests that an incorrect spelling, when believed to be correct, may still have high lexical quality within an individual, relative to other words in their vocabulary.

Beidas et al. (2013) found a strong correlation between decoding and spelling skill in their adult-learner sample. This strong correlation with decoding skill makes intuitive sense. Decoding involves processing spelling to sound; spelling involves encoding of sound to spelling. Swanson et al. (2003) found that word learning was best predicted by spelling and nonword reading over and above nine other person-level skills.

Partial information does appear to be used in spellings of adult-learners, however, it is related to language experience rather than reading experience. Treiman (2018) asked skilled adults to rate spellings by adult-learners and pre-instruction

children. The adult-learner spellings were rated as more plausible than the child spellings. Treiman (2018) inferred that adult-learners applied partial phonological knowledge from their greater language experience, having a larger repertoire of *sound-spelling* examples than young children given their older age. This finding could also imply that the orthographic knowledge of an adult-learner may be sufficient for partial recognition of words for reading, supporting the finding in Swanson et al. (2003) that spelling may contribute to word recognition for this sample.

On average, adult-learner spelling skills are often weaker than those of their reading age match peers (Greenberg et al., 1997) and adult peers (Beidas et al., 2013; Eme et al., 2014). Taken together with the observation of low word reading skills being truncated at approximately 11 years of age, the pivotal source of information and knowledge upon which words of high lexical quality depend, seems to be under-developed in the adult-learner population.

2.3.3 Reading Fluency

Fluent reading suggests both speed and accuracy. It implies automaticity of word recognition. It is important for the release of cognitive and attentional resources for higher order processing such as construction of meaning at the text level (Perfetti and Stafura, 2014). Conversely, reduced fluency impinges on the capacity to understand text level meaning (Stanovich, 1986). Fluency will vary under different reading conditions. Even the most skilled reader will slow their reading to decode and understand a difficult passage (Bell and Perfetti, 1994).

Fluency in the present context may additionally imply that the three sources of orthographical, phonological and semantic information are integrated, coherent and simultaneously available for word recognition. Dys-fluent reading or recognition then implies that the sources of information are not integrated.

Independent of word or nonword reading skill, general processing speed can be a contributory factor to fluent reading. A measure of general processing speed is the rapid automised naming (RAN) task. Across four versions of the task, individuals

are asked to name a small set of highly familiar items (letters, digits, objects or colours) that are randomly arranged across multiple rows on one page, as quickly and accurately as possible. A longer time to complete naming the items indicates slower processing speed of the individual.

The naming of objects and colours is considered a purer measure of processing speed than naming letters and digits since they are not confounded with orthographic knowledge (Kirby et al., 2010). Object and colour naming is less automatic, however, producing longer task completion times, on average, since labels for objects and colours vary between people (Beidas et al., 2013; Cattell, 1886; Meyer et al., 1998b; Sabatini, 2002).

After controlling for skill and experience, there is variation between people as to their speed of processing. Some people process all types of information faster than others (Seidenberg, 1985). This tends to be reliable across separate testing sessions. Yap et al. (2012) demonstrated that a large sample of university students were highly reliable in their reaction time profiles across two testing sessions.

Persistent low fluency can impact reading skill development. Kirby et al. (2010) suggested that in the context of low fluency, the processing of adjacent letters in a letter string may remain independent of each other instead of adjacent letters being processed in parallel as a sublexical unit. This diminishes the opportunity for strong associations to form between letters in a word. If associations between letters are not formed, the statistical distribution for those groups of letters will be slower to develop or be missing from a person's orthographical knowledge.

Processing speed is often measured in beginning readers (Meyer et al., 1998a,b; Scarborough, 1998). In a meta-analysis of 35 studies, Swanson et al. (2003) found that RAN showed stronger correlations for average- ($r = .42$) and more-skilled ($r = .40$) developing readers than less-skilled readers ($r = .22$). This difference may suggest that from a very early age, young learners who go onto to show lower reading skills demonstrate less integration of information sources than average- and higher-skilled readers.

Kirby et al. (2010) describes RAN as showing a curvilinear relationship with

typical reading skill development. In the early years, there is a strong correlation which attenuates as other reading-related skills come on line. In support of this observation, stable effect sizes for RAN (and PA) across the ages of 4 – 10 years have been observed in longitudinal studies in typical readers (Åvall et al., 2019). Yet for less-skilled readers, the predictive age range for RAN is prolonged, in a similar way to phonological skill measurement. Meyer et al. (1998a) observed that third grade RAN (9 years) predicted fifth (11 years) and eighth grade (14 years) reading for children reading at or below the 10th percentile in skills.

In adult-learners, RAN scores were the second most important predictor amongst seven for adult-learners (Mellard et al., 2012a). From 10 studies, Tighe and Schatschneider (2016) estimated a correlation of $r = .53$ between reading fluency and reading comprehension (Mellard et al., 2010; Nanda et al., 2010). Both found that the best fitting path model included an independent predictor for processing speed. Similar to Hulslander et al. (2010), RAN measures indirectly predicted reading comprehension through its direct relationship with word and nonword reading.

The independence of fluency as a predictor plus the strong correlation estimated in Tighe and Schatschneider (2016) in adult-learner models of reading, rather than the relationship attenuating as Kirby et al. (2010) reported, suggests that adult-learners have yet to achieve fast word reading and for it to be integrated with accurate word recognition (Beidas et al., 2013; Ben-Dror et al., 1991; Bruck, 1990). This is a further observation suggesting that critical skills for reading show a lack of or weak integration in adult-learners.

2.4 Semantics

While semantic information has a role in single word recognition, as a multi-dimensional construct it shares a much less predictable relationship with orthography than phonology (Frost, 1998; Steyvers and Tenenbaumb, 2005). Also, because lexical quality is a property of a word and not the person, it is likely that words within a person will vary as to their lexical quality. This between word

variation in lexical quality intuitively suggests that semantics has a role to play for the recognition of words (Andrews and Hersch, 2010; Perfetti and Hart, 2002).

A convergent line of evidence comes from fMRI data exploring how skilled readers' brains activate for words that differ for phonological information (consistency) and semantic information (imageability, Graves et al., 2014). Analyses demonstrated similar pathways of brain activation across individuals under consistency conditions but highly variable pathways of brain activation across individuals for imageability conditions. Differences for use of semantic information are much more variable than use of phonological information across individuals. This variability is symptomatic of the less predictable relationship between orthography and semantic information.

Evidence suggests that semantic information supports recognition where the orthographical-phonological information is not so strongly correlated, such as in low frequency or inconsistently spelled words, or in individuals of low phonological skill. Strain and Herdman (1999) demonstrated that imageability, as a proxy measure for semantic information, supported exception word reading for readers that differed in their phonological skills. Effects were present across exception words of high and low imageability for readers of low phonological skill but only for exception words of low imageability for readers of high phonological skill. Strain and Herdman (1999) interpreted this as semantic information compensating for weak phonological representations in word naming.

Woollams et al. (2016) measured individual consistency effects in a sample of skilled readers and created low and high semantic reliance (SR) groups based on the smallest and largest consistency effects observed in naming low imageability words, respectively. The high SR group demonstrated lower phonological skills as a function of nonword naming and rime production tasks, compared to the low SR group. Differences were located in the naming of inconsistent words, with no differences across error rates in consistent words. High SR readers were also slower in their reaction time measures across tasks compared to the low SR readers.

2.4.1 Vocabulary

Semantic knowledge in individual people is captured by vocabulary measures. Adelman et al. (2014) suggested that a measure of size of vocabulary could be interpreted as a person-level index of lexical quality. Participants with higher vocabulary scores often name words faster, are more accurate and show smaller psycholinguistic variable effects than participants with lower vocabulary scores (Yap et al., 2012). Strong vocabulary scores predict faster and more accurate recognition of words with inconsistent spellings (Steacy et al., 2017b; Ziegler and Goswami, 2005) and Bell and Perfetti (1994) found that skilled readers with higher vocabulary scores were better at nonword reading than skilled readers with lower vocabulary scores.

Katz et al. (2012) use vocabulary measures as their proxy for reading experience. Many studies who recruit skilled readers who vary in age find superior vocabulary skills in the older participants (Allen et al., 2002, 1995; Balota and Ferraro, 1996; Ratcliff et al., 2010; Spieler and Balota, 2000). Keuleers and Balota (2015) suggested that with continued exposure to diverse texts, vocabulary knowledge will continue to grow.

Yap et al. (2009) observed that skilled readers with lower vocabulary scores relied upon semantically related word primes to a greater extent than students with higher vocabulary scores in a lexical decision task. Andrews and Lo (2013) confirmed these findings. They categorised skilled adult readers into semantic and orthographic types of readers. A semantic reader displayed high vocabulary with low spelling skills and an orthographic reader showed low vocabulary with high spelling skills. The groups were dissociated by the effectiveness of the types of semantic primes and, overall, the semantic reader was slower than the orthographic reader to recognise words (cf. Woollams et al., 2016).

Intuitively, given the older age of adult-learners, one might hypothesise that on average, they have larger receptive vocabularies that may support word recognition, and that this may confer an advantage for word recognition over other learners of the same word reading skill. A narrative review by Bakhtiari et al. (2015)

found that vocabulary knowledge was important to word reading in adult-learners but of low skill level. Sabatini et al. (2010) found that expressive vocabulary skills in adult-learners of seventh grade level were only slightly higher than their reading skills.

Yet adult-learners do not seem able to capitalise on this relatively strong resource for reading gains. Mellard et al. (2010) showed that greater experience of spoken language in adult-learners did not boost the decoding and comprehension skills relative to those of younger typical readers. Greenberg et al. (1997) and Nanda et al. (2010) found that any advantage of oral vocabulary knowledge in adult-learners was no longer present when compared with fifth grade reading performance.

Braze et al. (2007) suggested that this may be because the familiar vocabulary originates from speaking and listening experience rather than print experience. While exposure to words in print yields the orthographic and the phonological code, exposure to the spoken form of words yields only the phonological code. Relative to print exposure, spoken language represents an impoverished source of information with diminished opportunities for impact on reading development over time.

Several studies suggest a close relationship between vocabulary knowledge and phonological knowledge in adult-learners. Hall et al. (2014) found that expressive vocabulary uniquely predicted exception word reading for adult-learners but not regular word reading which was significantly predicted by nonword reading skills. McKoon and Ratcliff (2016) found that their adult-learners were more reliant upon their nonword reading skills and language skills than their undergraduate reading sample.

Taken together, vocabulary knowledge appears to be a relative strength in reading-related skills for adult-learners. Further, semantic knowledge appears to work with nonword reading in effecting recognition of regular and irregular types of words. There is conjecture in the literature that the source of vocabulary knowledge in adult-learners is from spoken language rather than written language (Braze et al., 2007). If so, this may mean that semantic representations are weak. This may explain why, even though absolute vocabulary scores appear stronger than word reading scores, the development of orthographic learning and knowledge that would otherwise

be expected, is not observed (Mellard et al., 2010).

In summary, adult-learners tend to be lower on all reading-related skills, relative to their adult peers. We have some understanding that adult-learners look like younger readers. Crucially, however, study findings suggest that adult-learner word reading skills reach a ceiling level around the age of 11 years, and do not seem to progress beyond that.

Within their own set of reading-related skills, measures of semantic skills appear to be strongest, with orthographic skills next and phonological knowledge the weakest of the three sources of word reading information. Relationships between the three sources appear to be weak – weaker than those reported in the typically-reading younger samples.

Under these circumstances, the lexical quality hypothesis predicts that orthographic learning in adult-learners will be of low quality. Given the importance attached to strong correlations between orthography, phonology and semantic information, evidenced in studies with skilled readers, the ability for adult-learners to benefit from relative strengths in their skills profile seems limited, with learning rates slowed as a consequence.

Although adult-learners may resemble typically-reading younger people in their individual differences measure scores, it is not automatically clear that they read in the same way. They may operate reading strategies at the whole-word level or operate an analogical reading strategy. We know that words have psycholinguistic properties that adult-learners may use to a greater or lesser extent compared to other types of readers. For instance, they may show a difference in the N-size effect if their vocabulary is based upon spoken language experience and the shared spelling patterns that comprise a neighbourhood are weakly represented, compared to a typically-reading younger person. A comparison of psycholinguistic property effects for single word naming tasks would provide evidence for an evaluation of such a question. Yet none of the studies with adult-learner samples examined benchmark psycholinguistic effects. In fact, there are very few studies that go beyond individual difference measures of adult-learners.

There are studies that examine the relative contributions of person-level skills and item-level properties in typically-reading adult samples and these are eminently useful as a reference point. In the next chapter, we discuss several large scale studies with human participants that explore variation in reading-related skills and their impact upon psycholinguistic effects (Adelman et al., 2014; Balota et al., 2004; Davies et al., 2017; Yap et al., 2012). We also review implementations of the PDP computational models that simulate individual differences in word recognition that are observed in behavioural data for atypically-reading individuals.

3 Individual Differences in Psycholinguistic Effects

Assessing individual differences for word recognition is important as they are a significant source of variation in study data. For instance, Seidenberg and Plaut (1998) found that the correlation between the item means for an identical set of items used in Seidenberg and McClelland (1989) and Spieler and Balota (1997) was $r = .54$. If shared variance was solely due to item-level effects, then we should expect the correlation to be much higher.

As well as variation in experimental settings and equipment, differences within each sample are at play, mandating the measurement of person-level skills in conjunction with psycholinguistic variables. A greater understanding of the skills that structure individual difference variation may go some way to explain why estimated effects of psycholinguistic variables vary between studies and why interaction effects are present in some studies and not others, e.g. Seidenberg and McClelland (1989) and Spieler and Balota (1997) for frequency x consistency.

3.1 Computational Model Accounts of Individual Differences

Most accounts for individual differences in the computational modelling literature are located at the level of people with reading disorders or patients with neurological conditions (Dilkina et al., 2008; Harm and Seidenberg, 1999; Plaut, 1997; Plaut et al., 1996). If we assume that the observed behaviour from such individuals marks an extreme pole of a continuum of reading behaviour then an alternative version of a model may simulate data that accounts for milder behaviour observable across a range

of typical readers. Models can be adapted to test hypotheses from a range of experimental conditions with a range of human participants.

As a simulation of mature, skilled reading, the DRC (Coltheart et al., 2001) is not naturally designed for an individual differences approach. Both the lexical and non-lexical route are programmed to reflect perfect knowledge of words, graphemes and phonemes (Coltheart et al., 2001). The architecture of the two separate routes reflects the observed dissociations between acquired surface and phonological dyslexia reading data and turning either route off simulates individual difference reading behaviour at the level of either an acquired surface or phonological dyslexic reader.

The Self-Teaching Dual Route Model (ST-DRC, Pritchard et al., 2018) extends the DRC by including a mechanism by which unfamiliar word nodes are added to the orthographic and phonological lexicon. The assumptions of perfect knowledge from the DRC remain unchanged, however. Consequently, the ST-DRC simulates novel word learning in the context of mature, skilled reading. As such, the model's purpose does not fit our purpose and so we do not discuss it any further.

Implementations of the PDP model have sought to account for individual differences in relation to the observed behavioural patterns of atypical reading (Dilkina et al., 2008; Harm and Seidenberg, 1999; Plaut, 1997; Plaut et al., 1996). In the following, we refer to the models by the author initials and year for simplicity: Dilkina et al. (2008) as D08, Harm and Seidenberg (1999) as HS99, Plaut et al. (1996) model as PMSP96 and Plaut (1997) model as P97.

Underpinning each of these PDP accounts for individual differences is the division of labour hypothesis (Plaut et al., 1996). A division of labour for word recognition is predicated upon a strong interdependence developing between the semantic and phonological pathways as children listen to and speak their early language, prior to the onset of literacy instruction (Chang and Monaghan, 2018). Once the learning of printed text begins, words of consistent spelling-sound patterns forge strong and stable orthographical-phonological relationships and word recognition is less dependent on semantic-phonological information. For words that a reader knows, semantic information is used where the orthographical-phonological

relationship is less predictable. Then, co-activation of semantic information with the phonological information strengthens the phonological activation for the word pattern, assisting word recognition (Plaut, 1997; Plaut et al., 1996).

In the absence of a discrete layer of semantics, semantic information was approximated in the training set of PMSP96 via an additional "boost" of activation on the phonological units. The augmented signal was stronger for high frequency words. Under these circumstances, the phonological units began to specialise in consistently spelled and / or high frequency words. Semantic information became critical for recognition of inconsistent words, where phonological information was less predictable. This division of labour has also been observed in behavioural studies with adult and child readers (Steacy et al., 2017b; Strain and Herdman, 1999; Woollams et al., 2016).

Once trained, the model could be lesioned in different ways to simulate acquired and dyslexic reading behaviours. The damage to the interdependent relationship gave a computational account of the symptoms of surface and phonological dyslexia within a single route model. Lesioning the semantic pathway of the model simulated successful recognition for consistent words, high-frequency inconsistent words and nonwords but eradicated the recognition of inconsistent words. Further, Plaut et al. (1996) suggested that variation in patterns of behaviour within acquired dyslexia reading could be attributed to the strength of the interdependence between semantic and phonological information.

P97 tested the division of labour hypothesis further by varying the number of connections between units on pathways, attaching different strengths of decay on the weights and reducing the strength of the external semantic input to the phonological units. Within the model, the phonological pathway compensated for weak semantic input and produced variations in the quality of the phonological representations. This had a direct impact on recognition rates and patterns, reflective of variation within the behavioural data from human readers with acquired dyslexia.

The PMSP96 and P97 implementations describe reading behaviour across different types of atypical reading. Dilkina et al. (2008) described degraded reading behaviour observed in five patients of differing degrees of semantic dementia.

Although sharing the same diagnosis, each person displayed variable word recognition profiles. Working from the basis of a model trained for typical development, Dilkina et al. (2008) found that the site of the lesion to the model best described the individual pattern of word reading behaviour amongst the five people. Different training regimes and differing numbers of units with pathways also affected how reading behaviour declined, in line with the behavioural data of the separate patients. Both Plaut et al. (1996) and Dilkina et al. (2008) determined that the degree of interdependence between the phonological and semantic pathways, before injury or disease onset, determined some part of the behaviour of the system once damaged.

HS99 extended the remit of the PDP model into developmental dyslexic behaviour and by implication, typical developmental reading behaviour. Critical to the developmental account was a training phase for only semantics and phonology information to emulate the pre-literate language experience of a typical child.

Developmental phonological dyslexia was simulated by removing the hidden layer of units for the phonological pathway. The hidden layer's purpose is to push the activation and subsequent outputs of the phonological layer to a legal and precise representation. In the presence of hidden units, the input to the phonological layer can be less precise, which helps with generalised recognition behaviour – eminently useful for novel word recognition. In their absence, the input to the phonological layer must be much more precise for stable and accurate word recognition. Over time, the phonological units became specialised to word patterns as opposed to more general sublexical patterns, reducing the effectiveness of the model's recognition for nonwords – a pervasive symptom of phonological dyslexia.

An alternative behaviour was simulated by reducing connections between units within separate domain layers. This restrained the ability of units to learn the associations between frequently occurring adjacent units and forming sublexical chunks – i.e. the different grain sizes that are found in English. The optimal grain size under these conditions is single letters, constraining word recognition behaviour that is characterised by letter-by-letter decoding.

To simulate developmental surface dyslexia, Harm and Seidenberg (1999)

initialised the model with less training of the orthographical and phonological layers. Further, they decreased the learning rate parameter. Making the learning rate smaller impacted on the entire system's ability to capture the available learning within a cycle. A further change was to remove 80% of the hidden units between the orthographical and phonological layers. The change observed here was that exception word reading was impaired with nonword reading slightly impaired. Harm and Seidenberg (1999) interpreted this as the hidden units facilitating the system learning of larger chunks, with the activation patterns of exception words being their own individual chunk. With a reduced number of units, the resources work to an optimal design which is smaller units, i.e. single letters.

So far, what has been described are theory-driven implementations of computational models to simulate atypical word recognition behaviour observed in human participants. In this way, every lesioned model represents one possible individual. A base model without lesions represents a single typical reader. As such, it is impossible to capture the variability within a participant sample (Seidenberg and Plaut, 1998).

As computational power has increased, it has become feasible - and desirable - to implement multiple computational models to simulate multiple individuals. One such study was conducted by Adelman et al. (2014). Skilled adult readers ($n = 100$) completed a word naming task for 711 items. Each participant's data was used as an outcome variable in a multiple regression and model coefficients collected. The distribution of 100 coefficients for mutliple psycholinguistic variables demonstrated a wide range of effect sizes within a skilled-reading, adult population.

Moreover, the participants completed repeated sessions of data collection by which to measure the stability of effect sizes within individuals across time. Individual profiles of effects sizes were reliable across sessions with the only difference being that participants became slower in their responses from sessions one to three.

Adelman et al. (2014) analysed the variability of estimated effects within the participant sample for frequency, N-size, word and nonword length, exception effects, consistency and position of irregularity. They found significant differences in the size

of effects across all variables except N-size and consistency. The interpretation was that the majority of psycholinguistic predictors exert a differential impact across readers that are all assumed to be 'skilled'.

No person-level measures of individual differences were collected from the human sample. Individual differences were operationalised as the distribution of estimated effects for the psycholinguistic predictors in the human data. As such, the variation observed in the effect sizes is attributable to random sampling variation.

To complement the human data, Adelman et al. (2014) created multiple implementations of the DRC (and the CDP+ (Perry et al., 2007, not discussed here) by seeding 250,000 random model parameter sets, testing each model on the same items (minus the word dire and mould) as seen by the 100 skilled readers. Of the 250,000 models, those that made less than 60 errors were retained (n = 2,674). Estimated effects from each retained model's parameter set became predictors for each human participant's reaction time data, effectively finding the best parameter set for an individual.

Two problems occurred: an inhibitory N-size effect and under-estimation of consistency effects. This led them to an implementation of an adapted model, the DRC-FC. A critical change was the relocation of the frequency effect from the individual word nodes in the orthographic lexicon to connections between orthography and phonology. This made the most improvement on simulations for individual differences. More models were retained for the DRC-FC than in the DRC (n = 3,548). More importantly, the N-size effect was now in the correct direction, but the DRC-FC estimates for consistency remained weak, as with the original implementation.

The relevance of Adelman et al. (2014) is both the variability of effect sizes naturally occurring within a skilled reading population and also the mutability of computational models. The number of implementations to be able to represent 100 humans' reading behaviour illustrates the complexity of the reading system.

A much more modest endeavour that instantiates multiple models for learning within words was achieved by Zevin and Seidenberg (2006). They tested a PDP model's ability to capture the variation in pronunciation of nonwords. Zevin and

Seidenberg (2006) tracked the relative proportions of regular pronunciations for four different types of nonwords across 10 runs of their model. The primary argument was that nonwords of regular but inconsistent pronunciation (as defined by Glushko, 1979) are pronounced in different ways by human participants, and computational models need to be able to replicate this effect. Further, they hypothesised that variable pronunciations arose from the distributions of orthographic knowledge that were individual to a person, based upon their reading experience.

As a proxy measure for reading experience, frequency values for the words of the training set were scrambled for each run of the model. This was to approximate the hypothesis that the objective value of high frequency words may differ and be lower in subjective frequency within an individual (Perfetti, 2007). Zevin and Seidenberg (2006) tracked the inherent variability across multiple runs of the same model.

While words that have regular analogous words by which to model pronunciations were relatively stable across models, words with no regular analogous words were highly variable in their accuracy across the 10 models. This suggests that lower reading experience may be characterised by reading by analogy and produce errors across episodes for the same word.

3.2 Accounts of Individual Differences with Human Participants

There are several studies in humans that document individual difference accounts of word reading behaviour (Adelman et al., 2014; Balota et al., 2004; Davies et al., 2017; Yap et al., 2012). Each takes an explicit approach of sampling large numbers of human participants with large numbers of items, in direct contrast to smaller, more carefully controlled studies that use a factorial design. Each study uses large scale multiple regression methods to provide an account of how multiple psycholinguistic predictors behave when modeled simultaneously. Critically, each study operationalises

one or more individual difference (ID) measures and explores how psycholinguistic predictors may differ as a function of variation in the measure.

Balota et al. (2004) conducted a lexical decision and a word naming task with younger and older participants. The analyses of two data sets enabled a task and age comparison across several benchmark psycholinguistic effects, including a test of interactions to detect if psycholinguistic effects were significantly different across the age groups. Effects of frequency, length, N-size, consistency and semantic variables were estimated from item level and participant level analyses. For each participant, a multiple regression was conducted on their task outcomes. Standardised coefficients for each variable were then collected together as dependent variables in analysis of variance (ANOVA) analyses with age and task as independent variables.

Across tasks, there were reliable estimates for initial phonemes of words. This is interesting because lexical decisions do not require overt pronunciation. This supports the presence of phonological processes on apparently visual tasks where print is involved. Frost (1998) has suggested the strong phonological theory that all reading involves phonology; Share (1995) and Share (2004) places phonological recoding at the centre of the self-teaching hypothesis, written predominantly for novice readers. The observed effect in Balota et al. (2004) is tentative evidence that phonological processing is involved in processing visual stimuli where no overt pronunciation is required and extends beyond the earlier stages of reading development into mature, skilled reading practices.

Across both age groups, semantic variables were larger for lexical decision than word naming. In particular, a higher quantity of semantic connections (WordNet predictor, Miller, 1995) for a word gave faster responses across both tasks and participants. Additionally, an interaction between WordNet and age showed a larger effect for older participants than younger participants in reaction time measures. Younger adults showed greater effects of N-size and semantic variables than older adults. Semantic variable effects needs to be interpreted with caution, however, since there were statistically significant differences in vocabulary scores (older > younger) and no statistical adjustments were made for this difference.

For both age groups in word naming, the N-size effect was consistently facilitatory, showing a larger effect at low frequency than high frequency words. N-size behaved very differently in lexical decision. In young readers, N-size facilitated responses for low frequency but inhibited responses for high frequency words. In older readers, N-size was consistently inhibitory. This was further explained as being related to the generally slower responses from the older sample as N-size was consistently inhibitory for the slower younger reader also. Over the time course of slow readers, therefore, the presence of neighbour words needs to be reconciled before a response is made.

Older adults showed greater effects of objective frequency but smaller effects of subjective frequency than younger adults. Younger adults showed larger subjective frequency in lexical decision tasks. Effects of consistency were similar across age groups (c.f Adelman et al., 2014).

Older adults showed a stronger correlation between word naming and lexical decision outcome measures than younger participants. Balota et al. (2004) interpreted this as younger readers being more affected by task specific demands. They found a general slowing effect for older participants. Ratcliff et al. (2010) and Davies et al. (2017) have confirmed this effect of slowing for older participants.

Yap et al. (2012) draws data for repeated sessions of participants (n = 1,289) from the English Lexicon Project (ELP, Balota et al., 2007) and examines how psycholinguistic effects' estimates vary with individual differences in reading skill. Participants were recruited from six universities. Across the 1,289 participants (naming n = 470, lexical decision n = 819), 40,481 items were sampled with each participant contributing either 2,500 naming responses or 3,400 lexical decision responses, collected over two sessions. Additional to the item level responses, the ELP contains measures of age, years of education and vocabulary knowledge for each participant. Yap et al. (2012) operationalised reading skill as the vocabulary measure in their statistical analyses.

With measures occurring across two sessions, Yap et al. (2012) could estimate reliability within individuals and found stable reaction times across sessions for

participants. Adelman et al. (2014) found that participants generally slowed across sessions, however effects were reliable for participants. Reliability of measures across sessions increases confidence in the estimations of effect and their sizes.

Yap et al. (2012) performed a principal components analysis to mitigate collinearity between their 10 psycholinguistic predictors, retrieving principal component scores for each individual participant. The best solution was for three components: a word-structure component, a N-size component and a frequency-semantics component. The word structure component included length in letters, syllables and morphemes and measures of OLD20 and PLD20. Orthographical and phonological N loaded separately onto the second component.

Overall, readers of higher vocabulary scores were faster and more accurate than readers of lower vocabulary scores. Readers with higher vocabulary scores showed smaller effects across all three principal components for word naming and smaller effects for N-size in lexical decision, echoing Balota et al. (2004). There appeared to be no effect of differences in vocabulary for either word structure or frequency and semantic effects for lexical decision.

Davies et al. (2017) took a lifespan approach to word recognition and included children in addition to younger and older adults in their participant sample ($n = 535$, age range 8 – 83 years). They examined individual differences for age, reading skill and phonological skill in interaction with word-frequency and age-of-acquisition (AoA) for word naming and lexical decision tasks. Additional psycholinguistic predictors were also included (bigram frequency, imageability, length, N-size, regularity), ostensibly to control for their influence on the measurement of the primary variables of interest.

Davies et al. (2017) differed from the studies above by using linear-mixed-effects-models for data analyses. Rather than estimating single models per participant and pooling coefficients for further analysis, or rather than averaging across participant responses to create one mean response per item, the mixed-effect-model allows for trial level data to be modelled while exploring fixed effects of person- and item-level variables. The additional inclusion of random effects

terms estimate the variation per participant and per item per predictor. As a result, the fixed effects' estimates for primary predictors are robust to sources of within-participant and within-item variation and between sources of random sampling error attributable to sampling across multiple people and multiple items (Baayen et al., 2008).

Their primary finding was of curvilinear effects for age and skill across tasks. Word naming and lexical decision reaction times showed an independent effect of age where reaction times tended to decrease steeply from children to younger adults but slow again as the participants became elderly. Effects for AoA, imageability, concreteness and frequency also followed this pattern of diminished effects sizes from childhood to adulthood with the rate of decrease slowing into late adulthood (the interaction for frequency on lexical decision was marginally significant, however). In word naming only, there was a significant regularity x age effect showing the same curvilinear trend.

There were several significant interactions between reading skill and psycholinguistic predictors. Frequency and imageability showed significant interactions in word naming only, decreasing in size for those with higher reading skill. In both tasks, orthographic neighbourhood decreased with increased reading skill. Bigram frequency and regularity effects increased in word naming responses for readers of higher skill. Neither the interaction of reading skill x AoA nor reading skill x length were significant in either word naming or lexical decision.

Taken together, the four studies with human participants described above (Adelman et al., 2014; Balota et al., 2004; Davies et al., 2017; Yap et al., 2012) provide lines of converging evidence for the effects of individual differences on psycholinguistic variables. First, even within a skilled reading sample, psycholinguistic predictor effects can significantly vary in size between individual participants (Adelman et al., 2014; Davies et al., 2017; Yap et al., 2012). Second, across repeated sessions, reaction time (Yap et al., 2012) and patterning of effects (Adelman et al., 2014; Yap et al., 2012) are stable within participants. Third, older participants are generally slower than younger participants (Balota et al., 2004; Davies et al., 2017)

but that the assimilation of information for older adults is no different from that of younger adults (Davies et al., 2017; Yap et al., 2012). Fourth, participants who show strong skills in the individual differences measure, tend to be faster and more accurate than participants who are weaker on the measure (Adelman et al., 2014; Davies et al., 2017; Yap et al., 2012). Fifth, there is a tendency for the influence of psycholinguistic effects to diminish for human participants who show stronger skills in the individual difference being studied (Davies et al., 2017; Yap et al., 2012). Furthermore, by taking a lifespan approach, Davies et al. (2017) described a tendency for effects to change once more in late adulthood. Where, for the majority of mature, skilled readers, psycholinguistic effects may continue to decrease, the more elderly participants of Davies et al. (2017) reversed this trend, beginning to show an increase in effect sizes.

The fourth and fifth observations listed above suggest two things. The first is that scant assistance for word recognition from psycholinguistic predictors for skilled readers, in the presence of strong word frequency effects, can be inferred as support for the lexical quality hypothesis. It would suggest that for words of high lexical quality, orthographic information alone is enough for efficient word recognition. The second is that when exploring individual differences, modelling multiple interactions between person-level and item-level predictors yields both insights as to the structure of the reading system but also insights as to the stability of predictors in relation to each other.

The studies also provide divergent lines of evidence. For instance, the effects of regularity or consistency. Yap et al. (2012) omits a measure of consistency. Both Adelman et al. (2014) and Balota et al. (2004) found no differences in the size of consistency effects in their respective samples. Davies et al. (2017) confirmed this for effects in lexical decision, however age and skill related differences were detected for consistency effects in word naming. The inclusion of children in the Davies et al. (2017) sample may have amplified any such effect, if we assume that children (as young as 8 years old) have less orthographic knowledge than a mature, skilled reader.

Additionally, Davies et al. (2017) and Adelman et al. (2014) operationalise their regularity and consistency measures differently from each other while Balota

et al. (2004) trimmed their data set of all words that did not achieve 67% accuracy. Considering that when present, consistency effects are small, these differences in sample construction, operationalisation of the measures and data cleaning may explain the difference in findings.

N-size effects are also inconsistent across the studies. In word naming, Balota et al. (2004) reported only that N-size was facilitatory for both groups of young and old participants. Adelman et al. (2014) found an interaction with age such that older participants benefited less from larger neighbourhood words than younger participants. In by-items analyses, Davies et al. (2017) found that while the relationship was facilitatory, the N-size effect for younger adults was non-significant and the older adults' p value was borderline ($p = .049$).

In lexical decision, Balota et al. (2004) found a facilitatory effect for young participants and an inhibitory effect for older participants. Both interaction terms for age and N-size in Davies et al. (2017) were non-significant. Yap et al. (2012) and Davies et al. (2017) report that participants with greater skill (operationalised by vocabulary and reading skill, respectively) demonstrated smaller N-size effects in both lexical decision and word naming.

Reasons for variable effects for N-size are not clear. The relational properties of N-size may be at play. As described earlier, the N-size metric may suggest a reading by analogy strategy, the success of which may be reliant upon a reasonably sized vocabulary that contains enough words of similar patterns to assist word recognition. Only Yap et al. (2012) reports measures of vocabulary, however, so we cannot be sure. Alternatively, Yap et al. (2012) reported that of the three components, the component for N-size was less reliable than either the word structure or frequency / semantics component.

Zevin and Balota (2000) reported how particular primes were able to direct attention within a task and produce effects on reaction times. It is possible that with the large quantity of trials, weaker participants were primed by similar stimuli such that responses quickened and the N-size effect was not detected. A weaker reliability in the N-size component, and waxing and waning effects could be explained somewhat

by the different sets of items across studies and trial order within those sets.

These four studies of individual differences have tested established effects from smaller experiments in a less conservative setting. Findings have begun to further shape our understanding of the reading system and how variability in person-level skills may modulate the presence, size or direction of psycholinguistic predictor effects on word recognition.

Each of the studies of human participants describe reading behaviour for typical readers, while the computational modelling papers, in the main, focus upon accounts of disordered reading. This begs the question as to whether the accounts generalise to samples that are atypical but not disordered. It is not immediately clear that any of the results should be relevant to an adult-learner population. Like typical skilled adults, adult-learners are mature, but unlike their peers, they are relatively unskilled in their reading behaviour but not disordered in a clinical sense. Consequently, study findings for skilled adult readers may not generalise to adult-learners well and model predictions based on disordered reading may not be relevant either.

Chapter 2 detailed numerous studies that indicated that adult-learners do seem to resemble typically-reading younger participants in their person-level measures of reading-related skills. Adult-learners may use information from text in different ways, however. For instance, as a result of their age and by extension, greater language experience, adult-learners may rely, to a greater extent, on semantic information for their word recognition.

To complicate matters further, even in the context of large sample sizes and item sets, psycholinguistic effects' estimates are highly variable. Although using multiple regression methods, Balota et al. (2004) emphasises the complementarity and importance of measuring the presence and size of effects in factorial experiments. Seidenberg and Plaut (2006) reinforce this, stating that any computational models need to be tested against effects that have shown successful replication across many experiments. Nevertheless, effect sizes also differ across factorial studies.

The observation of variability of effect sizes both within participant samples

and across separate studies suggests that we conduct a meta-analysis. Collating and aggregating effects from multiple studies will estimate an effect size within a wider range of plausible effect sizes. Gathering data from studies involving child and adult participants will provide a basis for answering questions around differences between samples of participants, such as whether children show larger effects, on average, than adult samples and whether effects from one task really are larger than effects from another.

On completion, we would have a collection of effect sizes, estimated with their confidence intervals, plus an estimate for how much of the variability within the effect is due to differences between studies and not only random sampling variation (Adelman et al., 2014). Since we know very little about how adult-learners use psycholinguistic variables, such a data set would be a strong reference point against which to compare their reading behaviour in a behavioural study.

We conducted such a meta-analysis for the psycholinguistic predictors of age-of-acquisition, arousal, consistency, imageability, length, word-frequency, N-size and valence. We searched for study-level effects for child and adult samples on word naming and lexical decision reaction times and accuracy measures. The meta-analysis study findings are an openly accessible resource for use by the research community. We detail our methods and findings, with an example predictor report for word-frequency and summaries of the findings for the remaining seven psycholinguistic predictors, in the next chapter.

4 Meta-Analysis of Psycholinguistic Effects

We present a meta-analysis of effects for differences between groups in psycholinguistic effects across word naming and lexical decision tasks for reaction time and accuracy outcome measures. We conducted a systematic search of relevant articles in EBSCO, ProQuest Dissertations and Theses, and SCOPUS databases up until 2020. Additional articles were located through a search of citations in eligible articles. To be included, studies had to compare at least two groups, measuring performance in either single word naming or lexical decision tasks with a manipulation of a psycholinguistic variable(s) for reaction time and / or accuracy outcome measures. Studies with participants with clinical conditions or who were bilingual, or studies that were not available in English, were not eligible for inclusion.

To anticipate the results, 122 articles met our inclusion criteria, representing 155 studies ($n = 12705$) with 472 study-level effects across five discrete groups: adult or child samples contrasted by either age or skill. We recovered study-level effects for age-of-acquisition, arousal, consistency, word-frequency, imageability, length, neighbourhood-size and valence. Forty studies reported on both word naming and lexical decision outcomes, 73 studies reported on word recognition outcomes only and 32 studies reported only on lexical decision.

Each study was judged for risk-of-bias. We fitted multi-level random-effects or fixed-effects statistical models to return 32 global effects and 131 subgroup effects. We performed sensitivity analyses and where sufficient data and variability were present, moderator analyses were completed. Each estimate's data was evaluated and given a confidence rating.

Subgroup estimates for which we have moderate confidence are sparse ($n = 24$). They occur mainly for frequency and length estimates. For the majority of the

other predictors, there is insufficient evidence to suggest differences between groups and where data are present, estimation of differences between groups is imprecise and unreliable, lowering confidence in the subgroup effect.

Going forward, a consortial approach using matrix sampling designs for single word recognition research is suggested. Methods that increase efficiency of data collection coupled with appropriate methods of analysis will power the estimation of effects appropriately. Consequently, the precision of estimates and subsequent confidence will increase. The goal is to have robust estimates in place for each valid psycholinguistic predictor, such that each predictor's influence can be considered in the context of its peers for substantive contributions to theories of word recognition processes.

This work was supported by the ESRC. The project protocol, all data and code plus generated reports for eight predictors are shared via https://bit.ly/Meta-analysis-repository.

4.1 Introduction

The previous chapter described studies that demonstrate variation in psycholinguistic effects for word recognition within samples of skilled readers (Adelman et al., 2014; Davies et al., 2017; Yap et al., 2012) and samples of skilled readers that differ in age (adults: Balota et al., 2004; children and adults: Davies et al., 2017).

The variation in independent effect sizes within samples begs the question of the reliability of the measurement of differences for psycholinguistic effects between samples. A range of effect sizes is suggested when interaction effects are found in some studies but not others. Estimating the range of plausible effect sizes is also desirable if we want to make statistical inferences about practical differences in size of effects between tasks. We collected evidence from smaller studies to estimate summary effect sizes that can be used as a point of reference for a further study involving an adult-learner population.

Previous review articles exist that make important contributions to reading

research, however very few of them consider psycholinguistic variables as a primary focus. Brysbaert (2019) performed a meta-analysis for skilled, adult reading rates of passages. Lima et al. (1991) meta-analysed lexical decision reaction times from younger and older participants to account for general slowing in older participants' decisions. Swanson and Hsieh (2009) and Tighe and Schatschneider (2016) collate study-level effects to meta-analyse differences in reading-related skills between typically and atypically-reading adults. A meta-analysis by Laver and Burke (1993) considers semantic priming as a function of age. Reis et al. (2020) considers a measure of consistency as a measure of orthographic depth in their meta-analysis for differences between adult readers with dyslexia in reading-related skills across languages of varying orthographic depth.

In child samples, there are narrative reviews that focus on phonological awareness (Castles and Coltheart, 2004) and nonword reading (Rack et al., 1992). Meta-analyses are also available for phonological skills (Melby-Lervag et al., 2012; Swanson et al., 2003), rapid naming (Araujo et al., 2015), spelling instruction (Graham and Santangelo, 2014), oral language deficits in children at familial risk of dyslexia (Snowling and Melby-Lervag, 2016), print exposure (Mol and Bus, 2011) and nonword reading (van Ijzendoorn and Bus, 1994).

Within the field of psycholinguistics, there are narrative reviews of AoA (Juhasz, 2005), word-frequency (Allen et al., 1991; Ellis, 2002), neighbourhood size (Andrews, 1997), and for processing of emotion words (Rohr and Wentura, 2021), all describing skilled, adult readers. Metsala et al. (1998) meta-analysed study-level effects for regularity effects in readers with reading disabilities. Just as in the adult studies, very few of these reviews consider the impact of skills in interaction with psycholinguistic variables. Given our interest in the variation of effect sizes as a function of individual differences and their impact on word recognition, a meta-analysis that explores differences in effects for skilled and less-skilled, and younger and older readers, across a range of psycholinguistic predictors, is warranted.

Since so few studies have looked at *adult-learners* in relation to psycholinguistic variable effects (but see McKoon and Ratcliff, 2016), yet we have

studies that suggest greater difference from adult readers and similarity to younger readers, gathering studies that explore differences between typical and atypical readers in child and adult samples is a good place to start. Consequently, the present meta-analysis included studies that explored adult and child samples of typical and atypical skill on word recognition performance.

Our review was motivated by two sets of assumptions and a methodological requirement. Assumption one is that evidence on the modulation of psycholinguistic effects by group differences enriches theoretical accounts of the development of word recognition skill, and the ways in which reading can vary across individual differences in the population.

Assumption two is that variation in effect sizes does to some extent depend on design. Word recognition enjoys a long history. While the experimental tasks are largely unchanged, new procedures, equipment and materials are likely to have refined the measurement which may be observed in smaller or more precise effect sizes in more recent studies compared to older studies (Meehl, 1967), but we are unsure.

Aggregating multiple study-level effects and generating confidence intervals for groups will provide reference points for those groups. Understanding average effect sizes and the credible range over which they can vary may inform future research design. To this end, the data, code, figures and meta-analysis reports for eight psycholinguistic variables are openly accessible at an Open Science Framework (OSF) repository for interested researchers to view and use.

Finally, our methodological requirement: the number of psycholinguistic variables available to inform accounts of word recognition research is growing. Each claim a portion of the variance within outcome measures. Intuitively, this demands that they be modeled simultaneously. This suggests the use of larger and more complex statistical models. Such models often experience convergence problems. Bayesian inference models can be more robust to convergence problems. An integral part of Bayesian inference modelling is the prior knowledge that is entered into the model with the observed data. The summary effects from this meta-analysis will be used as strongly informative priors in Bayesian inference model analyses in the main

study (see next chapter).

Below, we explain our study rationale for focusing upon a contrast group design, many variables rather than a single variable, and our chosen tasks. We also detail our assumptions and principles for the evaluation of confidence in the study findings.

4.1.1 Contrasting Groups

Individual differences research often contrasts two or more groups along a single dimension. A common contrast is age. For instance, comparing younger and older skilled adults to explore how frequency effects vary, addresses the question of whether and how changes brought about by the typical ageing process affect cognitive reading processes (Davies et al., 2017). A suite of papers by Allen and colleagues (Allen et al., 1991, 2004, 2002, 1993) claimed that, for the lexical decision task, there was no difference between the younger and older adults for a word-frequency effect, whereas Balota and Ferraro (1996) and Balota et al. (2004) found larger effects in their older participants compared to their younger participants. In contrast, Davies et al. (2017) found a smaller frequency effect in older compared to younger adults. Later still Cohen-Shikora and Balota (2016) found no differences for a word-frequency effect between younger and older participants, concluding that, once skilled reading development is attained, the word-frequency effect remains fairly stable across the lifespan.

The second type of contrast is reading skill. A well-practised approach is to recruit a group of typical readers and a group of dyslexic readers of the same age and test them on the same items to identify variation in psycholinguistic effects. Findings from these age-matched samples may inform a model of typical reading development or describe a word recognition profile of skills for the atypical reader. For instance, Jorm (1981) found a difference between two groups when asked to name regular and exception words. The impaired readers experienced a greater difficulty with accurate exception word reading compared to regular word reading. Jorm (1981) argued that

this difference indicated greater reliance on grapheme-phoneme correspondence rules for exception word reading. In contrast, Treiman and Hirsh-Pasek (1985) and Gottardo et al. (1999) found no differences between their typical and atypical readers in the impact of consistency on word naming accuracy outcomes. The variable findings contributed to debate upon how words were represented in the cognitive reading system for verbal and computational models of word recognition.

Studies that compare groups along the skill dimension often include a further group that is matched to the atypical group's reading skill level, referred to as a reading-matched sample. Generally, this means the third group is younger in age than the atypical group, and has typical reading skills (Backman et al., 1984b). Any lack of difference in skill between the atypical and reading-matched group allows assertions to be made about reading delays or deficits with regard to the atypical readers, dependent upon the pattern of findings. As an example, Bruck (1988) and Jimenez Gonzalez and Valle (2000) both implemented a reading-match design and found no difference between older atypical readers and younger typical readers for frequency effects on word naming reaction time. Both drew the conclusion that the older readers are delayed in their skill development, and show a younger, immature reading style, thus giving confidence to a theory that skill levels will be attained over a longer period of time for the atypically-reading individual.

From only a couple of examples, it is clear that the evidence for a difference in the effect of psycholinguistic variables between groups is far from consistent. Each of the examples above claimed an effect's presence or absence depending on the significance or non-significance of a corresponding hypothesis test. It is well established that a null finding does not equate to no effect, however. Effects may be present but too small to detect in studies that are often under-powered for a significance test (Gelman and Stern, 2006).

Meta-analysis offers the opportunity to reappraise a study finding by estimating an effect size with confidence intervals rather than performing a threshold test of significance (Cumming and Finch, 2001). The presence and reliability of an effect for a predictor may be better supported when aggregated from multiple

study-level effects (Cooper et al., 2009; Shadish and Haddock, 2009). Further, the meta-analytic approach allows an estimation of the amount of variability between studies (Borenstein et al., 2017; Ellis, 2010). With a sufficient amount of variability and number of included studies, moderator analyses could also be performed that may explain heterogeneity across the included study-level effects, informing accounts of variation in effect sizes across studies.

4.1.2 Psycholinguistic Variables

There is an array of distinct psycholinguistic variables vying as candidate predictors of single word recognition reaction time (RT) and accuracy. Balota et al. (2006) list nine separate psycholinguistic variables (consistency, length, frequency, familiarity, age of acquisition (AoA), orthographic neighbourhood, phonological neighbourhood, concreteness / imageability and meaningfulness). Foreshadowing our results, we found sufficient study-level effects from multiple studies to allow the estimation of group differences in eight distinct psycholinguistic effects through meta-analysis (AoA, arousal, consistency, frequency, imageability, length, neighbourhood size (N-size), valence).

Research into the effects of psycholinguistic variables such as word-frequency, length and consistency has a long history with many published studies. In contrast, for more recently developed variables (e.g. arousal, valence), studies are sparse. However, discussing the range of variables together in the same space will allow for an evaluation of their relative influence on word recognition. Collation of findings for multiple variables will also enable an appraisal of the current quality of the data and understand the rate of missing data in the field.

4.1.3 Word Recognition Tasks

We have explored word naming and lexical decision tasks for two reasons. Firstly, we believed that these tasks, of all word recognition tasks, were the most likely to yield a

numerous sample. They are some of the earliest tasks performed in the exploration of cognitive models for word recognition (Cattell, 1886) and they are relatively cheap and easy to use compared to other methods such as eye-tracking. Secondly, the tasks inform each other in terms of shared and distinct processes (Andrews, 2012; Balota et al., 2004) which can inform theory and model building for word recognition processes.

Both tasks share the process of taking orthographical information from a printed item that stimulates the activation of representations. Dependent upon the person, item and task however, activation may be of different levels or different kinds. For instance, word naming assumes the complete identification of a word for successful naming. For lexical decision making, it may not be a requirement for a correct response that a complete identification of the specific item is made (Balota and Chumbley, 1984; Grainger and Jacobs, 1996; Seidenberg and McClelland, 1989), only that the item is more likely to be a word than not. This decision could be made with a greater input from semantic information for words as opposed to nonwords. Larger effects for semantic variables in lexical decision findings compared to word naming findings is a well-established pattern (Balota et al., 2004).

By estimating effects for a range of psycholinguistic predictors for both tasks, we can build a picture of the extent of any substantive differences between the kinds or levels of information that each task uses across a body of evidence rather than at a single study-level. This may strengthen conviction in certain processes and discern between verbal theories and computational models of word recognition.

4.1.4 Risk-of-bias and Confidence

A critical component of the meta-analytic process is evaluating the quality of the evidence and information for both the study-level effects and the summary effects produced by the meta-analysis. Study-level effects undergo a risk-of-bias (RoB) evaluation; summary effects are reported with a confidence rating (Higgins et al., 2011; Schunemann et al., 2011).

Evaluating the presence of systematic sources of bias for a study-level effect through a RoB process may explain why a study is detected as an outlier in a sensitivity analysis. RoB judgements can be a predictor in moderator analyses to explain high heterogeneity values in any summary effect. Awarding a confidence rating to a summary effect may help its interpretation and in turn, help an end-user interpret it for use in future studies.

While detection of RoB within a study-level effect is important, we need to be aware of downstream effects on a summary effect. Systematic sources of bias in study-level effects are propagated within aggregated effects, leading to high levels of heterogeneity (Higgins et al., 2011). For instance, each study involves instructions to participants. A set of instructions requiring children to refrain from speaking until they are sure of the answer may systematically induce longer reaction times for poorer readers than good readers in one study than a study with no such instructions, producing different effect sizes.

Propagated error within a summary effect may also arise from lack of statistical power. Cohen (1988) and Sedlmeier and Gigerenzer (1989) suggest that power in some areas of psychological research is low, where power is defined as the probability of detecting an effect when an effect is present. Recently, Vasishth and Nicenboim (2016) simulated 1000 data sets of a two group condition (2 x 20 participants) for reaction times to 16 items. The known, group difference in the original study was 4 ms. The estimated power to detect the effect was 9% with 11% of the simulations showing an opposite direction of effect to the true effect. This illustrates the unreliability that can be present in small sample sizes.

Brysbaert and Stevens (2018) recommend that to estimate an effect size of $d = 0.4$, a sample needs at least 40 people x 40 items for a sufficiently powered study to yield a robust main effect. Power to detect effects arising from interactions mandate even larger sample sizes. Estimating an interaction that is half the size of main effect needs 16 times the sample size (Gelman, 2018).

Additional to the possibility of reversals of directions of effect, small samples may inflate the size of study-level effects (Button et al., 2013). The Open Science

Collaboration (2015) found that 100 replicated effects were approximately half the magnitude of original study effects. Vasishth and Nicenboim (2016), based upon two groups of 20 participants and responses to 16 items, estimated a mean exaggeration rate of 5.08, i.e., the mean effect size across models was inflated by approximately five times the size of the known true effect.

Inflation of study-level effect sizes can also arise due to the choice of data analysis method. Brysbaert and Stevens (2018) mirrored an analysis-of-variance (ANOVA) analysis workflow of a pre-existing data set with a linear-mixed-effect model and recovered an effect size almost one tenth of the reported estimate of the original study effect.

Bias in study-level effects is inevitably carried forward into the aggregated summary effect. We can evaluate the quality of evidence for a summary effect and indicate how close we believe the size of the summary effect is to the true effect through the application of a confidence rating. The Grading of Recommendations, Assessment, Development and Evaluations process (GRADE, Schunemann et al., 2011) recommends evaluation of evidence across five domains:

- *Imprecision* focuses upon the width of the confidence interval as a measure of uncertainty around the size of the effect.
- *Indirectness* refers to the relevance of the samples and outcomes within the meta-analysis to the population to which the study wishes to generalise.
- *Inconsistency* considers the level of residual unexplained heterogeneity of the finding after sensitivity and moderator analyses are completed.
- *Publication bias* takes into account whether missing studies are likely, as indicated by the outcomes of the Egger's Test or the Rank Correlation Test.
- *RoB* at the summary effect level. A composite indicator of RoB taken from the study-level effects, describes whether limitations in study design or execution feed forward as a source of bias for the summary effect.

Our research aim is to estimate effect sizes that describe the difference between groups in the effect sizes for various psycholinguistic variables. Where a

difference-of-differences effect is reliable, we can assume that one group has a larger effect size than another. Where the finding is unreliable, we can assume that the effect of the psycholinguistic variable is equivalent across groups. We detail our methods for searching, collating, estimating and evaluating our findings next. All code, data and reports for the eight variables are available for use at the OSF repository.

4.2 Method

The meta-analysis is guided by PRISMA guidelines (Preferred Reporting Items for Systematic reviews and Meta-Analyses, Moher et al., 2009; Page et al., 2021). The PRISMA-P document (Moher et al., 2015), including agreed revisions and updates is available at OSF repository. We searched EbscoHost, Scopus and ProQuest Dissertations and Theses on three separate occasions over the space of four years (November 2016 - March 2020). We used a systematic search string constructed with the help of an information specialist. An example of one of the most recent search strings is included in Appendix A.

4.2.1 Eligibility Criteria

4.2.1.1 *Article Eligibility Criteria*

Articles written in English are included from peer reviewed journals and unpublished theses and dissertations. Only studies using languages of alphabetic scripts were included. Studies within the field of L2 learning, or where one of the groups was of participants with a clinical condition (e.g. dementia or aphasia) were excluded. Each of these conditions would add a layer of complexity that was out of scope of the current research question.

4.2.1.2 *Study Eligibility Criteria*

Individual studies were considered eligible if the design involved a) two or more contrasting participants groups with the contrast being along dimensions of age or reading skill; b) where participants were asked to recognise words through a visual word naming and / or lexical decision task; c) the items were selected to vary across one or more psycholinguistic variables; d) reaction time and / or accuracy was the outcome measure.

In eligible articles that contained multiple studies, the first eligible study was always included. Subsequent studies were also eligible for inclusion if they involved a change of participants, items or both. We adjusted for the potential of greater correlation between study-level effects reported in the same publication by employing a nested-model analysis (see analytical model section below).

4.2.2 Study Selection

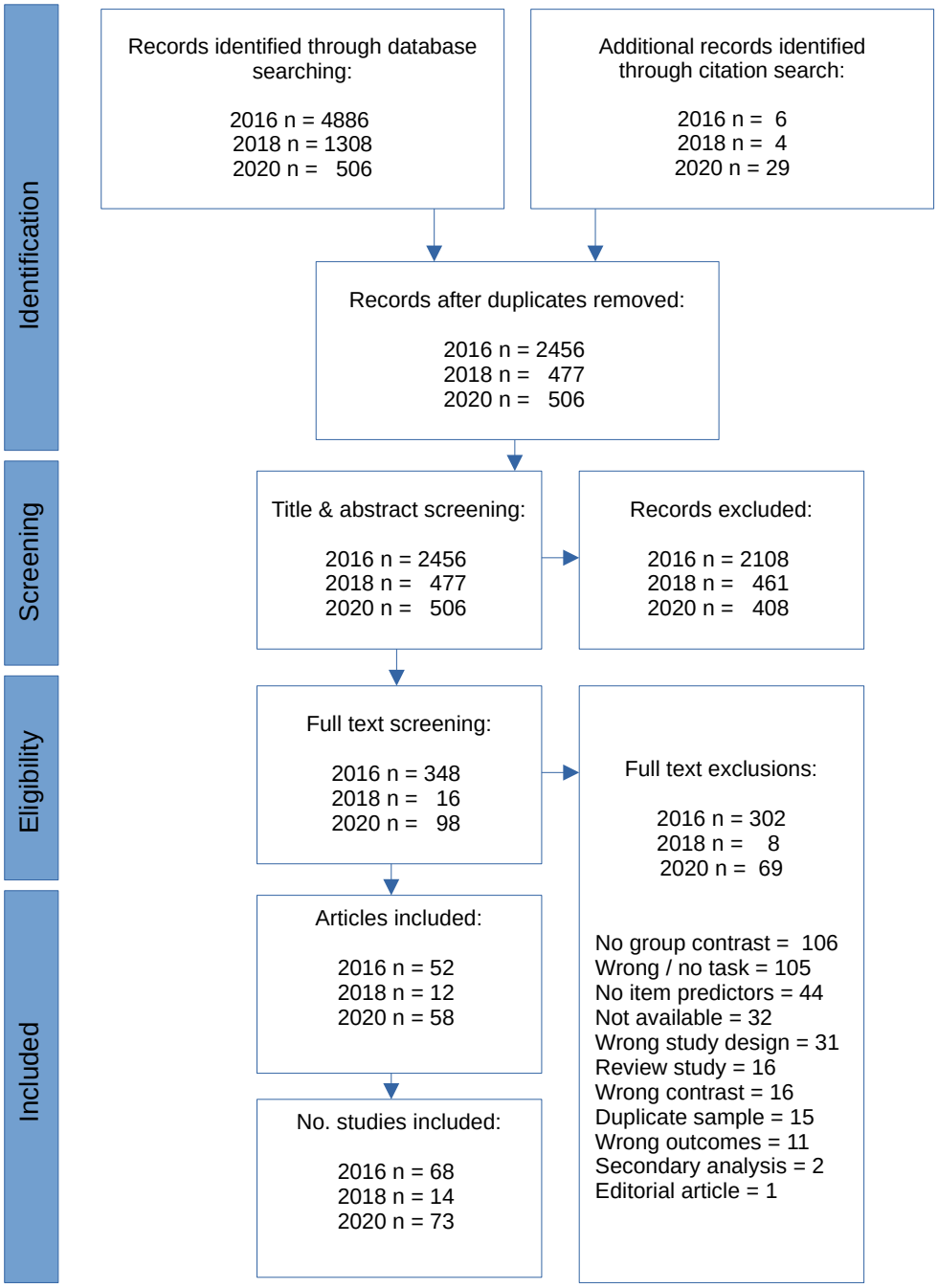
A flow diagram describes the study selection process in Figure 4.1. The author conducted the screening of the literature search records, coding and data extraction for all included studies at each time point. A second researcher reviewed a random sample of 20% of the included studies for data extraction and RoB reliability checks. Disagreements arising from this process were resolved by discussion. All code and data are available for inspection or use at the OSF repository.

4.2.3 Data Extraction

We coded contrasts for studies in the following way. An ‘experience’ contrast is where two groups of participants differ in age with typical reading development for that age, i.e. younger and older participant scores are compared. An ‘ability’ contrast is where two groups of participants are of the same age but one has atypical reading skill for that age. An ‘age’ contrast compares two groups of participants that have the same level of reading skill but differ in age (usually the older group is demonstrating

Figure 4.1

Flow Diagram for Systematic Search Returns



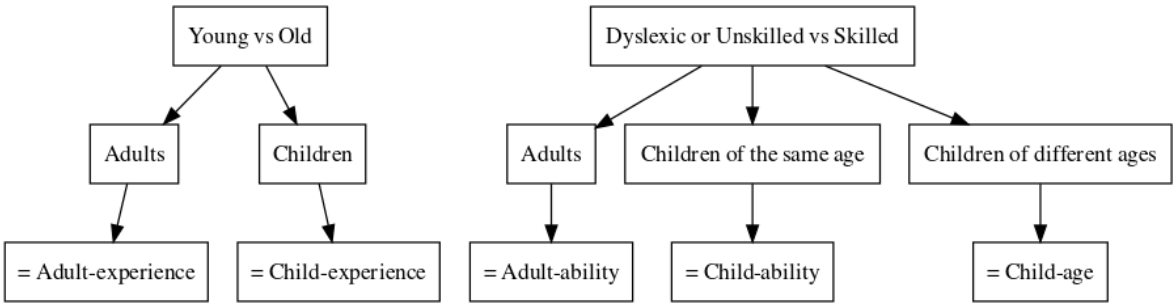
atypical reading skills). We chose to use these labels to keep the focus on the level of contrast (see Figure 4.2).

Occasionally, some studies used multiple groups for one contrast, e.g. five separate age groups of typical readers. In this case a subset of the sample was collected, matched as closely as possible to how the contrast was reflected across the rest of the extracted data.

When a paper using an age *and* ability contrast (i.e. three groups) reported a single omnibus test statistic for an effect, we merged two of the groups, respecting the key group contrast. For instance, if an omnibus test statistic was presented for an ability contrast, we kept the older, typical readers as one group and merged the younger typical readers with the older, atypical readers because they were matched on reading skill level, both being lower in skill than the older, typical reading group.

Figure 4.2

Flowchart Showing How the Participants were Subdivided to Capture the Range of Contrasts Available in Adult and Child Participant Studies.



Similarly, where more than two levels of a psycholinguistic variable were employed, for example, items of four different lengths, and the article reported means and standard deviations for each level, we extracted the levels that most closely reflected the levels used by other studies. For instance, in a study that operationalised length of words as 3-, 4-, 5-, 6-, 7-, and 8-letters, we extracted data for words of 3- and 6-letter lengths as these psycholinguistic variable levels closely resemble a common operationalisation of length amongst studies.

Where studies from unpublished theses were eligible and the same data was

included in a published article at a later date, the published article was taken as the source of data. If other studies were available in the thesis / dissertation that were eligible for inclusion, these data were also extracted and the thesis citation entered as the data source.

We approached six authors of eligible studies asking that they share the study data and code when the study findings were unclear. Four authors replied with data but not code. We followed the methods in the original study to estimate means and standard deviations from the data as a proxy for the published study-level effect. The studies for which we obtained data or data and code are marked in the list of studies given in Appendix B.

Information was extracted for articles (authors, year, publication and country of first author), study design (design[1], setting, sample type, number of participants and number of items), participant sample construction (type of contrast, measure used for contrast, test scores / ages, number of items), task level data (word naming or lexical decision, outcome measure, psycholinguistic variables and effects tested) and reported statistics (means and standard deviations for reaction time, proportions or totals correct / incorrect for accuracy or summary values for statistical tests, directions of effect and status of results).

4.2.3.1 *Missing Data*

Many studies had missing data for descriptive statistics of the study experimental conditions. Therefore, we collected a range of statistics (e.g., F-ratio, t-value, regression coefficients) by which to estimate the study-level effects.

Sometimes, a non-significant result was verbally reported in the text without supporting statistical information. At other times, entire conditions reported in a method section were missing from a results section. Chan et al. (2004, as cited in Pigott, 2012) found that results were twice as likely to be missing if they were non-significant results.

[1]The research question implicitly assumes a quasi-experimental study design because of the nature of the contrasts at the person-level.

For each missing datum, a p value of .1 was imputed into the datasheet. We created an additional variable indicating the type of missingness. Verbal reports of non-significance were coded as 'I' for 'inferred'; absent verbal reports for conditions were coded as 'M' for missing.

4.2.3.2 *RoB for Study-Level Effects*

RoB has six domains: selection, performance, detection, reporting, attrition and other sources of bias (Higgins et al., 2011). We reviewed articles for reports of practices covering the six domains and ascribed a level of bias to each domain, either 'high', 'low' or 'unclear'. Finally, we gave each study-level effect an overall level of bias.

A judgement of high RoB in any one of the six domains resulted in an award of overall 'high' RoB. Where all domains are adjudicated as low, a 'low' value was ascribed. Where any domain is listed as 'unclear' but the majority of domains are otherwise 'low', RoB was listed as 'unclear'. For brevity, overall RoB judgements for study-level effects are presented in forest plots (Figures 4.7, 4.9, 4.11, 4.13). In the eight full predictor reports, RoB plots display judgements for each domain for each included study.

The author evaluated all included studies; a second researcher sampled 20% of the included studies. No studies were excluded from the review on the basis of RoB decisions however, RoB judgements are taken into account if the study is indicated as influential in sensitivity analyses and are also used as a predictor in moderator analyses when performed.

4.2.3.3 *Measures of Subgroup Estimates*

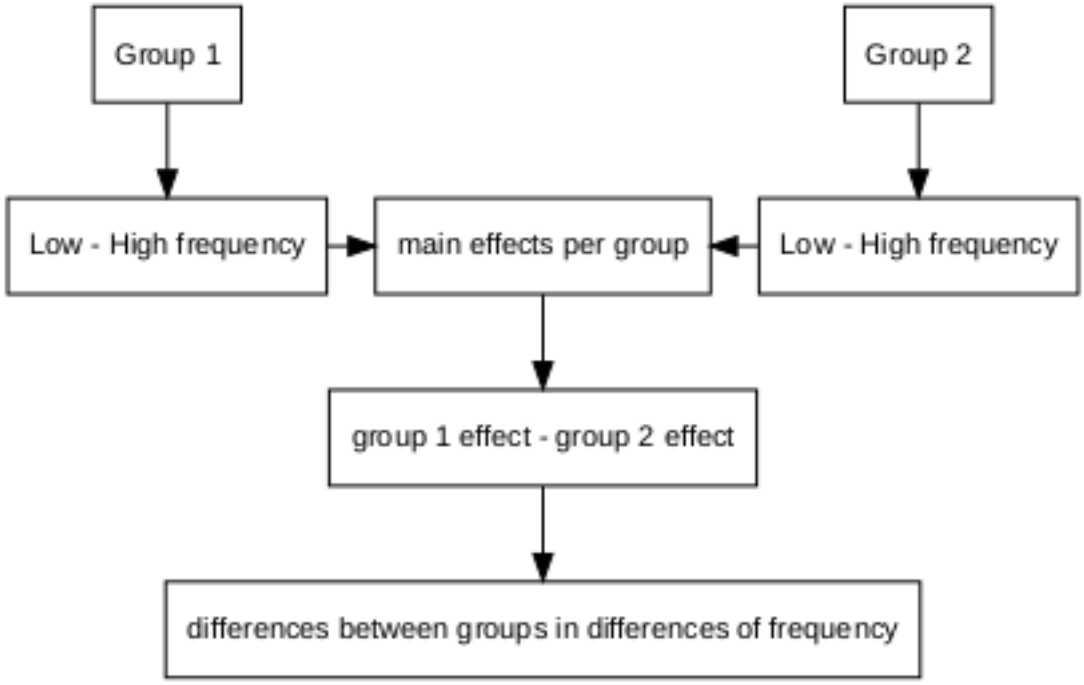
We extracted data reflecting main effects as well as interaction effects but since our primary focus is on the estimation of the differences in psycholinguistic effects between groups, summary effects for main effects are not reported here. They are included in each full predictor report at the project OSF repository.

For clarity, the summary effects reported represent the differences of the

differences between two groups and levels of a psycholinguistic variable. For example, the difference in the size of a word-frequency effect for younger vs older children in lexical decision accuracy (see Figure 4.3).

Figure 4.3

Flowchart Showing How Raw, Study-Level, Interaction Effects Were Calculated from Condition Means.



Included studies used a highly disparate range of items, measures and statistical tests. There was also evidence of unbalanced samples sizes. Therefore, rather than use raw study-level effects, we used standardised study-level effect sizes for estimation purposes (Baguley, 2012).

If mean and standard deviation values at the condition level were available, we calculated a standardised mean difference value for each study-level effect. Our effect size unit is Hedges' g with 95% confidence intervals. Hedges' g is an alternative to Cohen's d that adjusts for unbalanced sample sizes between groups[2].

[2]Hedges' g accommodates differences between the two groups' standard deviation *and* differences in sample size where Cohen's d assumes equivalent standard deviations and equal sample sizes. In Hedges' g, each group's standard deviation value is weighted by its sample size before pooling the standard deviation to calculate the effect (Ellis, 2010; Lakens, 2014).

We report absolute values of Hedges' g as a consequence of some study-level effects supplying only omnibus statistics for effect size estimation. Although we coded a variable to note the direction of the effect as verbally reported in study results, sometimes this was missing information. Reporting absolute differences is a conservative adjustment to prevent errors in interpretation, however it limits our capacity to talk substantively about directions of effect for a summary effect.

We interpret values of Hedges' g using Cohen's (1998) commonly used thresholds, i.e., Hedges' g of 0.2 - 0.49 as small, 0.5 - 0.79 as medium, 0.8 - 1.30 as large and higher values as very large effect size categories. We also use "very small" for estimates below 0.2.

We report 95% confidence intervals for each study-level and summary effect. Confidence intervals reflect a range of possible effect sizes that are compatible with the data with a long-run interpretation that the interval will contain the true value of the effect 95% of the time, given a sequence of valid models (Cumming and Finch, 2001). This application of confidence intervals is also consistent with the use of random effects models that assumes that estimates vary along a continuum. Consequently, we interpret effect sizes as 'reliable' when confidence intervals do not cross zero, as the limits of the range exclude the possibility of no effect or a reversal of direction of effect (Gelman and Carlin, 2014). When confidence intervals do cross zero, we interpret the effect size as 'unreliable' because the data suggest that no effect or a reversal of direction of effect is also plausible.

A further consideration when using confidence intervals is their relative width. In plotting study-level effects and their confidence intervals, and summary effects with their confidence intervals, we will be able to provide a visual representation of the precision of measurement of the effect for differences of a psycholinguistic variable. Where the width of a confidence interval crosses the threshold of two effect sizes, we interpret the measurement of the effect as 'imprecise'.

4.2.4 Data Synthesis

4.2.4.1 *Person-level and Item-level Contrasts*

We use the labels of *subgroup* and *global* effects to denote two levels of estimation in results. There are potentially five subgroup summary effects for each psycholinguistic variable and one global summary effect across each task outcome. The global effect is an average of all the study-level effects for that predictor, task and outcome.

As a reminder, we define child samples as participants who are younger than 18 years of age; adult samples are defined as 18 years of age and above. We define ability contrasts as children or adults of the same age who differ in levels of reading skill. We define experience contrasts in adult and child samples that are groups of younger vs older participants who show age-appropriate reading skills. We define age contrasts as children of different ages who have the same levels of reading skill.

4.2.4.2 *Analytical Models*

We know we are missing study-level effects. Consequently, the data set within this paper represents a sub-sample of the potential population. We assume that, given the different contrasts between studies, an effect may vary in size along a continuum rather than be one size across the alternative sample constructs (Baguley, 2012; Pigott, 2012; Shadish and Haddock, 2009). In some cases, we have included multiple effects per study and / or multiple studies per research paper. We assume smaller variation between effects reported within the same study or article than variation between effects from different studies in different research papers (Nakagawa and Santos, 2012).

These properties motivate the use of a random effects (RE) model rather than a fixed effects (FE) model for the following reasons: They represent a level of conservatism needed to represent the missing studies as they provide wider confidence intervals than FE models (Baguley, 2012). The RE model assumes the sample is a subset of the full population and accounts for a continuum of effect sizes. Choosing a

RE model also allows a grouping term to account for inter-correlation between study-level effects generated from the same article if and when they occur in the same analysis.

Consequently, where two or more studies contributed study-level effects for a psycholinguistic variable, we estimated an effect size using a random effects (RE) model. At the subgroup level, if only one study-level effect was available, we present a fixed effects (FE) estimate. Essentially, the single study-level effect acts as a placeholder until further study-level effects are accrued.

All data analysis is conducted in R (R Core Team, 2020). Study-level effects were transformed to standardised effect sizes using the `compute.es` package (Re, 2013)[3]. Meta-analysis RE models, sensitivity and moderator analyses and forest plots use the `metafor` package (Viechtbauer, 2010). The multi-level-random-effects model function in the `metafor` package (Viechtbauer, 2010) takes account of sampling variance, within-study heterogeneity, between study heterogeneity and covariance between the effects. We use restricted maximum likelihood (REML) methods to estimate predictions. Two further packages support P-Curve analyses: `meta` (Harrer et al., 2019a) and `dmetar` (Harrer et al., 2019b).

4.2.4.3 *Measures of Consistency*

A measure of between-study variability is of critical importance to understand how dissimilar study-level effects are from each other. Low between-study variability would mean that we could interpret a summary effect as being stable across varying research designs, with the inverse being true . We assess the presence of variability between study-level effects using two tests, Cochrane's Q and I^2.

Cochrane's Q is a null-hypothesis significance test that follows a chi-squared distribution with K-1 degrees of freedom. Assuming that between-study variance equals zero, a p value $< .05$ would be interpreted as a significant amount of variability between study-level effects and we would reject the null hypothesis of zero

[3]This package follows formulae recommended in Cooper et al. (2009)

between-study variance. One caveat of Q is that it inherits the properties of the chi-square distribution, which is well known to have difficulties with type I and II errors at low and high samples sizes (Borenstein et al., 2009).

I^2 is not affected by the number of study-level effects in a meta-analysis. I^2 equals Cochrane's Q minus the degrees of freedom, divided by Q and multiplied by 100 to reflect a proportion (Borenstein et al., 2017; Moher et al., 2009). Values for I^2 above zero percent tell us that between study variance is greater than that due to sampling variation alone, with values above 25% needing to be explored and explained. I^2 levels have been assigned thresholds to assist with interpretation: low heterogeneity $\geq$ 25%, moderate heterogeneity $\geq$ 50% and high heterogeneity $\geq$ 75% (Higgins & Thompson, 2002 cited by Borenstein et al., 2009). Higgins and Thompson (2002) note that these thresholds are not universal but it seems reasonable that values $\geq$ 25% provide sufficient variability (assuming sufficient data) for an exploration of possible moderators. These thresholds will be used to interpret estimated I^2 values and dictate whether moderator analyses are warranted.

4.2.4.4 *Additional Analyses*

Sensitivity Analyses. While tests for heterogeneity indicate the presence of between study differences, sensitivity analyses will perform several smaller statistical analyses to help identify individual study-level effects that may be exerting a large influence on the estimate. We perform sensitivity analyses of all subgroup effects using the `influence()` function from the `metafor` package (Viechtbauer, 2010). This performs a case deletion routine that provides leave-one-out diagnostics per study-level effect. Where indicated, we compare the identified study-level effect against its sample peers for methodological and design differences. We also take into account its RoB judgement. Where differences are apparent and the summary effect's I^2 value is above 25%, we may choose to remove the study-level effect. We describe this process within text, and update a subgroup effect if study-level effects are removed due to this process. An example of this process is included in the example report (Section 4.3.2).

Moderator Analyses. Where I^2 values for global effects remain at moderate or high levels after sensitivity analyses, and there are sufficient study-level effects ($k > 10$), we perform moderator analyses to consider whether heterogeneity in the estimate can be further explained by other properties of the sample. For brevity, none of the global effects in the example report (Section 4.3.2) warranted a moderator analysis so we do not report the process any further here. Full details are available in the full predictor reports at the project OSF repository.

4.2.5 Estimation of Bias Across Studies

Although we explicitly imputed p values at the .1 level for missing study-level effects, this does not help us estimate if entire studies are missing from the sample. We use statistical and graphical methods to help adjudicate the presence of missing studies.

4.2.5.1 *Statistical Methods*

Where more than three study-level effects contribute to a subgroup effect, we use the Egger's Test and the Rank Correlation Test (Begg and Mazumdar, 1994) as statistical measures of missing study-level effects. These tests measure the association between a study-level effect size and its standard error. Van Aert et al. (2016) recommends adopting a p value of $< .1$ for an Egger's Test and Rank Correlation Test.

4.2.5.2 *Graphical Methods*

Funnel Plots. We draw contour-enhanced funnel plots for global effects at the task x outcome level (Figures 4.8, 4.10, 4.12, 4.14) when there are 10 or more study-level effects (Lau et al., 2006). Each study-level effect size in the data set for task and outcome is plotted as a function of their standard error. We use different colours to represent groups. We centre the triangle at zero rather than the global effect value, to be able to see more clearly how the data relates to a null hypothesis of zero.

The contour-enhanced aspect of the plot involves shading areas of the funnel to denote different levels of statistical significance. The largest, white central region denotes p values $> .1$. Moving outward, the next shaded area represents data with p values between .1 and .05, the next, between .05 and .01 with study-level effects falling outside of the funnel boundaries having p values $< .01$. Relatively few or no study-level effects in the central, white section of the plot may suggest missing studies and bias in effect sizes, introducing bias in both publication rates and the global effect (Palmer et al., 2008).

P-curve Analysis. We also construct and inspect a p-curve for each task x outcome sample of studies (Figures 4.8, 4.10, 4.12, 4.14). P-curve analysis (Simonsohn et al., 2014) is a systematic method of evaluating the evidentiary value of a data set. From any task-outcome data set, only study-level effects with $p \leq .05$ are entered into p-curve. Visually, the shape of the distribution of p values in the subset of data can be indicative of strength of evidence for an effect. A right skewed distribution means p values are clustered close to .01. This indicates strong evidentiary value for an effect. An even spread of p values between .05 and .01 indicates a lack of evidential value. A left-skewed distribution means p values are clustered close to .05 and may be indicative of p-hacking.

P-curve analysis is supported by two inferential tests. The first is the 'test of right-skewness'. This tests that the number of $ps < .025$ is greater than the number of $ps > .025$. The subsetted data are split into two sets at the 0.025 level and tested against the uniform null (50% high) with a binomial test. If we can reject the null hypothesis of equal numbers or more $ps > .025$, we can infer that the data sample contains evidential value.

When we fail to reject the test of right skewness, we move to the second test, the 'test for flatness'. This tells us whether the data sample is under-powered to detect a very small effect or that more studies are needed. The data are re-analysed as if the power to detect an effect is now 33%. If we reject this test, we can infer that the data lack evidential value due to an effect that is too small to detect under current sample

sizes. If we fail to reject the null hypothesis of 33% power, we must infer that there is not enough information at the current time and more data is needed.

4.2.6 Confidence Judgements

Each summary effect is accompanied by an evaluation of how much confidence we hold that the value reflects a 'true' effect size. In our evaluations, we were guided by the GRADE process (Schunemann et al., 2011). The process involves evaluating the strength of evidence over five domains and lowering confidence levels when the evidence falls below a threshold. There are four confidence levels: *high, moderate, low* and *very low*.

As all included studies operate a quasi-experimental design, we initially placed each of the subgroup effects at 'moderate', which represents a belief that 'the subgroup effect size is *probably close* to a true effect size'. A 'low' confidence rating is explained as 'the estimated effect *may be markedly different* from the true effect'. A judgement of 'very low' is interpreted as 'the estimated effect is *probably markedly different* from the true effect' (our emphasis). We concealed summary effect labels during the evaluation process. See Appendix C for an overview of each domain in the adjudication process; code and data for confidence judgements is available at https://bit.ly/ConfidenceData

Threshold criteria for the five domains are:

Imprecision: We lowered the confidence rating by one level when two or more of the following criteria were met:

- statistical power to replicate an effect of half the size was below 80% *and* magnitude of the replication effect is above 2. Power and magnitude values are produced by `retrodesign` function (Gelman and Carlin, 2014, see 4.2.6.1 below).
- average participant sample size and total number of items for each effect are lower than a 40 x 40 *group by item condition* design

(Brysbaert and Stevens, 2018);

- an effect's confidence intervals cross more than two effect size categories (categories as defined by Cohen (1988), by visual inspection of forest plots).

Indirectness: We lowered the confidence rating by one level if an effect would not generalise to the population at this time, indicated by whether random effects models' credible intervals included zero (visual insepction of RE credible interval upper and lower limits).
Inconsistency: We lowered the confidence rating by one level if the I^2 value was high ($> 75\%$) after sensitivity and moderator analyses were performed (visual inspection of I^2 value).
Publication bias: We lowered the confidence rating by one level if the p values of either the Egger's or Rank Correlation Test $< .1$ (visual inspection of test p values).
RoB: We lowered the confidence rating by one level if an effect carried a high RoB.

We present confidence levels in the text and in summary tables. The information used to generate confidence levels for each summary effect is also presented in the summary tables (Figures 4.16-4.20).

4.2.6.1 *Estimating Replication Power*

We use the `retrodesign` function (RDF, Gelman and Carlin, 2014) to estimate statistical power to replicate a future effect of half the size of a summary effect, and estimate the potential for exaggeration in that effect size. Gelman and Carlin (2014) recommend retrospective *design* analysis based on an effect size "that is determined from literature review or other information external to the data at hand" (p.2) with standard error values taken from the current study. The `retrodesign` function takes four arguments:

1) d^{rep}, 'a random variable to be the estimate that would be observed in a hypothetical replication study with a design identical to that used in the original study' (Gelman and Carlin, 2014, p. 3). We define d^{rep} as half the size of each summary effect, after the findings of Open Science Collaboration (2015);
2) the standard error value, taken from each related summary effect;
3) a statistical significance threshold: $p < .05$[4];
4) the degrees of freedom, set as infinite within the function.

The function returns three outputs:

1) statistical power, range 0 - 1, multiplied by 100 for reporting as a percentage;
2) type S error rate, a probability of the replicated estimate producing a sign error (i.e. different direction of effect, range 0 - 1);[5]
3) an exaggeration ratio value. An exaggeration ratio value of 1 suggests that d^{rep} is returned, a value of 2 has returned the original summary effect size. Values > 2 reflect even larger effect sizes.

We use the statistical power value and the exaggeration ratio value in our confidence evaluation process.

4.3 Results

Results are presented in three parts. First, we give an overview of the data set. Second, we present subgroup effects for the word-frequency variable. We present results for frequency as it is the most represented effect of the set and due to its

[4]By calculating each set of RDF values as if d^{rep} is significant, we may be accused of holding unreliable findings for task outcomes to an inappropriate standard, and should assess power for an insignificant finding, however the *p*-curve analyses for most meta-analysis samples were inconclusive due to lack of available data rather than indicative of a minute effect size, consequently it is too early to make that assumption, and while these samples are explored, we should ensure adequate power.

[5]At this time, we cannot use the type S error rate. Due to some of the reporting of some study-level effects being incomplete or missing, we transformed our effect sizes to absolute values before beginning the meta-analysis.

longevity, is well measured[6]. Third and finally, we present summary effects for the psycholinguistic variables for which we recovered data and indicate which subgroup effects are missing at this time (Figures 4.16 - 4.20). All of the data and code to reproduce the meta-analysis findings are available at the project OSF repository.

4.3.1 Overview of the data set

We located 122 articles that were eligible for inclusion. These articles reported 155 studies in which the analyses estimated a total of 472 interaction terms, where the interactions were between 1) the effect of one of five different kinds of group contrasts and 2) the effect of one of eight psycholinguistic variables. Appendix B lists the articles and the study-level data needed to be able to reproduce the meta-analysed estimates.

The oldest study in the sample occurred 48 years ago with the most recent in 2019, demonstrating the longevity of this line of research. Articles originate from 14 distinct countries representing 10 languages. The majority of studies reported were conducted in English ($n = 71$), with Italian ($n = 15$), Spanish ($n = 12$), French ($n = 10$) , Dutch ($n = 7$) and Turkish ($n = 3$) languages also represented. Two studies contain participants native to Germany. One study has a Finnish sample. Most documents were peer-reviewed journal articles, however a non-negligible amount of eligible studies were found in doctoral theses ($n = 8$)[7].

4.3.1.1 *Psycholinguistic Variables*

Most of the studies comprised a quasi-experimental design where items were sampled to vary on two psycholinguistic variables, e.g., frequency (high vs low) and AoA (early vs late acquired), and where participants were sampled to vary on one dimension e.g., experience: young vs older children. Consequently, for most studies, we could extract

[6]The reports for arousal, AoA, consistency, frequency, imageability, length, neighbourhood-size and valence are also available in .pdf format at https://bit.ly/Meta-analysis-repository)

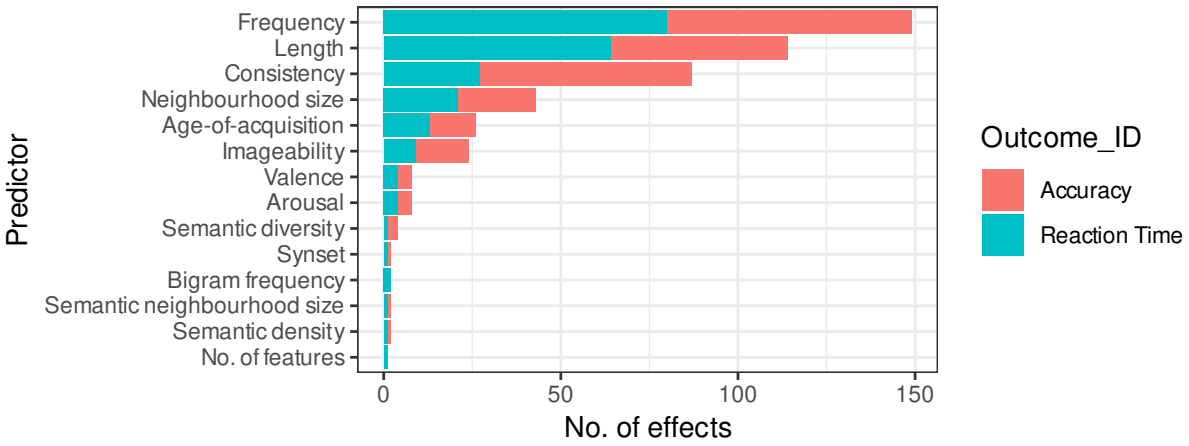
[7]More doctoral theses were eligible for inclusion, however, full text was not available, nor were the authors contactable to retrieve data

data for the difference between groups for more than one psycholinguistic variable. Figure 4.4 displays how the extracted 472 study-level interaction effects are distributed across psycholinguistic variables for reaction time and accuracy.

The most prevalent psycholinguistic variable is word-frequency ($n = 149$) with length ($n = 114$) and consistency ($n = 87$) next. N-size ($n = 43$), AoA ($n = 26$) and imageability ($n = 24$) are also present. Arousal and valence account for eight effects each. We refer to these eight variables as a core set from this point. Bigram frequency, no. of features, semantic density, semantic diversity, semantic N-size and synset each represent four or less effects each.

Figure 4.4

Distribution of Study-Level Interaction Effects Across Psycholingusitic Variables for Reaction Time and Accuracy.



4.3.1.2 *Participants*

Study-level effects were collected for adult-experience and adult-ability samples, child-age, experience and ability samples. Notably, the ability contrast contained a contrast with skilled readers vs either unskilled- or dyslexic-readers. By "unskilled" we mean that reading skills were lower than expected for their age. The "dyslexic" categorisation derives from formal diagnosis or multiple assessments carried out during sample screening procedures to detect dyslexic reading profiles. We coded for

these two levels in the ability contrast for use as a potential predictor variable in moderator analyses.

There is a lower number of study-level effects involving contrasts between adults than children (162 vs 310 respectively). The majority of adult studies focused upon age (100 vs 62) while ability holds a majority in the child samples compared to experience and age (153 vs 99 vs 58).

4.3.1.3 *Tasks*

The majority of papers focused on a sole task (32 for lexical decision; 73 for word naming) with the remaining 17 papers using both tasks. Given the longevity of the research period, there are differences in task administration. We detail these briefly next.

Item Presentation. Researchers used index cards, paper-based lists, slides or computer display presentation of items. Research teams involved in the earlier studies may have used index card or paper list presentation rather than slides or computerised presentations, and it is less likely for the paper based presentation method to report reaction time outcomes.

Word Naming Protocol. In computer based presentation, each naming trial involved a fixation point before item presentation, often with a voice key relayed to the presentation program that captured the response onset latency. Blocking methods for item presentation varied across studies with most using some level of randomisation either within and across blocks or, if using a fixed order within blocks, counterbalancing of blocks was reported. Time out thresholds also differed across studies. Error categorisation also showed variation: in one study, repeated attempts were allowed; in another, if the first articulation made was not the word (e.g. an "um", "ah" or a stumbled sound), this was counted as incorrect and discarded before data analysis. In all studies, analysis of reaction times was for correct responses only.

Lexical Decision Protocol. A computerised lexical decision trial involved a fixation symbol appearing before the item was presented which would disappear upon the participant's response. Gross differences centering around whether word and nonword items were mixed, blocked as separate items with randomisation within blocks and counterbalancing of blocks across participants occurred. Methods of recording the responses also varied between button boxes and keyboards.

4.3.1.4 *Outcomes*

In the word naming task, outcomes were either response accuracy or reaction time. There was almost an even split across the data (243 and 229, respectively). Reaction time is defined as the time elapsed between item onset and response onset, where the latter is usually registered by a voice key for word naming or a key press for the lexical decision task, respectively. Most studies report mean reaction times in milliseconds, with some reporting in seconds for lists of words, in which case a per word measure was derived by dividing the total list time by the number of list items. Accuracy is defined as the correct pronunciation of the word (word naming) or the correct identification of the item as a word or nonword (lexical decision). Many studies reported error rates. We subtracted these values from item totals (raw counts) or from 100 (percentages) to obtain accuracy rates. The distribution of reaction time and accuracy rates are displayed for psycholinguistic variables (Figure 4.4).

4.3.1.5 *Missing Data*

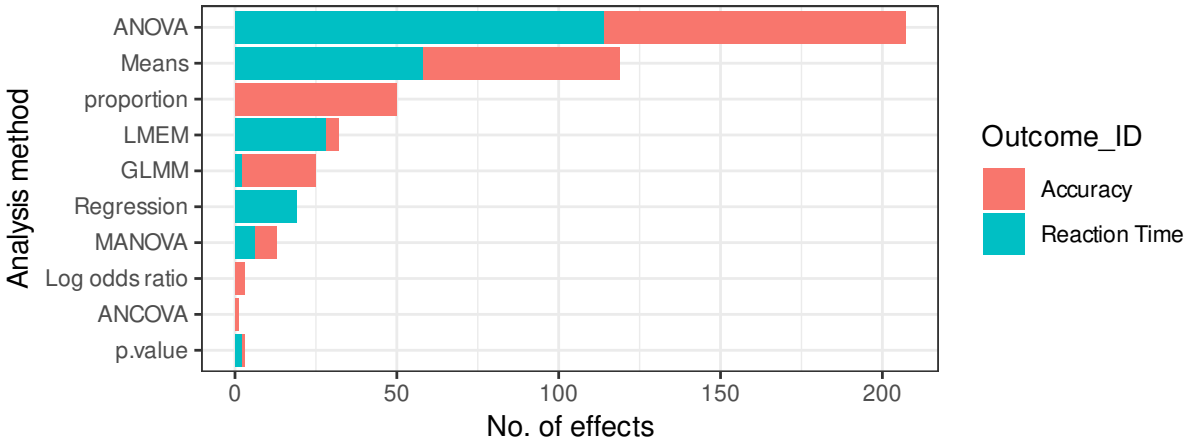
Of the 472 extracted effects, 386 were explicitly reported with statistical data. Where primary psycholinguistic variables had no statistical results reported at the study-level, we imputed a missing effect at a value of $p = .1$. There were 63 inferred and 23 missing observations. Of the missing data, reaction time has a higher level of imputed data points than accuracy (58 vs 28).

4.3.1.6 *Analysis Methods*

Figure 4.5 displays the distribution of statistical analysis methods for outcomes across the included studies. ANOVA is by far the most popular method of analysis (n = 207). Means for reaction time and accuracy was the second most prevalent (n = 119), with proportions correct for the accuracy measure third (n = 50) Linear mixed effects models for reaction time (n = 32) and generalised linear mixed effects models for accuracy (n = 25) were the next most prevalent with simple regression (n = 19) ranking the fifth most popular method.

Figure 4.5

Prevalence of Analysis Methods for Study-Level Effects.

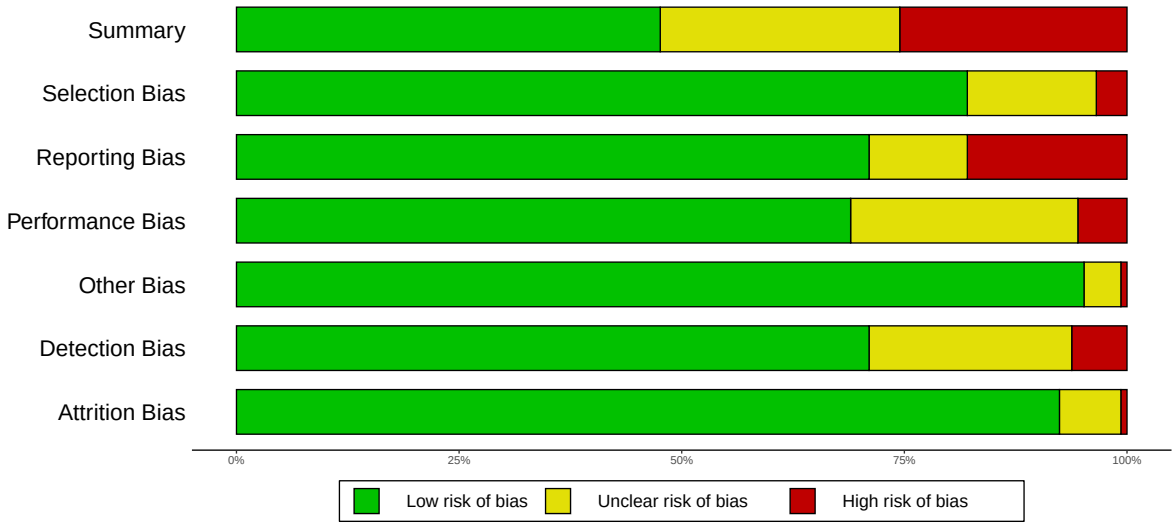


4.3.1.7 *RoB Within Studies*

Overall, just under half of the included studies were assessed as having a "low" RoB (n = 69). Thirty-seven were given a "high" rating of RoB and 39 were given a judgement of "unclear". Figure 4.6 shows how the three judgements are spread within each separate domain across the sample.

Figure 4.6

Adjudication of Risk-of-Bias Across Included Studies.



4.3.2 Group Differences for the Word-Frequency Effect: An Example Report

This section presents the meta-analysis of study-level effects that involve differences between groups for the word-frequency effect. We present word naming results first for reaction time and then accuracy, then lexical decision outcomes. Each section follows the same pattern: first, estimates are described. There is a global effect that aggregates all study-level effects and five subgroup effects, one for each group contrast (see Figure 4.2). As a reminder, the thresholds for interpretation of Hedges' g are 0.2 - 0.49 as a small effect, 0.5 - 0.79 as a medium effect, 0.8 - 1.30 as a large effect and any size larger than 1.31 as very large. Effect sizes below 0.2 are very small.

Each subgroup effect captures the aggregated magnitude of the difference *between* the paired, contrasted groups for how differently a group uses levels of word-frequency (see Figure 4.3). For instance, between the typical and atypical readers in the child-ability contrasted group in word naming reaction time, there is a difference in the word-frequency effect of $g = 0.6$, which tells us that for one of those groups, the difference in reaction times is greater between high and low levels of

frequency than for the other group, and that the difference between the two groups' word-frequency effect is 0.6. of a standard deviation.

The estimates are presented in a forest-plot. Each subgroup has its own section that displays the subgroup effect below the study-level effects of which it is comprised (explained in more detail below). We do this to visualise both the heterogeneity between studies within a subgroup and also heterogeneity between subgroups. Displaying study-level effects also gives the reader a strong picture of how well an estimate is populated. Additionally, we include the global effect that is an aggregate of the entire task-outcome data set.

After the first estimation of effects, we perform sensitivity analyses and moderator analyses, where indicated. Study-level effects may be removed at this stage. In which case, the relevant subgroup effect and the global effect for the task outcome are updated. Finally, we describe the results of tests for publication bias and *p*-curve analyses, presenting funnel plots and *p*-curve plots where data permits. A brief summary ends each section.

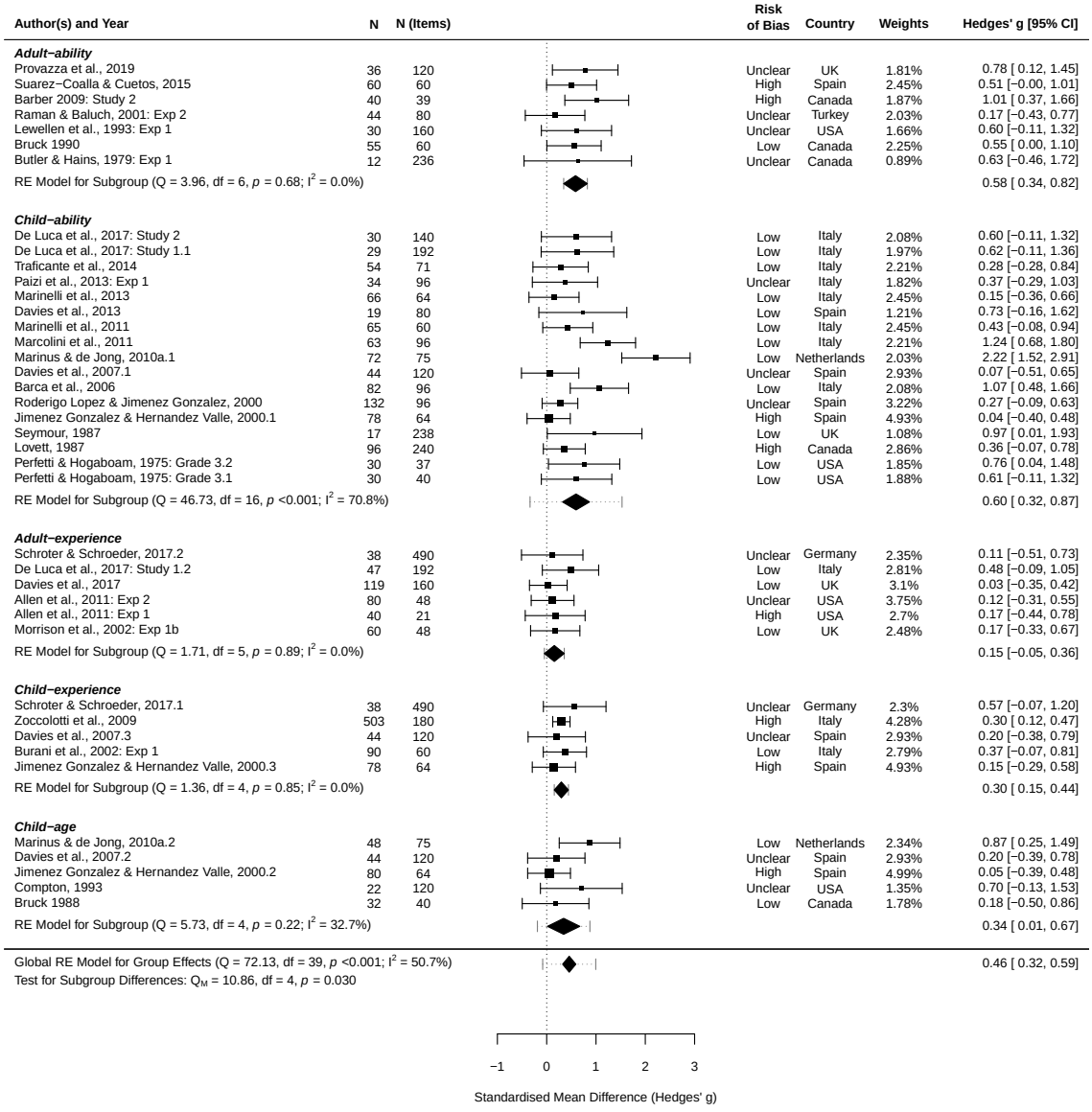
4.3.2.1 *Word Naming Reaction Time*

Thirty-two papers contributed 40 study-level effects for group differences in word-frequency effects on word naming reaction time. A global effect estimate for this sample is Hedges' $g = 0.46$, 95% CI [0.32, 0.59], $I^2 = 50.72\%$. This suggests that, on average, when two groups were compared in how much word-frequency affected response reaction time, the groups' frequency effect differed by almost half a standard deviation. Figure 4.7 displays the sample forest plot, which we explain next.

The plot is arranged in sections, one for each subgroup contrast. Each section displays the included study-level effects arranged in chronological order of year of publication. A study-level effect is visually represented by a solid square, the size of which denotes its weighted contribution to the subgroup effect. Bars to the left and right of each square denote its 95% confidence interval. The values for each study-level effect size are also given in text in the right-hand side column.

Figure 4.7

Standardised Mean Differences Between Groups for Frequency Effects on Word Naming Reaction Time



An estimate for the subgroup effect is at the bottom of each section shown by a solid diamond shape. The centre of the diamond represents the effect size value, while the width of the diamond extends to the 95% confidence intervals. Bars that extend to the left and right of the diamond, represent the credible intervals indicating if generalisation to new samples is reliable. To the left of each subgroup effect, we indicate whether the model is a random effects (RE) or fixed effects (FE) model alongside Q and I^2 values.

A further five columns in the forest plot presents sample size, item sample size, the study-level RoB judgement, country of origin for the first author and the relative weight that a study-level effect carries in the global effect for the task outcome.

Toward the very bottom of the forest plot, a further solid diamond shape represents a global effect, following the same interpretation as the subgroup diamonds. Additional to the tests for heterogeneity for the global effect, we display information for a formal test of statistical difference between subgroup effects. A p value $< .05$ suggests that at least one of the subgroup effects is significantly different in size from the other four.

We describe the subgroup effects and the results of sensitivity analyses next.

Seven studies explored adult-ability differences for frequency, including 277 adults (median $n = 40$, range = 12 to 60). A group difference of Hedges' $g = 0.58$ [0.34, 0.82], $I^2 = 0\%$, indicates a reliable, medium size effect with 0% heterogeneity between studies. Lower reading skill in adults is associated with a continued word-frequency effect in maturity that is larger than their adult, typically reading peers. None of the seven studies were indicated as influential in the sensitivity analysis. We have moderate confidence in this subgroup effect.

For the child-ability sample, 941 children were tested across 17 studies (median $n = 54$, range = 17 to 132) to give a Hedges' g of 0.6 [0.32, 0.87], $I^2 =$ 70.78%. This is a medium sized, reliable effect. Lower reading skill is associated with 0.6 of a standard deviation difference for frequency effects on reaction times.

Two studies were identified as influential in this sample: Marinus and de Jong

(2010) and Marcolini et al. (2011). Comparing these studies to others revealed that the frequency manipulation in Marinus and de Jong (2010) occurred in a post hoc analysis; the original items were all of high frequency values. Consequently, the frequency values within the original list may represent a restricted range of frequency, compared to other studies in this sample.

The primary interest of Marcolini et al. (2011) was in frequency effects on morpheme-based reading, manipulating both the root and the suffix frequency of their items and creating a contrast both within and across items. No other study manipulates frequency within the structure of the target item.

Given these differences, we removed the two study-level effects and updated the subgroup effect to a new Hedges' g that is small but still reliable (0.39 [0.24, 0.55]). Cochrane's Q is no longer significant ($Q = 14.24$, $p = 0.432$) with variability reduced to very low levels: $I^2 = 9.64\%$. We have low confidence in this subgroup effect resulting from both Egger's Test and the Rank Correlation Test showing $p < .1$.

Six studies looked at experience differences in adults (remember that we define experience as a pure age contrast in typical readers) and word-frequency with a total sample size of 384 people (median $n = 53.5$, range = 38 to 119). Hedges' g for this subgroup effect is 0.15 [-0.05, 0.36], $I^2 = 0\%$, a very small, unreliable effect. No study-level effects were indicated as influential. We have very low confidence in this subgroup effect as both confidence and credible intervals cross zero.

An estimate for the subgroup effect of the child-experience contrast (i.e. younger vs older children of typical reading skills) is Hedges' $g = 0.3$ [0.15, 0.44], I^2 at 0%. The estimate is generated from five study-level effects from 753 children (median $n = 78$, range = 38 to 503). It is a small sized effect and is reliable. The picture of any direction for this effect is not clear as some studies did not fully report or interpret the direction of their study-level effect due to its non-significant status.

In sensitivity analyses, Zoccolotti et al. (2009) was indicated as influential, however the study-level effect size ($g = 0.3$) mirrors that of the subgroup effect. It is also the only study-level effect of the subgroup with confidence intervals that do not cross zero. With I^2 at 0%, we decided to retain the study-level effect. We have very

low confidence in this subgroup effect.

Five study-level effects pertain to the child-age subgroup effect - samples of the same reading skill who differ in age. This sample, comprising 226 children (median $n = 44$, range = 22 to 80), produces a subgroup effect that is small and reliable: Hedges' $g = 0.34$ [0.01, 0.67], $I^2 = 32.65\%$. For the most part, we see the older students of atypical reading skill showing a larger effect for frequency than the younger, typical reading participants.

Two study-level effects were indicated as influential for the child-age subgroup effect: Jimenez Gonzalez and Valle (2000) and Marinus and de Jong (2010). The Jimenez Gonzalez and Valle (2000) effect is very small ($g = 0.05$). This study-level effect is in the opposite direction compared to the rest of the data. In this study, younger typical reading participants are first grade children. Much younger than the other younger child participants. Marinus and de Jong (2010) may show an influence due to a post hoc manipulation on frequency being applied (as described above). Although Jimenez Gonzalez and Valle (2000) continued to be indicated as influential after the removal of Marinus and de Jong (2010) (Hedges' $g = 0.19$ [-0.1, 0.48]), we chose to retain the study-level effect since the measure of heterogeneity between the remaining four studies was reduced to zero percent and Q was no longer significant ($Q = 1.88$, $p = 0.598$). We have low confidence in this subgroup effect.

The updated global effect is now Hedges' $g = 0.34$ [0.25, 0.44]. It remains small and reliable. We conducted a formal test of the subgroup effects to see if they were statistically different from each other. Recall that the reference level effect for this analysis is the adult-ability subgroup effect. The test was not significant ($QM_4 = 9.01$, $p = 0.061$). With heterogeneity reduced to 10.36% in the global effect, we did not perform a moderator analysis.

Tests for Small Study Bias and Publication Bias. Only the child-ability subgroup tested significantly for small study biases (Egger's $p = 0.043$, Kendall's $\tau = 0.46$ $p = 0.011$). Figure 4.8 shows a funnel plot (left) and a p-curve analysis plot (right) for the study-level effects contributing to the global effect of the differences in

frequency on word naming reaction time.

As a reminder, a study-level effect is plotted by its effect size (x-axis) and standard error (y-axis). The funnel plot contours relate to levels of statistical significance (central white region = $p > .1$, dark grey = p value = .1 - .05, light grey = p value = .05 - .01, beyond funnel boundaries = p value <.01) and an absence of study-level effects in the central white region is suggestive of publication bias.

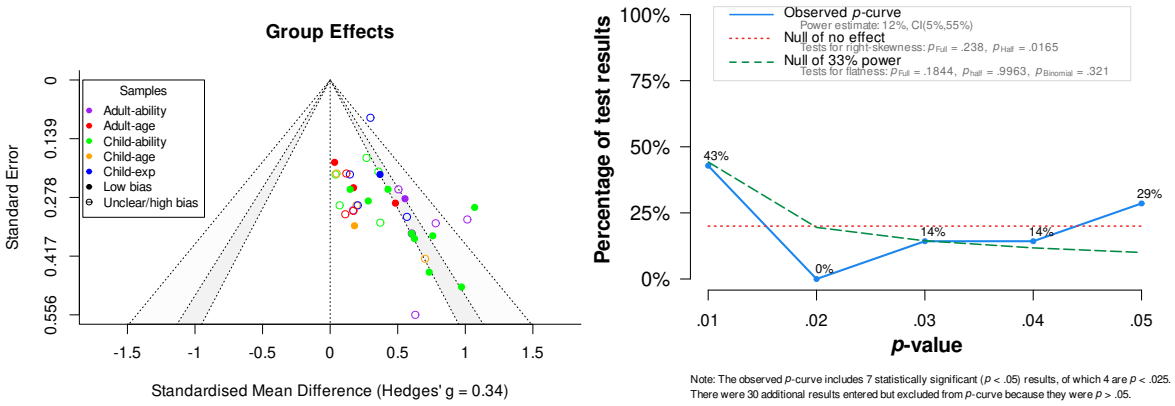
In reading a funnel plot, it follows that a study-level effect that is large and precise will be far away from the central line of the plot (zero effect) and towards the top (small error). A study-level effect that is close to the central line and towards the bottom of the plot has a small effect size and a large standard error value. Given that our Hedges' g effect sizes are all positive (we coerced absolute values, see Section 4.2.3.3), a strong and precise global effect will be characterised by study-level effects located in the top right quadrant of the plot. If study-level effects move down, either into the centre of the plot or remain in the bottom right quadrant, a global effect may be less precise and less robust.

Notice in Figure 4.8 how the study-level effects for adult-ability and child-ability, (purple and green circles), move down the plot as they move away from zero, showing a trend of less precision in the larger sized study-level effects. Of particular interest is the variability in the precision, especially for the green child-ability points, with some of the smallest and largest standard error values in the cohort. Given this spread of child-ability study-level effects in the lower right quadrant of the plot, we would expect to see some study-level effects in the central white region aswell, however there are none. This could suggest publication bias, as indicated by the Egger's Test and Rank Correlation Test result for the subgroup.

We have indicated study-level effect risk-of-bias (RoB) crudely using solid, filled circles to denote low RoB judgements and unfilled circles to denote high or unclear RoB judgements. Notice the spread of unfilled points as a band of study-level effects spanning all contour sections of the plot. This may suggest that consideration of methods to reduce sources of systematic bias at the study design stage may be useful for future research.

Figure 4.8

Funnel Plot and P-Curve Analysis Plot for Frequency Effects on Word Naming Reaction Time



Finally, we conducted a *p*-curve analysis as a check of the evidentiary value of the sample. Of 37 study-level effects, only seven had p values $< .05$ and were eligible for use in this analysis (3 study effects were removed from the initial sample of 40 as a result of the sensitivity analyses). Consequently, estimated power to detect an effect for word-frequency on word naming reaction time when the effect is present is critically low: 12% [0.05, 0.55]. The result of the test for right-skewness reflects this ($p = .238$), meaning we must fail to reject the null hypothesis of no effect. The blue curve in the right hand plot in Figure 4.8 shows that the seven values are distributed throughout the .01 - .05 range, three of the seven values are $p > .025$, i.e., it is not right-skewed, as we would hope in a data sample that has evidentiary value.

We move to the second test of the *p*-curve analysis. Statistical power is reset to 33% and the distribution of data is tested against this null hypothesis of 33% power. If the result is $p < .05$, we can reject this hypothesis and interpret the finding as if a small effect *is* present, however it is too small for the current data to detect and adjust sampling rates for future studies. Where $p > .05$, *p*-curve is inconclusive.

Tests for flatness for this data are all $p > .05$. We fail to reject the null hypothesis of 33% power to detect a small effect if it is present, so the judgement about evidential value within this set of data is inconclusive and we must wait for

more study-level effects to accrue before a decision is made.

Word Naming Reaction Time Summary. The difference between contrasted samples of participants in the effect of frequency on word naming reaction time is very small for adult-experience and child-age groups. Both of these subgroup effects cross zero. Hence, there are limits in our capacity to distinguish small effects as the results are compatible with no effect. The subgroup effect is small for child-experience and child-ability groups and medium sized in the adult-ability contrast group. On average, the impact of the difference between low or high frequency words on word naming reaction time is smaller in skilled compared to less-skilled children or adults, and it is smaller in less- compared to more experienced children. There could be publication bias or small study bias for the child-ability sample as indicated by its Rank Correlation Test result and funnel plot. If corrected, this suggests the small subgroup effect could be moderated downward. At this time, the *p*-curve analysis results suggests that the data sample is too limited to be conclusive about its power to detect an effect. We return to these points in the general discussion.

4.3.2.2 *Word Naming Accuracy*

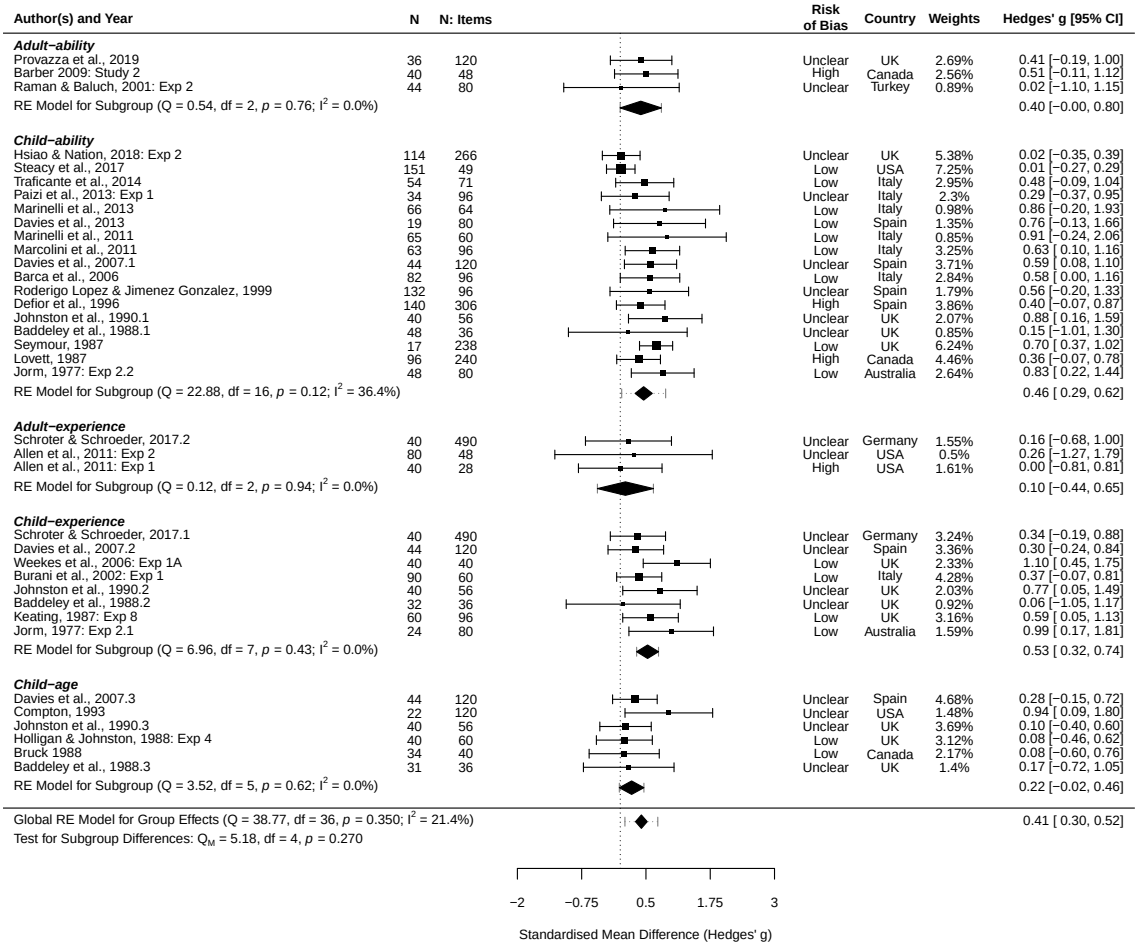
The analysis for word naming accuracy includes 37 study-level effects from 29 studies. A global effect for group differences in word-frequency for accuracy rates is Hedges' g = 0.41 [0.3, 0.52], I^2 = 21.42%. The test for subgroup effects' differences within the sample is non-significant ($QM_4 = 5.18$, $p = 0.269$), suggesting that each subgroup effect size for differences are similar in size. The forest plot detailing study-level effects and subgroup effects is shown in Figure 4.9.

But for the child-ability subgroup, I^2 values for the following subgroup analyses are all at 0%, indicating that the variability due to extraneous variables is equal to or lower than the variability due to random sampling variation.

An estimate of the interaction between the effect of frequency and the difference between adult-ability groups is a small and unreliable Hedges' g of 0.4 [0 ,

Figure 4.9

Standardised Mean Differences Between Groups for Frequency Effects on Word Naming Accuracy



0.8], given a sample of three study-level effects (n = 120, median n = 40, range = 36 to 44). We have very low confidence in this subgroup effect.

The difference in word-frequency effect for the child-ability subgroup is estimated as a reliable and small Hedges' g = 0.46 [0.29 , 0.62], for a total sample of 1213 children (median n = 63, range = 17 to 151), given 17 study-level effect estimates. This is the only subgroup to show any level of heterogeneity between studies (Cochrane's Q = 22.88, p = 0.117, I^2 = 36.45%). It is also the only subgroup to present with an influence from one study-level effect. The items for Steacy et al. (2017b) were a mix of exception and strange words, where every other study, if exploring consistency at the same time as frequency, contains a mix of regular and irregular items. This could represent a difference amongst the larger set of studies. We elected to retain the study effect as the subgroup Q test for heterogeneity was non-significant. We have a moderate level of confidence in this subgroup effect.

The adult-experience sample showed a very small and unreliable Hedges' g of 0.1 [-0.44 , 0.65]. This cohort is also small, including only 160 adults (median n = 40, range = 40 to 80) across three study-level effects. We have very low confidence in this subgroup effect.

The child-experience subgroup effect is estimated as a Hedges' g of 0.53 [0.32 , 0.74]. This is a medium size effect from eight studies with a total of 370 children (median n = 40, range = 24 to 90). More experienced children show a smaller frequency effect than less experienced children. We have a moderate level of confidence in this subgroup effect.

The child-age subgroup effect is a small and unreliable Hedges' g of 0.22 [-0.02 , 0.46] from six study-level effects, representing 211 children (median n = 37, range = 22 to 44). We have very low confidence in this subgroup effect.

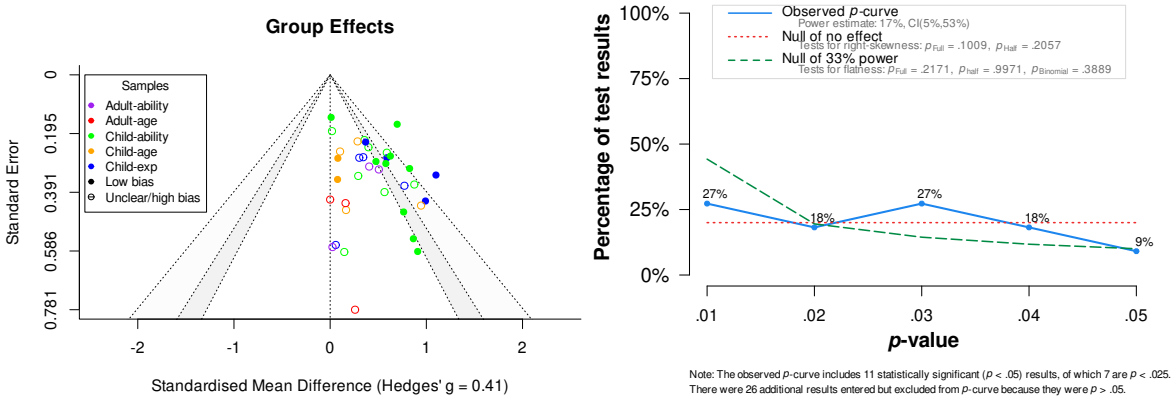
Given the very low levels of heterogeneity across the sample, we did not perform a moderator analysis.

Tests for Small Study and Publication Bias. None of the subgroup effects gave any indication of publication bias in the Egger's Tests or Rank Correlation Tests (ps

all > .1). The funnel plot for the 37 effects and a *p*-curve analysis are displayed in Figure 4.10. As with the reaction time funnel plot, generally, the two ability subgroups show the wider array of standard error values.

Figure 4.10

Funnel Plot and P-Curve Analysis Plot for Frequency Effects in Word Naming Accuracy



Eleven of the 37 study-level effects had $p < .05$ and were eligible for *p*-curve analysis (right hand plot of Figure 4.10). As with reaction time, the accuracy effect sample appears grossly under-powered (17% [5, 53.4]). The test for right-skewness is non-significant ($p = .100$) as is the test for flatness ($p = .217$) so we must conclude that at this time, the data set lacks evidentiary value of a true effect and more information is needed to draw valid conclusions.

Word Naming Accuracy Summary. Group differences for the word naming accuracy outcome cover a range of effect sizes. The child-experience subgroup shows a medium sized difference in word-frequency effects. Three groups show small size effects: child- and adult-ability and child-age. Although in different thresholds, the direction of effects is the same. For identical items, older children, and participants with higher reading skill should show greater accuracy than younger children or participants of lower reading skill, with the difference being that the older and more-skilled participants find words of lower frequency easier to name correctly. The

adult-experience subgroup effect is very small. Furthermore, the adult-experience, child-age and adult-ability effects are unreliable, compatible with an effect of no difference. An interpretation of the *p*-curve analysis suggests caution around any substantive interpretation of effects, with low power and insufficient information by which to generate valid inferences.

4.3.2.3 *Lexical Decision Reaction Time*

Thirty-three studies giving 40 study-level effects are eligible for meta-analysis of the interaction between the effects of word-frequency and group differences on lexical decision reaction time. There is a global effect size of Hedges' $g = 0.35$ [0.25, 0.45], $I^2 = 31.88\%$. It is a small sized, reliable effect with a low amount of heterogeneity. The test for differences between the subgroup effect estimates is non-significant ($QM_4 = 3.03$, $p = 0.552$), indicating that the subgroup effect sizes are similar to each other. Figure 4.11 shows the forest plot of the study-level effects.

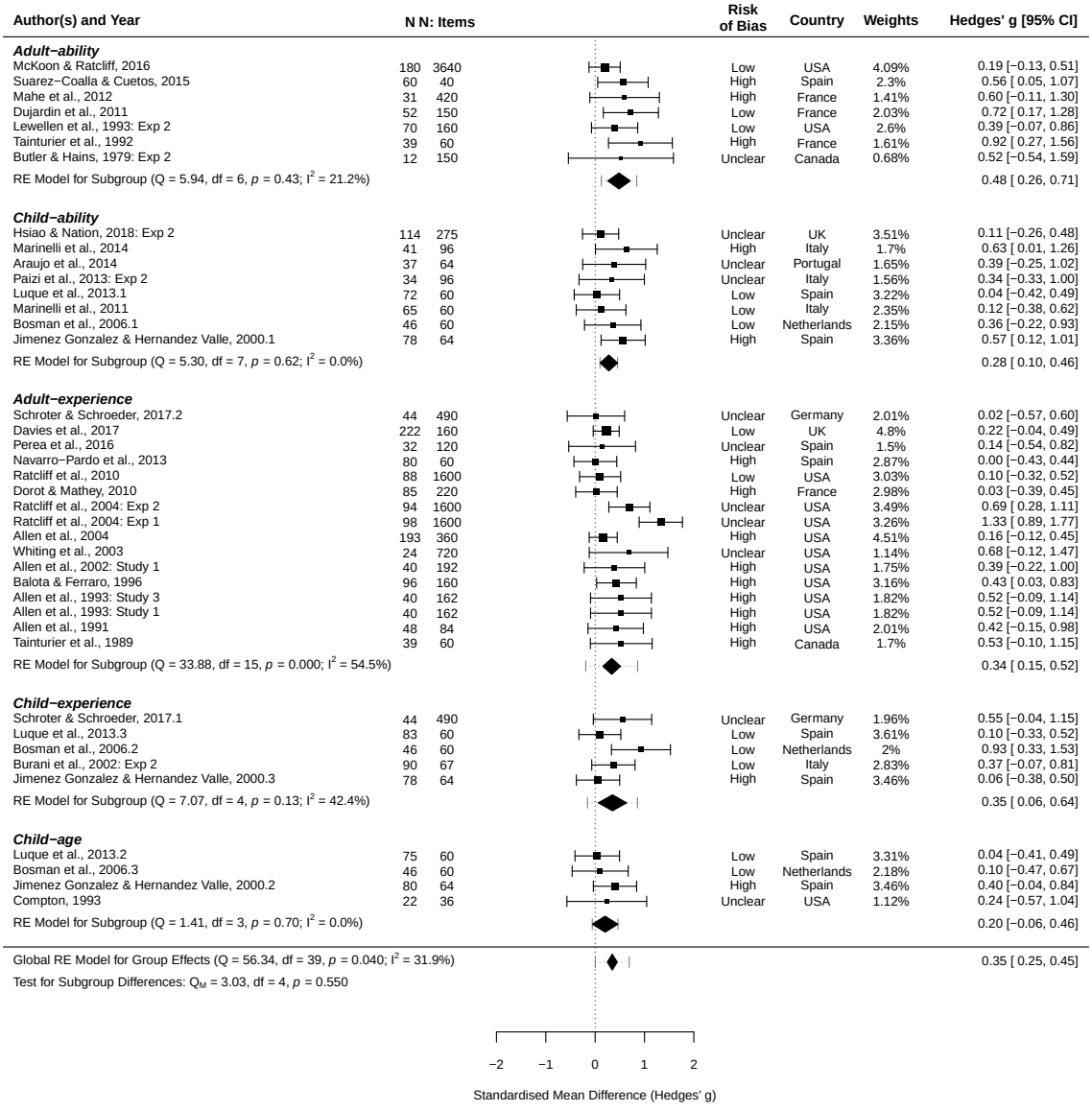
Seven studies (444 adults (median $n = 52$, range = 12 to 180) reported the interaction effect for word-frequency in the adult-ability subgroup for lexical decision reaction time. The subgroup effect is a small, reliable Hedges' $g = 0.48$ [0.26, 0.71], $I^2 = 21.2\%$.

The sensitivity analysis for the adult-ability subgroup identified McKoon and Ratcliff (2016) as an influential effect. It carries the greatest weight within the sample as a function of its higher precision and higher sample sizes. Heterogeneity values of the full sample were already very low and removing the McKoon and Ratcliff (2016) effect increased the value of the subgroup effect. We retained this study-level effect. We have very low confidence in this subgroup effect.

The child-ability contrast sample has eight studies (487 participants, median $n = 55.5$, range: 34 to 114). The difference between these groups is Hedges' $g = 0.28$ [0.1, 0.46], $I^2 = 0\%$, a small, reliable effect suggesting that, on average, children who are similar in age but differ in reading skill will tend to present different effects of word-frequency on lexical decision reaction time of approximately a quarter of a

Figure 4.11

Standardised Mean Differences Between Groups for Frequency Effects on Lexical Decision Reaction Time



standard deviation. Children observed to be lower in reading skill were found to show greater differences in reaction time between responses to low versus high frequency words. We have a moderate level of confidence in this subgroup effect.

A small and reliable interaction effect between the effects of word-frequency and group differences is also seen for the contrast between adult-experience groups, Hedges' g of 0.34 [0.15, 0.52], $I^2 = 54.47\%$. This group generated the largest sample of study-level effects ($k = 16$), involving 1263 participants (median $n = 64$, range: 24 to 222).

Ratcliff et al. (2004): Experiment 1 is indicated as influential for the adult-experience subgroup effect. With a Hedges' g value of 1.33, it is clearly much larger than the subgroup effect. This study-level effect is generated from a table of means within the article that presents a range of SE values rather than specific values, from which the mid-point was used to generate the study-level effect size. This may have introduced a bias so the study-level effect was removed from the sample and the subgroup effect updated to Hedges' g of 0.27 [0.16, 0.39], that remains reliable and small in size. I^2 was reduced to 0% (Cochrane's Q-test = 12.88, $p.$ = 0.536). We have very low confidence in this subgroup effect.

Five studies for child-experience, involving 341 participants (median $n = 78$, range = 44 to 90) contributed to a small, reliable group difference of Hedges' $g = 0.35$ [0.06, 0.64], $I^2 = 42.4\%$. Across these studies, younger children of typical reading skill generally presented the word-frequency effect to a greater extent than their older, typical reading peers.

Bosman et al. (2006) has a much larger study-level effect size than the subgroup effect and is identified as influential in the sensitivity analysis. The younger readers made twice as many errors as the older readers on the low frequency words in the sample; they made four times the errors on low frequency words compared to their own error rates on high frequency words. Since these errors were trimmed from the reaction time analysis, low frequency words may be under-represented in the sample. We removed the study-level effect and updated the analysis to Hedges' $g = 0.23$ [0, 0.46], a small sized effect with the lower confidence interval at zero. Heterogeneity

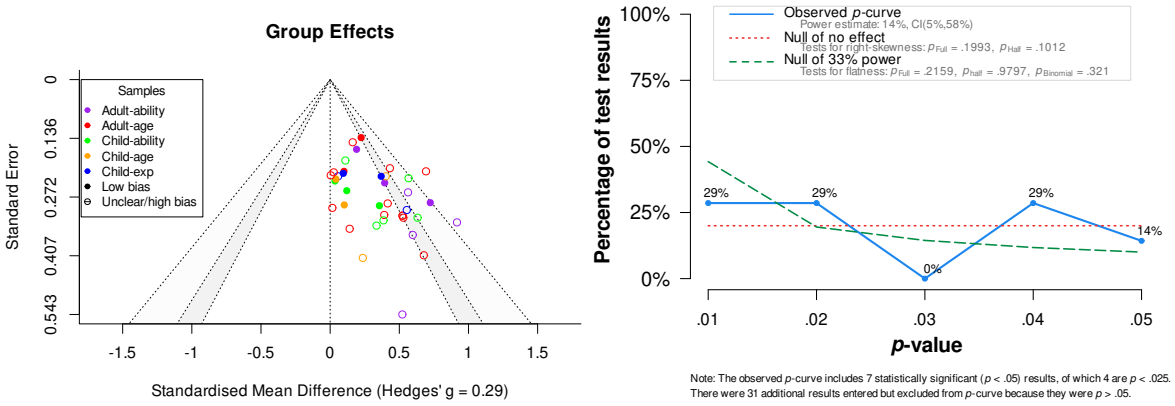
levels within the sample were reduced to $I^2 = 0\%$ (Cochrane's Q-test = 2.52, $p. = 0.472$). We have low confidence in this subgroup effect.

In the child-age subgroup four studies involving 223 children (median $n = 60.5$, range = 22 to 80) contributed to a small but unreliable group difference of Hedges' $g = 0.2$ [-0.06, 0.46], $I^2 = 0\%$. We have very low confidence in this subgroup effect.

With the removal of two study-level effects, the updated global effect is now Hedges' $g = 0.29$ [0.21, 0.37], $I^2 = 0\%$. The estimate remains small and reliable. The test for subgroup effect differences remains non-significant ($QM_4 = 3.58$, $p = 0.466$). We did not conduct a moderator analysis.

Figure 4.12

Funnel Plot and P-Curve Analysis Plot for Frequency Effects on Lexical Decision Reaction Time



Tests for Small Study and Publication Bias. Study-level effect sizes are plotted in the left-hand side plot of Figure 4.12. While tests of publication bias on the whole sample were significant (Egger's $p = 0.044$, Kendall's Tau = 0.26, $p = 0.02$), this was driven by one group, the child-experience subgroup (Egger's $p = 0.021$). There are very few study-level effects for the child-experience subgroup (blue circles) but there is a potential gap of study-level effects indicated by fewer blue circles in the white region of the plot compared to the outer contours. Also notable, the study-level effects with judgements of unclear and high risk of bias (unfilled circles) populate the

outer edges of the distribution with the greater quantity of low risk of bias study-level effects towards the top half of the collection in the white region of the plot.

In *p*-curve analyses, 7 of the 38 study-level effects were eligible for analysis (being $p < .05$). Of the seven, four were $< .05$, such that the distribution of p values looks fairly even (see right-hand side plot in Figure 4.12), suggesting that the sample lacks evidential value. To support this, the test of right-skewness is non-significant ($p = .199$) and the test for flatness is non-significant ($p = .216$). We fail to reject both null hypotheses of the *p*-curve analysis and conclude that the sample lacks evidentiary value of a true effect. More data is needed.

Lexical Decision Reaction Time Summary. Findings for group contrast effects on lexical decision reaction time suggest that, where reliable, group differences in the impact of word-frequency on lexical decision reaction time are small. The child-age and child-experience subgroup effects are both small and unreliable effects. The Q test for differences between the sizes of the effects is non-significant. The data set is sparse for robust findings, as indicated once more by the *p*-curve analysis results.

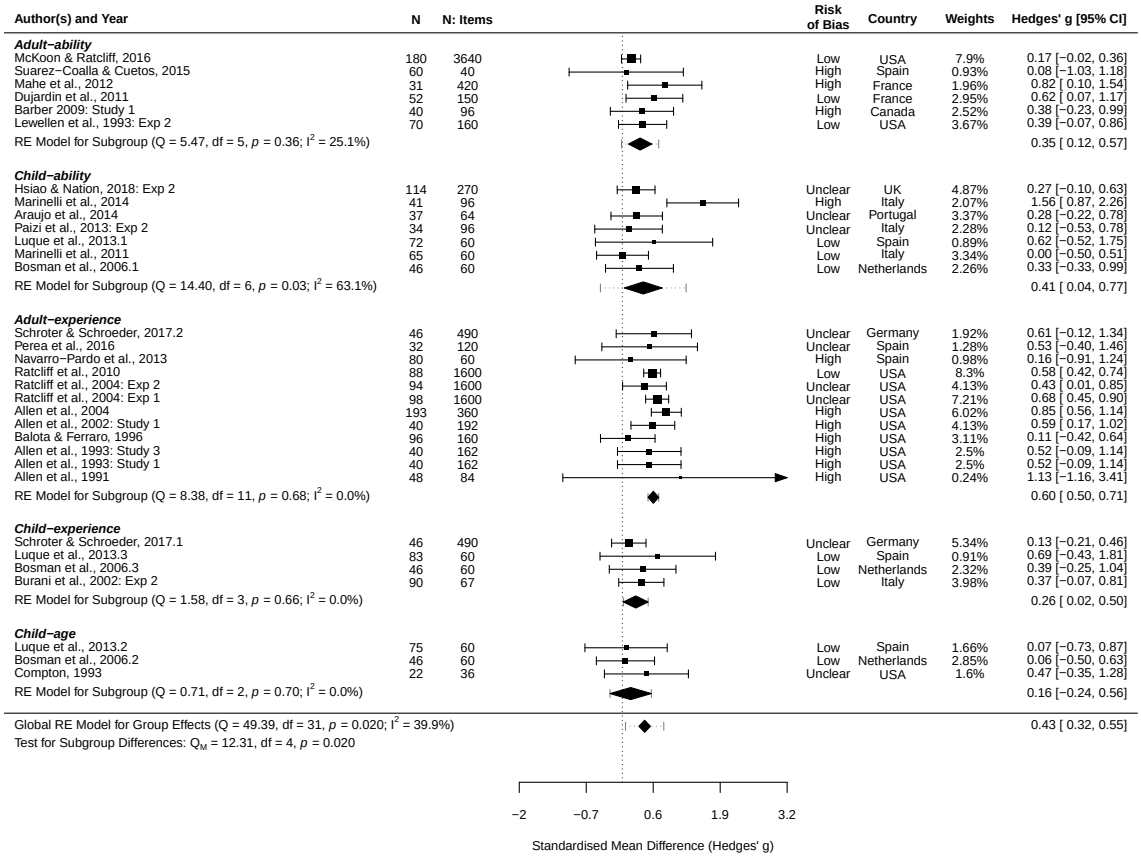
4.3.2.4 *Lexical Decision Accuracy*

Group effects for differences in frequency on the effect of lexical decision accuracy outcomes are estimated by 32 study-level effects from 27 papers. The global effect of the interaction between frequency and group differences is estimated as a Hedges' g of 0.43 [0.32, 0.55], $I^2 = 39.94\%$. It is small and reliable. This estimate suggests that, on average, word-frequency's effect on accuracy rates in lexical decision differs by approximately 0.4 of a standard deviation for contrasted groups. A test for subgroup differences is significant ($\text{QM}_4 = 12.31$, $p = 0.015$). The adult-experience subgroup effect is significantly different from the other subgroup effects.

For group differences in adults contrasted by ability, 433 participants from six studies (median $n = 56$, range = 31 to 180) give a Hedges' $g = 0.35$ [0.12, 0.57], $I^2 = 25.06\%$, a small, reliable subgroup effect. McKoon and Ratcliff (2016) was suggested

Figure 4.13

Standardised Mean Differences Between Groups for Frequency Effects on Lexical Decision Accuracy



as influential in a sensitivity analysis, however the study effect was comfortably within the confidence intervals of the subgroup effect (see Figure 4.13) and the Cochrane's Q test was non-significant ($p = .36$). We retained the study-level effect within the sample. We have low confidence in this subgroup effect.

Sample sizes for child-ability groups have a median size of 46 participants, (range = 34 to 114) with a total sample size of 409. The subgroup effect size is a Hedges' $g = 0.41$ [0.04, 0.77], $I^2 = 63.06\%$), a small, reliable effect size. Marinelli et al. (2014) is presented as an influential study-level effect. It is very large (Hedges' $g = 1.56$) and far away from the upper confidence levels of the subgroup effect. Marinelli et al. (2014) are the only study in the subgroup that included unpronounceable letter strings in their item sample. Given this difference, we removed the study-level effect and updated the subgroup effect to Hedges' $g = 0.22$ [0, 0.44], $I^2 = 0\%$. Group differences in children, contrasted by ability, for word-frequency in lexical decision accuracy are now small, and just reliable. We have very low confidence in this subgroup effect.

The Hedges' g value for the adult-experience sample falls into the medium sized category with a low level of heterogeneity (0.6 [0.5, 0.71], $I^2 = 0\%$). This is the largest interaction effect amongst the five sets of subgroup contrasts, derived from the highest number of study-level effects: 12 studies ($n = 895$, median $n = 64$, range = 32 to 193). The sensitivity analysis for adult-experience effects suggested that the study effect of Balota and Ferraro (1996) is influential. This is the smallest study effect within the subgroup sample. RoB is adjudicated as high for the reporting domain of the study due to a general trend of brief reporting using p values rather than summary statistics of results. However, this study-level effect is generated from accuracy proportions and so reflects a reported result. Since Cochrane's Q was non-significant ($p = 0.679$), and heterogeneity values were in the low range, we opted to retain the study-level effect within the sample. We have very low confidence in this subgroup effect.

The median sample size for child-experience groups is 64.5, (range = 46 to 90) with a total sample size of 265. The subgroup effect is a small Hedges' g value of

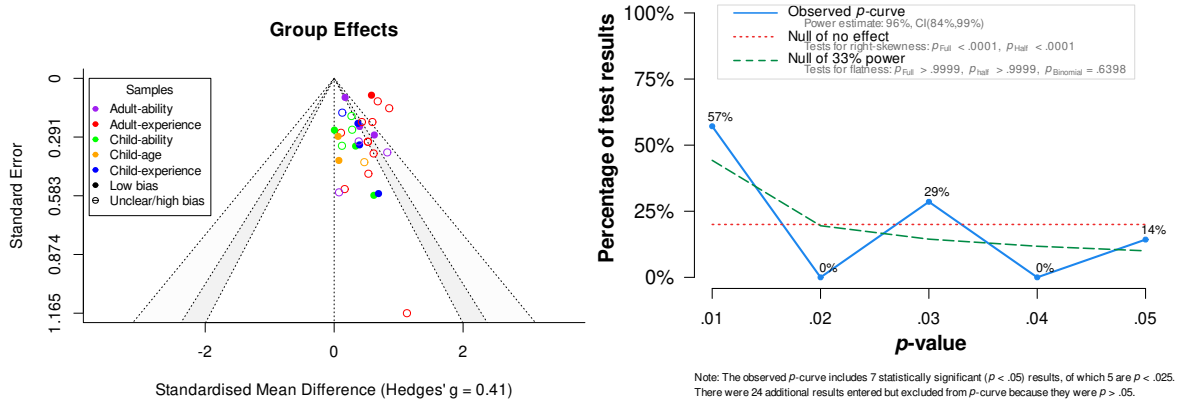
0.26 [0.02, 0.5], $I^2 = 0\%$, just on the right side of zero to be reliable. The effect from Schröter and Schroeder (2017) is suggested as influential but as Cochrane's Q indicates no heterogeneity above and beyond random sampling variation ($p = 0.66$) and the study-level effect was comfortably within the confidence interval of the subgroup effect, we retained it in the sample. We have a moderate level of confidence in this subgroup effect.

The median sample size for the child-age group is 46, (range = 22 to 75) with a total sample size of 143. Younger and older participants of similar reading skill show a very small and unreliable Hedges' g value of 0.16 [-0.24, 0.56], $I^2 = 0\%$. We have very low confidence in this subgroup effect.

One study-level effect was removed due to sensitivity analyses so we updated the global effect. There was a slight reduction in the effect size to Hedges' $g = 0.41$ [0.3, 0.52], $I^2 = 34.79\%$, however it remains small and reliable. The test for subgroup differences also remains significant ($\text{QM}_4 = 17.44$, $p = 0.002$), the adult-experience effect still being larger than all the others. No moderator analysis was completed, given the low level of heterogeneity.

Figure 4.14

Funnel Plot and P-Curve Analysis Plot for Frequency Effects in Lexical Decision Accuracy



Tests for Small Study and Publication Bias. Neither the Egger's Tests nor the Rank Correlation Test were significant for any of the subgroup effects of

word-frequency for lexical decision accuracy. The funnel plot and *p*-curve analysis plot for this sample are shown in Figure 4.14. Of the four funnel plots within the frequency report, this looks the healthiest with respect to precision of measurement, with points coalescing towards the top of the funnel (but note one adult-ability point lying at the bottom). We suggest that this is an artefact of the coarser level of measurement between letter strings at the word / nonword level combined with yes / no decisions, which then makes it easier to provide more accurate estimates over different samples and over time. The *p*-curve analysis has a significant result for right-skewness ($p < .001$) and an estimated power of 96% [0.84, 0.99]. While 24 of the 31 study-level effects were not eligible for the *p*-curve analysis, the seven that were provide evidentiary value of an effect being present in the data.

Lexical Decision Accuracy Summary. Estimates of effects of the interaction between word-frequency and group differences for lexical decision accuracy tend to be small with the exception of a medium sized effect for the difference between younger and older adults. This medium sized effect is a result of younger adults showing larger differences between responses to high and low frequency words than the older adults. The very small effect for children of different ages but similar reading skills is unreliable. The interaction between frequency and group differences are smaller for group contrasts in children than for group contrasts in adults. There was no indication of publication bias for the subgroup effects. The *p*-curve analysis estimated power to detect a true effect at 96% and gave a significant test of right skewness, suggesting that there is evidentiary value of an effect within this data set.

4.3.2.5 *Overall Summary*

There were sufficient study-level effects comparing the word-frequency effect across groups to estimate a complete set of subgroup effects for word naming and lexical decision outcomes, i.e., there are no missing data for the word-frequency predictor. We present a summary table of the subgroup effects across tasks and outcomes in

Figure 4.15. We explain the information contained within the table briefly next.

A summary table represents all subgroup effects across tasks and outcomes. Subgroup effects are presented with confidence intervals and standard error values, plus the information that contributed towards the confidence evaluation process and the confidence rating itself.

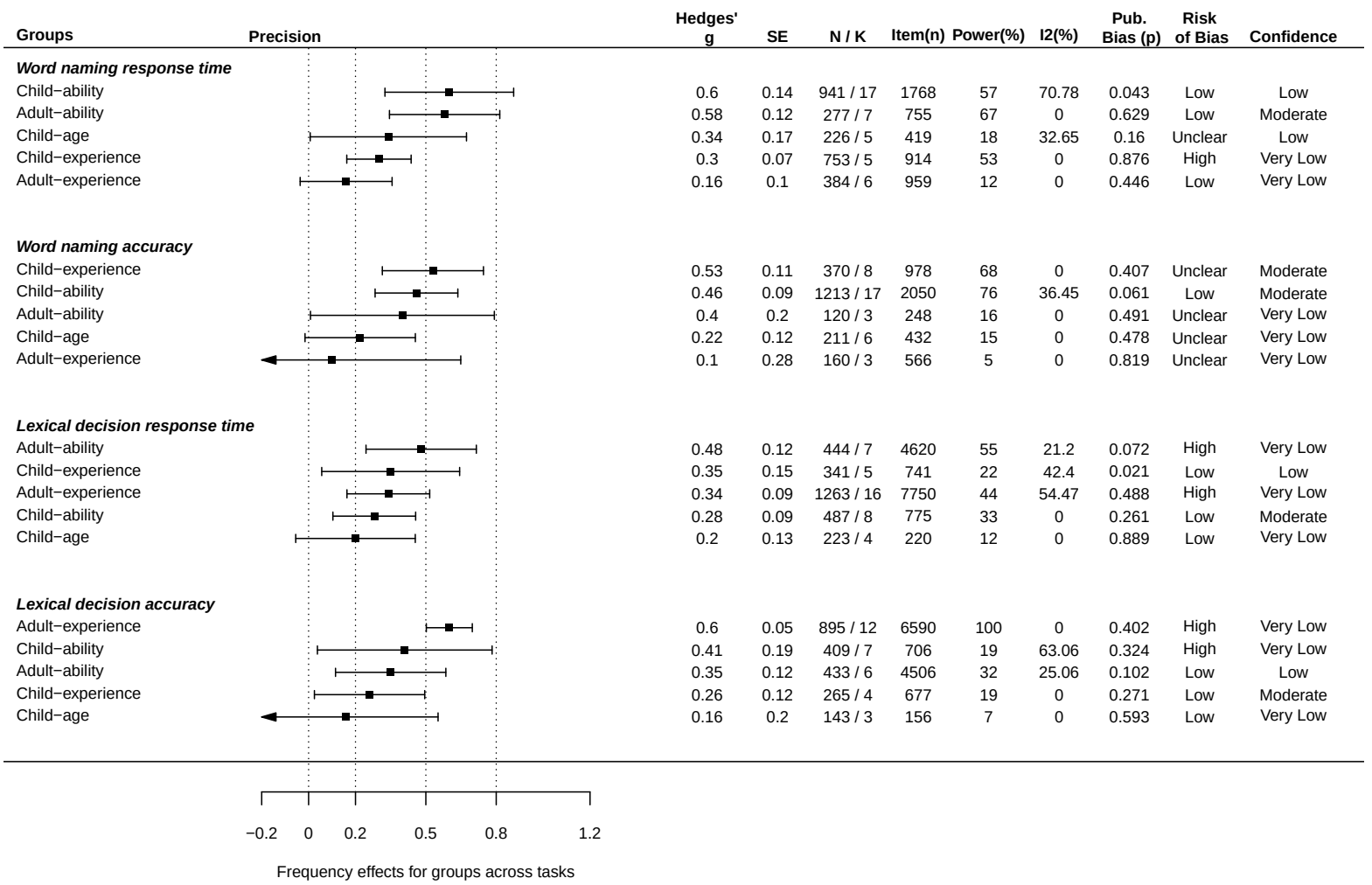
Subgroup effects are plotted as well as presented in text. Plotting the effects visualises how individual effects are distributed across the other subgroups. The value of each estimate is represented by a filled black square with 95% confidence intervals extending either side. Contrary to the earlier forest plots, each square is identical in size as there is no weighting performed here. Where a lower or upper limit of a confidence interval is presented with an arrow, this indicates that the range of that limit extends beyond the range of the x-axis. We draw dashed vertical lines on the x-axis to mark thresholds of effect sizes for Hedges' g, beginning at 0 and marking the lower limits of small (0.2), medium (0.5) and large (0.8) categories of effect size from Cohen (1988).

To the right of the plot, we include the Hedges' g value and its associated standard error in text. We then include the information that is appraised in the confidence evaluation process: the total number of participants and number of studies that generated the estimate, total number of items, power to replicate an effect of half the size, the residual heterogeneity value (I^2), the lowest publication bias p value, the RoB judgement and finally, the confidence level associated with the subgroup effect.

We briefly comment on subgroup effects for differences in the word-frequency effect across task outcomes next. We focus at the subgroup level, pulling together results of p-curve analyses and retrospective design analyses for the word-frequency interaction effects across task and outcome.

Figure 4.15

Summary of Findings for Differences in the Frequency Effect By Task and Outcome and Across Groups



Adult-Ability. Reflecting skilled readers with lower-skilled readers, differences in word-frequency effects are present in this contrast. The lower limit of the confidence interval for word naming accuracy rests on zero, with the remaining three effects showing reliability. The adult-ability estimates for reaction time outcomes are largest: word naming ($g = 0.58$) and lexical decision ($g = 0.48$); the other task outcomes present with small effect estimates. Recall that Hedges' g units are in standard deviations, so the differences in the word-frequency effects for less- and more-skilled reading adults is approximately half a standard deviation. Our confidence evaluations are moderate for word naming reaction time and low for lexical decision accuracy but very low for word naming accuracy and lexical decision reaction time estimates. This may indicate larger sample sizes are needed for greater precision in estimation. Further, in word naming outcomes the differences in adult-ability groups appears stronger than that of adult-experience groups, and the difference effect sizes appear more similar to the child contrast groups. In lexical decision, the adult-ability and adult-experience effect sizes look much more similar to each other.

Child-Ability. The differences in frequency effects between child-ability participants on accuracy appear to be equivalent across accuracy outcomes. Reaction time effect sizes are different, with a larger difference in word naming than lexical decision. We are moderately confident in the findings for word naming accuracy ($g = 0.46$) and lexical decision reaction time ($g = 0.28$). Word naming reaction time ($g = 0.6$) shows evidence for publication bias suggesting that in the presence of more data, the effect would be moderated downward. Consequently, our confidence rating is low for this estimate. We have very low confidence in the lexical decision accuracy effect size ($g = 0.41$). The plot of the effect in Figure 4.15 clearly shows that its confidence intervals span more than two effect size thresholds, and the RoB is adjudicated as high.

Adult-Experience. The adult-experience group is a contrast of typically skilled readers who differ in age. The reliable effects for this group are for lexical decision

outcomes. In lexical decision reaction time, the effect estimate is reliably small ($g = 0.34$) and for accuracy, medium ($g = 0.6$). Our confidence for lexical decision effects are low due to imprecision for replication of effects, low participant sample sizes and a high risk of bias. In word naming outcomes, estimates fall below the small threshold with confidence intervals crossing zero. Our confidence for word naming estimates is very low. Each effect describes either a very small difference in word-frequency effects for word naming, or potentially, no difference between the two groups. The effect sizes are larger in lexical decision outcomes than word naming outcomes for adult experience.

Child-Experience. There is a range of reliable differences in word-frequency effects for children of different ages but with typical reading skills for their age. The strongest estimate is of medium size, for the word naming accuracy outcome (Hedges' $g = 0.53$). Word naming and lexical decision reaction time and lexical decision accuracy show small sized estimates. Our confidence in the word naming reaction time and accuracy estimates is moderate. We have low confidence in the estimate for lexical decision reaction time as publication bias was indicated by the Egger's Test.

Child-Age. The child-age subgroup reflects a contrast of younger and older child readers who have equivalent reading skills. Essentially, the older readers are showing lower reading skills than would be expected for their age. Each of the subgroup effects is unreliable and tend to be at the lower end of small or very-small in size. The unreliability of the estimates could suggest no difference in the word-frequency effect between the groups, however, this subgroup is the least well powered of the five subgroups. We have very low confidence in these results, due to small samples for both participants and items, low power to replicate the effect and non-significant tests for out-of-sample prediction.

P-Curve and Retrospective Design Analyses. No moderator analyses were performed on word-frequency group level data for any task outcome since, after

sensitivity analyses, levels of heterogeneity were low. *P*-curve analyses showed that only lexical decision accuracy data showed evidentiary value, the remaining three outcomes failed to reject the test for right-skewness. Power to detect d^{rep} across the subgroup effects ranges from 5% to 97.9%, with only lexical decision accuracy above the desirable threshold of 80%. When we calculated the participant-per-group and item-per-condition rate for each effect, with a 40 x 40 design used as a threshold suggested by Brysbaert and Stevens (2018), one effect was above this threshold for participant sampling (child-experience group for word naming accuracy), while 17 were above the item thresholds.

4.3.3 Subgroup Estimates for All Predictors

In this final section, we present the findings of the meta-analyses of the eight variables. Of a potential 160 summary effects, we are able to present 131. Twenty-nine summary effects are missing data at the time of writing.

Twenty-four effects have a moderate confidence rating. Eleven effects have low confidence ratings with the remaining 96 effects carrying a confidence rating of very low. Moderate and low confidence ratings tended to be given to the psycholinguistic effects that are more established, having had a longer opportunity to be studied. Predominantly, moderate confidence ratings were for frequency, length and consistency variables.

Information for the summary effects are in Figures 4.16 to 4.19. The figures follow the same format as the summary figure for word-frequency presented in Figure 4.15 but now each row represents a Hedges' *g* estimate for one of the eight psycholinguistic variables. Empty rows in a figure represent missing data for that psycholinguistic variable to ensure we create a picture of our search return and to indicate gaps in our current knowledge.

As a reminder, an effect describes the magnitude of the difference between two groups in the difference of how a group responds to levels of the psycholinguistic variable, estimated with 95% confidence intervals. Additionally, confidence ratings

represent our belief of how closely the size of an estimated effect reflects the size of the true effect: 'moderate' confidence should be interpreted as *probably close*, 'low' confidence as *may be markedly different* and 'very low' confidence as *probably markedly different* from the true effect' (our emphasis).

4.3.3.1 *Adult-Ability*

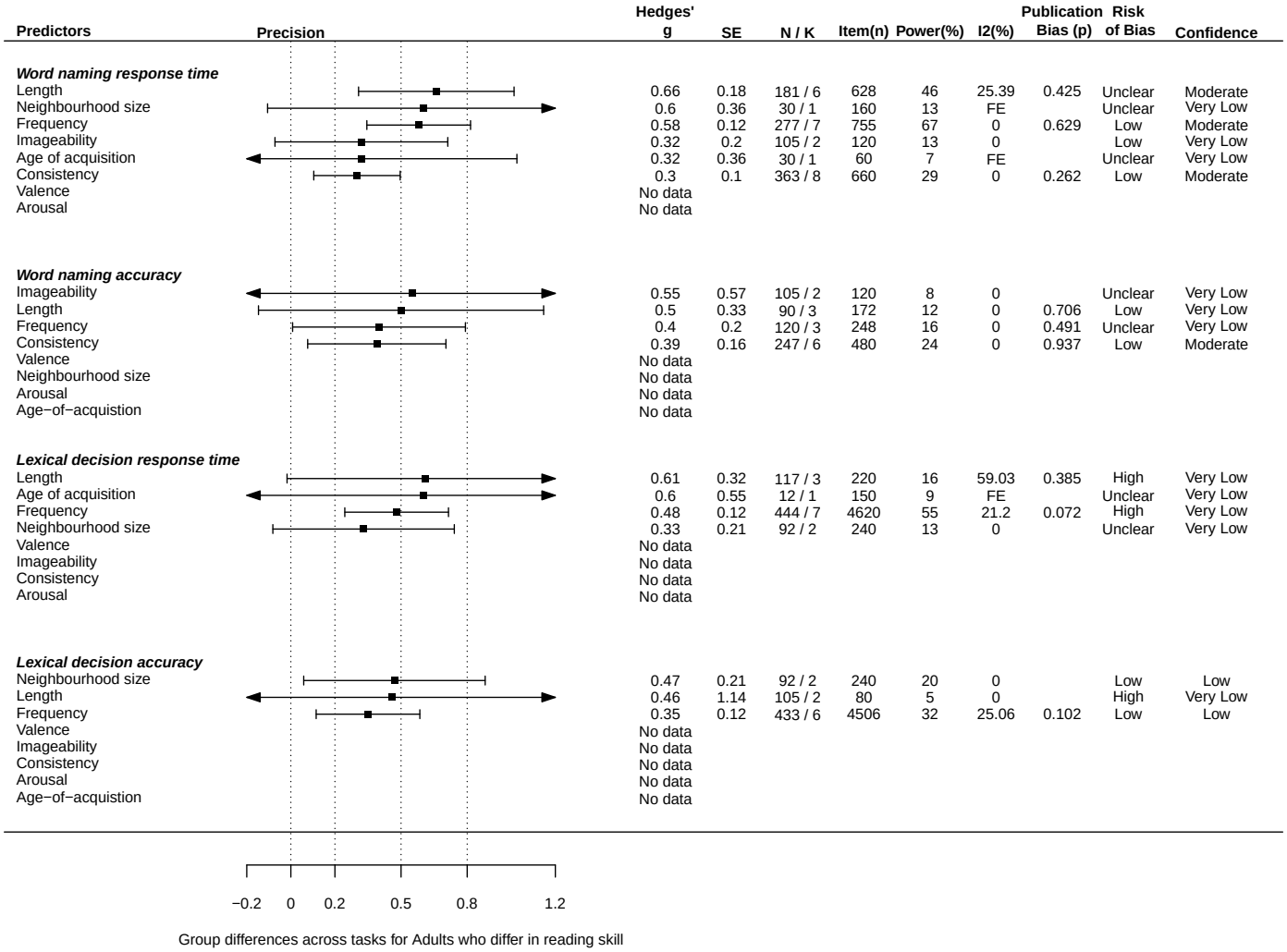
We estimated 17/32 interactions between the effect of different levels of a word property and how this differs between adults who differ in reading skill (Figure 4.16). The present effect sizes range in size from $g = 0.3$ - 0.66. Eight of the effects have confidence intervals that are reliably different from zero, and in those eight, three of those estimates remained at moderate levels of confidence. More data is needed. If we accept that, given study design (small samples, extreme group analyses) and analytical methods may be inflating effect sizes and conservatively halve the observed estimates to calculate power to replicate the effect, then range of power to replicate effects is between 5 – 67%.

Across this contrast (in both adults and children), the sampling from a population of readers with dyslexia was twice as common as for readers of lower-skill without dyslexia. It could be that the size of effects is being driven by properties specific to a dyslexic style of reading as opposed to a contrast of higher and lower reading skill. There was either insufficient heterogeneity or number of study-level effects to estimate this at the present time.

In word naming reaction time, reliable effects' estimates are available for length, word-frequency, and consistency. Effects for N-size, imageability and AoA are also estimated, however they are unreliable at this time so more data is needed to be more certain of sizes and directions for those effects. Length and frequency show medium sized effects, while consistency shows a small sized effect.

Figure 4.16

Summary of Findings for Differences in Predictor Effects for Adult-Ability Contrasts by Task and Outcome



In word naming accuracy, frequency and consistency effects are reliable. They show small sized effects. Length and imageability are estimated as medium sized effects however at this time, they are unreliable, needing more data for more precise estimation. Of the four available estimated effects for lexical decision reaction time, only a small sized frequency effect is reliable. Length, AoA and N-size are all unreliable. Length is a medium size however word naming reaction time has larger participant and item samples. The AoA effect is a fixed effect at this time since we recovered only one study-level effect that had reported measures for AoA.

Lexical decision accuracy has three estimates for differences in how groups use the variables. Two are reliable, N-size and frequency. Length is unreliable. All three estimates are small sized.

4.3.3.2 *Child-Ability*

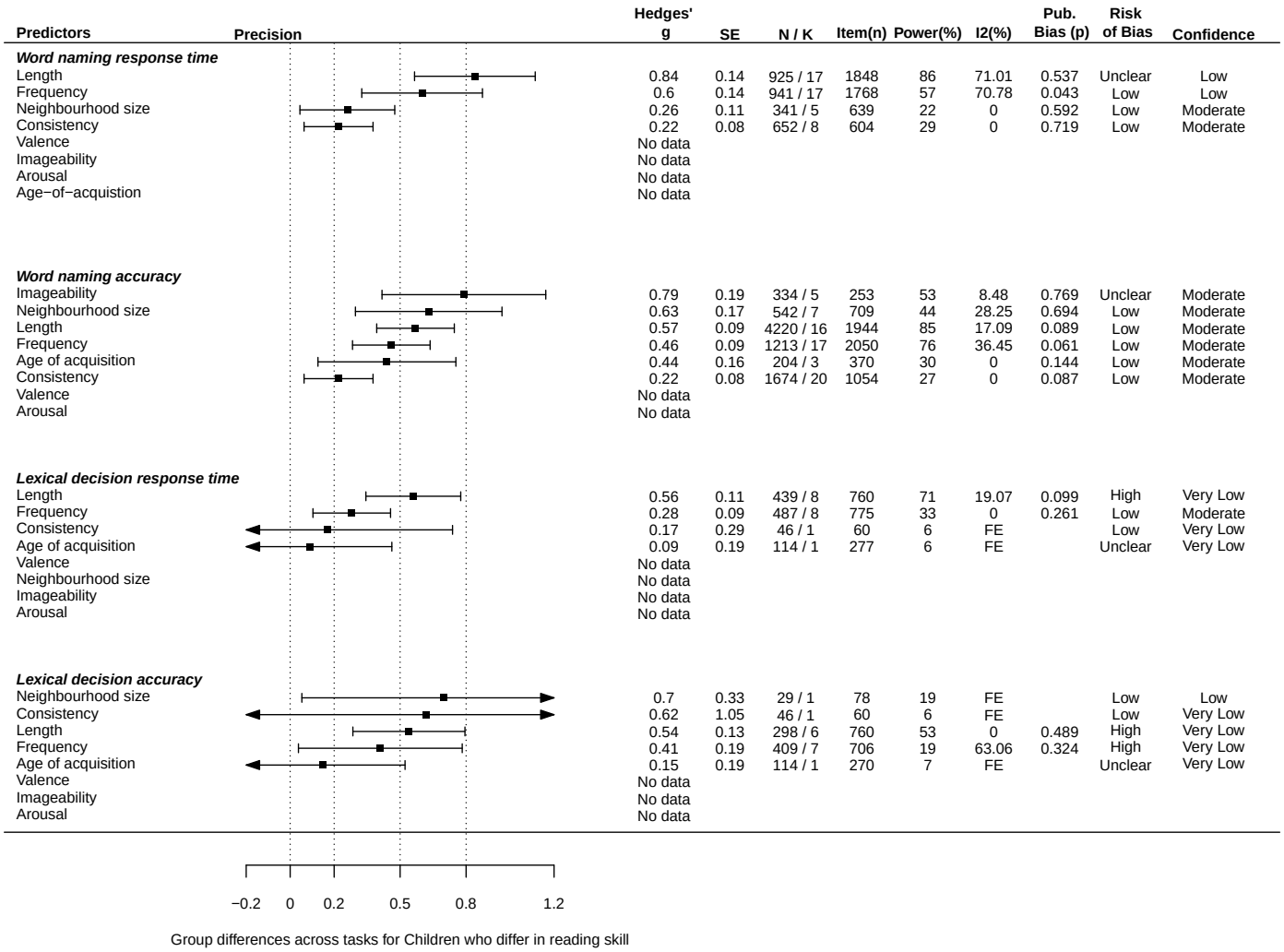
We found sufficient data to calculate 22 / 32 estimates for the child-ability contrast, of which five estimates are generated from FE models (see Figure 4.17). This contrast is one of the most frequently studied, but for a narrow set of variables. Missing data are present. Effect sizes range between $g = 0.09$ - 0.84 in size. Fifteen estimates are reliably different from zero and we have moderate confidence in nine. Power to replicate effects estimates ranges between 6% - 86%.

Four estimates are returned for word naming reaction time, all of which are reliable. Length shows a large sized estimated effect and frequency shows a medium sized effect. Confidence is low due to the I^2 values for both of these effect remaining at a high level. N-size and consistency estimates are small and reliable, and confidence ratings are moderate.

Six reliable estimates are reported for word naming accuracy, and all carry a confidence rating of moderate. Imageability, N-size and length are all estimated as medium size effects while frequency, AoA and consistency are all estimated as small size effects.

Figure 4.17

Summary of Findings for Differences in Predictor Effects for Child-Ability Contrast, by Task and Outcome



Worth noting are the two small effects for differences in consistency in the child-ability group for word naming. Other predictors show a range of effects. Given that we know readers with dyslexia were sampled twice as much as low skilled readers without dyslexia, and that difficulties with phonological information are symptomatic of dyslexic reading, we found this surprising. We split the child-ability subgroup into two smaller groups of skilled vs readers with dyslexia ($k = 7$) and skilled vs lower-skilled but without dyslexia ($k = 13$) and re-analysed the study-level effects.

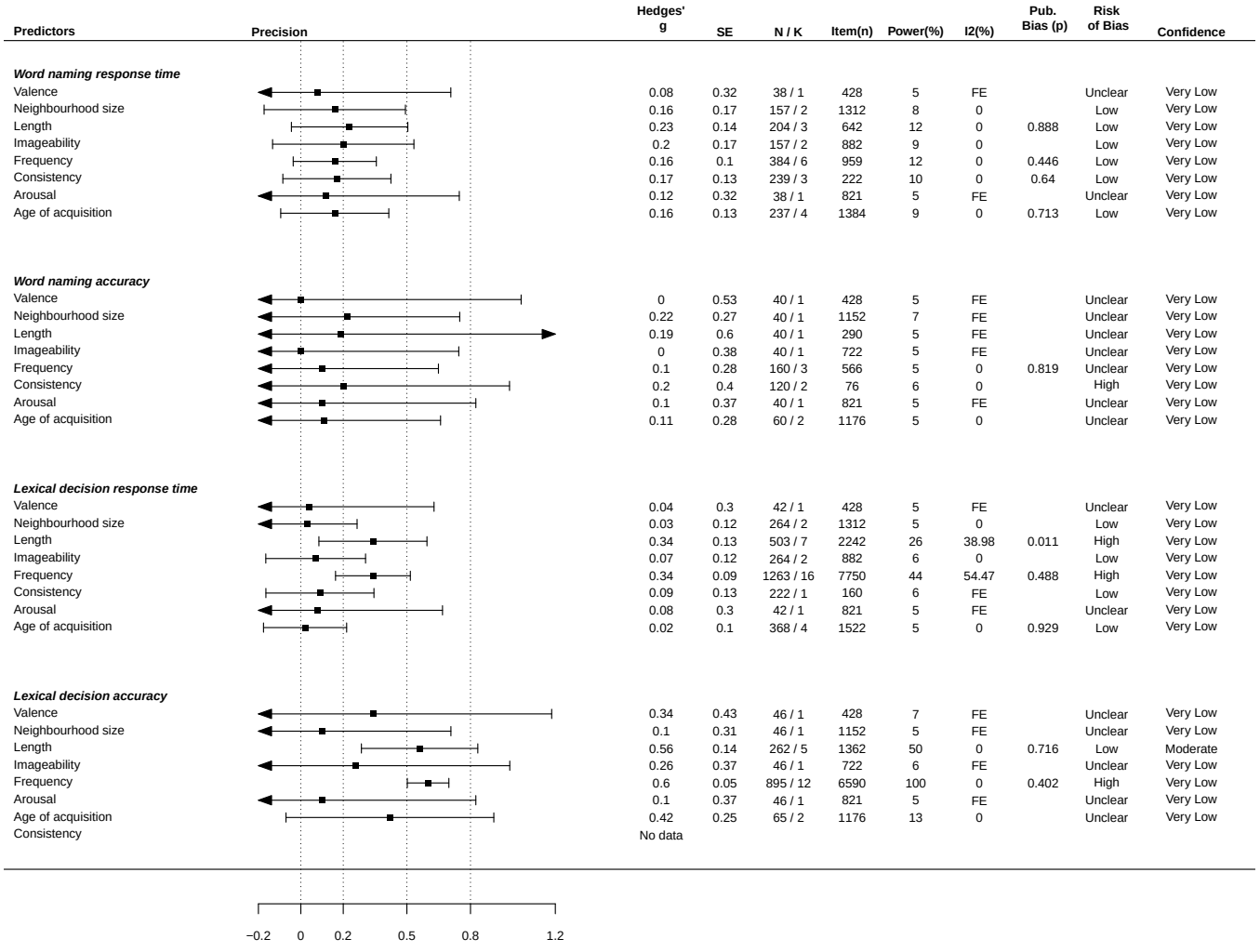
The estimate for the dyslexia contrast was small ($g = 0.37$) and the estimate for the low-skilled contrast was very small ($g = 0.16$). The original aggregated accuracy result is small due to very small effects from the low-skilled group weakening the overall effect in the dyslexia subgroup. Aggregation over the two groups has cancelled out the impact of length on the dyslexia sample. Going forward, refining the child-ability subgroup may be important for a clearer understanding of relationships.

Length and frequency have reliable estimates for lexical decision reaction time. Length is estimated as a medium size effect and frequency as a small effect. We have very low confidence in the length effect due to a significant p value on the test for publication bias and a high RoB rating. This suggests that in the presence of more data, the length effect could be smaller. We have moderate confidence in the frequency effect size. Consistency and AoA are present but unreliable at this time. Both are estimated from FE models.

Five estimates are returned for lexical decision accuracy. N-size, length and frequency are reliable estimates. N-size and length are both medium sized effects and frequency is small. We have low confidence in the N-size effect. Figure 4.17 shows the upper CI limit extending beyond the limits of the X-axis. As an effect from one single study, it is imprecise. Length and frequency effects carry very low confidence ratings. Both are adjudicated as having high RoB. Frequency is an imprecise measurement and length has a low level of participants. Consistency and AoA are unreliable at this time. N-size, consistency and AoA estimates are generated from FE models.

Figure 4.18

Summary of Findings for Differences in Predictor Effects for Adult-Experience Contrast, by Task and Outcome



4.3.3.3 *Adult-Experience*

The mean age for the younger adults is 22.4 years, range = (18.6 - 31.3) and for the older adults is 70.0 years, range = (47.9 - 86.4). Of 32 potential meta-analysis models, only an estimate for differences in the consistency variable for lexical decision accuracy is missing. The ability to make inferences is blurred, however, because almost half of the estimates come from FE models. Figure 4.18 displays the summary effects.

Generally, but for lexical decision accuracy, differences in effects are small or very small and, at the time of writing, most of the estimates are unreliable. Consequently, the majority of our confidence ratings of very low. Most estimates for the word naming task fall on or below $g = 0.2$ and none are reliable. We have very low confidence in each of the estimates. The power to replicate the effects ranges between 5% - 12%. The same pattern of findings is present for the word naming accuracy outcome. Seven of the estimated effects for word naming outcomes are from FE models.

Length and frequency estimates are measured with some reliability and precision in lexical decision reaction time, such that their lower confidence interval limits do not cross zero. Both effects are small. Length has a significant test for the presence of publication bias and a high RoB judgement, and consequently a very low confidence rating. Frequency also shows a high RoB judgement and a moderate level of residual heterogeneity within the subgroup estimate.

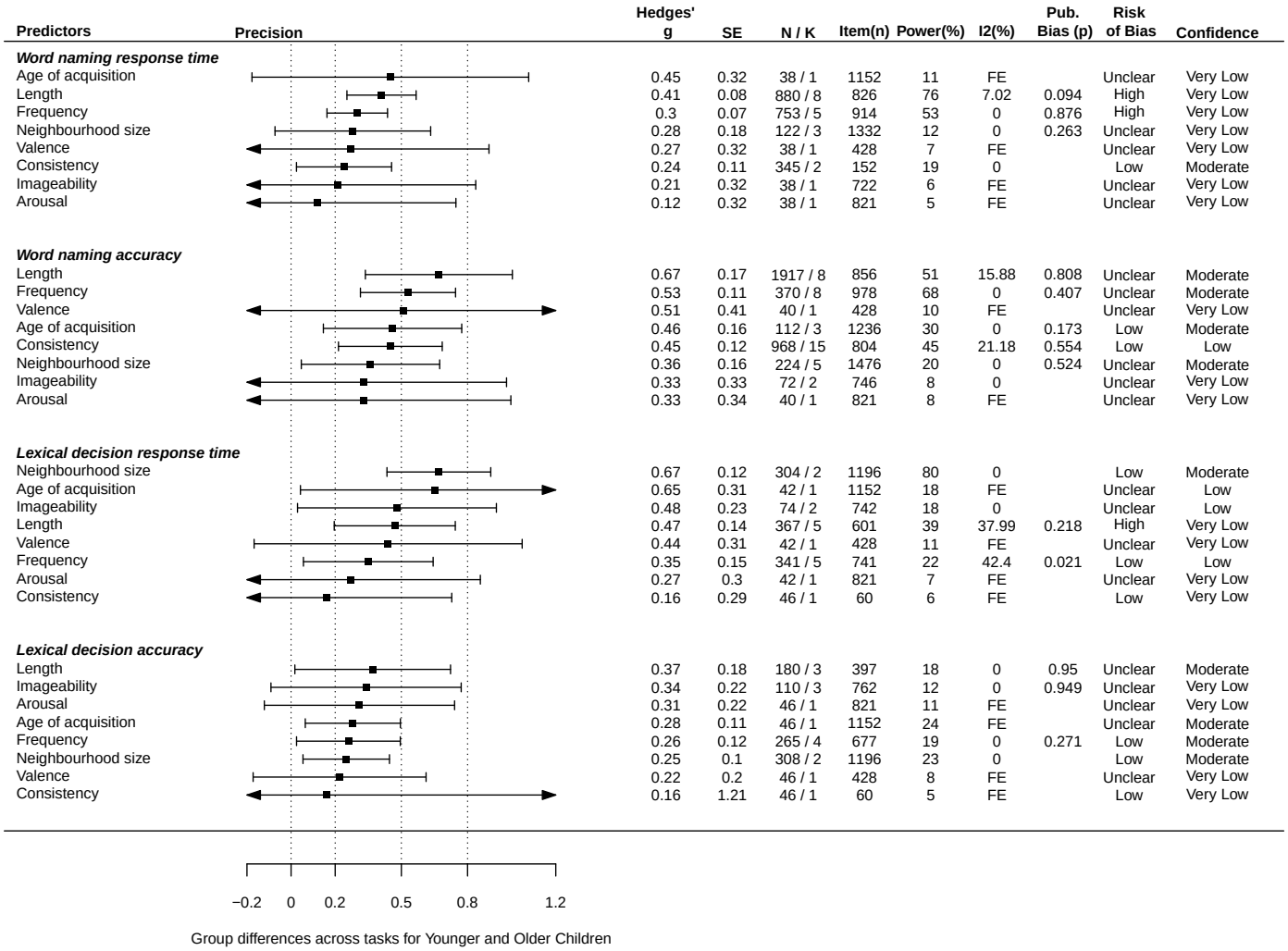
Length and frequency are also reliable estimates of effects in lexical decision accuracy, both showing medium sized effects. We have moderate confidence in the length estimate and very low confidence in the frequency estimate.

4.3.3.4 *Child-Experience*

There is a complete set of estimates available for both tasks and both outcomes (Figure 4.19). On average, the participants differed in age by approximately two-and-a-half years (mean age: younger = 8.2, older = 10.6 years), representing a reading skill difference of approximately two and a half years. Just as with the

Figure 4.19

Summary of Findings for Differences in Predictor Effects for Child-Experience Contrast, by Task and Outcome



adult-experience effects, almost half are generated from FE models which limits our ability to talk with certainty about them. Effect sizes range between g = 0.16 - 0.67. Fifteen estimates are reliably different from zero of which 10 have been given a confidence rating of moderate. Power to replicate effects estimates ranges between 5% - 80%.

Word naming reaction time has three reliable estimates. Length, frequency and consistency all show small size effects. We have moderate confidence in the consistency estimate but very low confidence in the length and frequency estimates, probably due to the judgement of high levels of RoB being present in the estimated effects. Four of the unreliable estimates are generated from FE models, so more information could quickly improve precision and alter this number of reliable measurements.

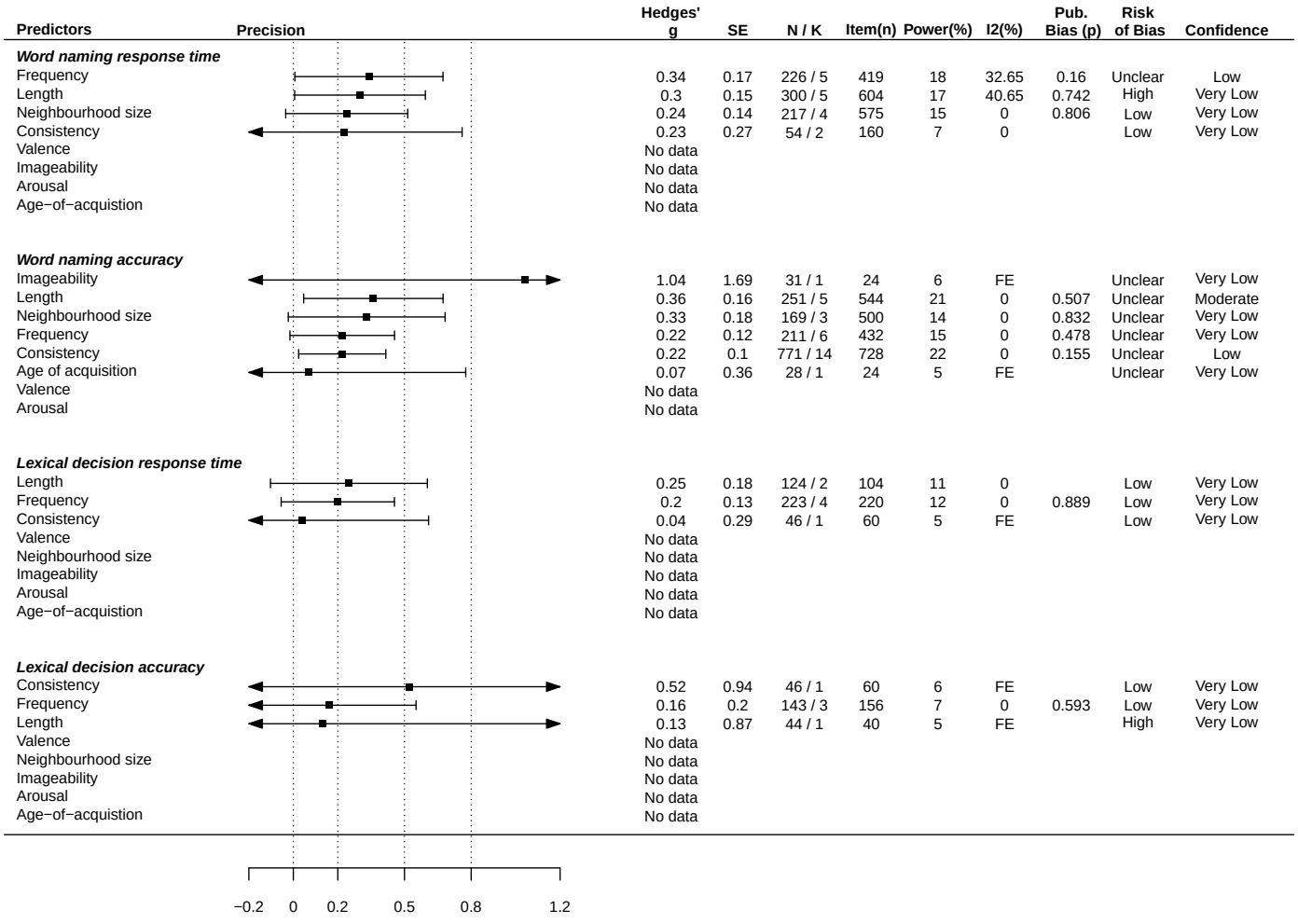
Length, frequency, AoA, consistency and N-size are all reliable estimated effects for word naming accuracy. Length and frequency show medium sized effects and show moderate ratings of confidence. AoA, consistency and N-size show small sized effects. AoA and N-size have moderate ratings of confidence, while consistency has a low rating. Valence, imageability and arousal are all unreliable at this time.

Five of the eight estimates are measured as reliable for lexical decision reaction time. N-size and AoA are medium sized effects. Imageability, length and frequency are small in size. Only N-size has a moderate level of confidence. The length estimate has been given a very low confidence rating. AoA, imageability and frequency have low confidence ratings. Valence, arousal and consistency are all generated from FE models and are estimated as unreliable at this time.

Four of the eight estimates in lexical decision accuracy are reliable. Length, AoA, frequency and N-size are all small in size and carry moderate confidence ratings. Imageability, arousal, valence and consistency are all unreliable at this time. It is worth noting that the two very small (and unreliable at the time of writing) difference effects in the whole set are for consistency on lexical decision outcomes. More data is needed, however, this could be relevant to an understanding of how developing readers approach lexical decision if differences in the consistency effect are very small.

Figure 4.20

Summary of Findings for Differences in Predictor Effects for Child-Age Contrast, by Task and Outcome



4.3.3.5 *Child-Age*

In the child-age group, younger readers' average age was 8.0 years; older readers' average age was 10.3, representing a reading skill delay of just over two years for the older participants. This contrast is often included in experiments under the assumption that there is no difference between the younger typical and older atypical, readers. A direct prediction then is that there are minimal differences in psycholinguistic variable effects between these two groups.

We were able to estimate 16 / 32 effects (see Figure 4.20), of which five estimates are generated from FE models. Contrary to an assumption of no difference, estimates range in size (range = 0.04 - 1.04), however only four are reliable. Twelve of the estimates' lower confidence interval limits cross zero. Only one estimate is given a moderate confidence rating (length on word naming accuracy). Power to replicate effects estimates ranges between 5% - 22%.

Four estimates are available for word naming reaction time. All are small in size. Frequency and length are reliable estimates and N-size and consistency are unreliable at this time.

Six estimates are available for word naming accuracy. Length and consistency are reliable estimates and both are small sized. Imageability, N-size, frequency and AoA are all unreliable at this time. Imageability and AoA estimates are generated from FE models.

Length, frequency and consistency estimates are returned for both lexical decision reaction time and accuracy outcome measures. All are unreliable and all have been given a confidence rating of very low. Length and frequency are estimated as small effects in reaction time and very small in accuracy. Consistency shows a very small sized effect in reaction time (this is the same threshold as the child-experience group) but a medium sized effect for accuracy. Both consistency effects are generated from FE models. Consequently the level of imprecision is very high.

4.4 Discussion

We conducted a systematic search to find studies that compared groups for their differences in performance for a psycholinguistic variable in word naming or lexical decision tasks. One hundred and fifty-five studies met our inclusion criteria. This set of studies yielded 472 interaction effects across a core set of eight variables for five types of group comparisons. We used meta-analytic methods to aggregate the study-level effects and presented a picture of the current state of our knowledge about the size, reliability and confidence for these summary effects.

What have we learned? Despite a large number of unreliable estimated effects, differences between groups are present for some psycholinguistic variables across some task outcomes. Study-level effects for word-frequency, length and to a slightly lower extent, consistency and N-size, are fairly well represented in the literature, such that aggregated effects can be estimated with some confidence.

The difference in the word-frequency effect does appear to be larger for lexical decision than word naming for adult readers that differ in age. In contrast, adult-ability differences for word-frequency effects across word naming and lexical decision appear to be similar. Interactions for the word-frequency effect in child readers is driven by larger differences for younger and less-skilled readers than older and skilled readers.

Differences in the length effect between groups are present for child subgroups (all outcomes and all medium sized) and the adult-ability subgroup. Verbal interpretations from the included studies suggest that the difference is located, once more in the younger and less-skilled readers, with generally, no length effect detected in the older-skilled readers. The nature of the effect is described differently across studies with some studies describing letter by letter increments and other studies describing effects at certain points in words, for instance between the third and fourth letter for younger / less-skilled readers but the sixth and seventh letters for older and more skilled readers. This suggests that ability to chunk or capacity to compress visual information develops with age or stronger skill.

Length shows a medium sized difference in effects for typically-reading younger and older adults on lexical decision accuracy. The larger length effect here is for younger skilled adults, following the pattern of the child and less-skilled adult subgroups.

When we turn to consistency, lexical decision outcomes have very few effects; word naming tasks show a higher level of representation. The effects are driven by slower and greater error for low consistency words. There is a high level of imprecision in the consistency effect estimates, however. Given its importance in verbal and computational accounts of reading, the consistency variable is a good proposition for replication studies.

For other variables, we probably have to say that the current state of our knowledge is highly uncertain. The representation of evidence is quite unbalanced between both psycholinguistic predictors and types of group. There are scattered results for AoA and imageability estimates. The most recently constructed variables, valence and arousal, returned study-level effects in child samples only. Obviously, time is a confound here. Older variables are more likely to have greater coverage and so be estimated with greater precision which gives greater probability of a moderate level of confidence. For new variables, it does beg the question of how to efficiently build an evidence base.

This point is linked to missing data points ($n = 29$) and the estimates that are generated from FE models ($n = 41$). Furthermore, gathering more information as efficiently as possible will also raise confidence in the 127 estimates that are rated either low or very low. Below we explain what we mean by efficiency and put forward an approach to future research design within the field that may focus and accelerate the collection of data for psycholinguistic variables that are currently either unreliable or under-represented.

First, estimates generated from multiple small studies convey reliability and confidence within this data set. Contrast Figure 4.19 and Figure 4.18 with Figure 4.17. The first two figures document adult- and child-experience summary effects, for which there is good coverage of predictors. In both groups, however, most of the

estimates are unreliable, with almost half of the estimates reflecting only one study-level effect and the confidence ratings marred as a result. Contrast this with the child-ability summary effects (Figure 4.17). Missing variables are present. However, where present, multiple small studies have been conducted (0/32 FE estimates), giving greater confidence in the estimates that are present (9/19 ratings are moderate). Clearly, the multiple small study approach has value.

Gathering more data is clearly mandated. *P*-curve analyses for the word-frequency effect suggested that only lexical decision accuracy contained evidentiary value of an effect. One possible solution to this is to source all single sample studies for corresponding groups and perform network meta-analyses on those study-level effects. Advances in statistical modelling and decreases in computational power costs make this plausible. This does not raise precision in the measurements, though. Across the networks, participant samples are using different sets of items, whereas here, the two participants groups within a study saw the same items. We are introducing a further level of variability, which may impact confidence ratings. At most, performing a network meta-analysis would be complementary and (hopefully) a converging line of evidence rather than additional data for this line of evidence.

Also, to perform such a review would not solve the dilemma of missing data for the newest variables. While we know that such variables exist and contribute to word recognition – but not by how much – the ability of the field to develop a richer or fuller theory for word recognition is constrained. So too, is it constrained if estimates for some groups are more present than others (compare the adult-ability and adult-experience effects). There is a risk of developing a theory that does not generalise outside of a reading behaviour for a specific group. On the contrary, having a more representative sample of data will give us clarity around how experience or skill moderate effects.

Which brings us to an entire group that is missing. Only one study included samples between the age of 12 - 18 years. Studies with contrasted groups for adolescent readers, that manipulate psycholinguistic variables, are not represented here. In the previous chapters, we detailed an adult-learner literature that was also

extremely sparse in psycholinguistic effects. We know that there are adult readers who left school without a functional level of literacy, yet we are not looking at the later school years to understand how this may happen.

Going forward, gathering *new* data is mandated. Efficiency will be improved when designs are sufficiently powered to detect interaction effects. Recall that Gelman (2018) stated that powering detection of an interaction effect needed 16 times the sample size for detection of a main effect. We must also remember that we are sampling at the level of both participants and items.

With these meta-analysed estimates in hand, researchers now have an estimate of an effect size which can inform an a priori power analysis. Brysbaert and Stevens (2018) suggested a sample of 1600 observations per condition for robust estimation of effects. Most studies explore two conditions at least, suggesting a minimum of 3200 observations, which can be thought of as 40 participants naming 80 items that represent two levels of a psycholinguistic variable. Yet this is still a single sample design. Adding a condition onto the participant sample changes the design to an 80 x 80 participants by items design for an effect size of $d = 0.3$ (Brysbaert and Stevens, 2018; Westfall, Jacob et al., 2014)[8].

To detect a Hedges' $g = 0.1$ (the smallest effect for word-frequency in the adult-experience estimates), the power of an 80 x 80 participant x item design falls to $\sim$ 17%. A massive 650 x 650 participant x item design gives a power of 81.2 %. Such levels of participants have only been observed in megastudies (Adelman et al., 2014; Balota et al., 2007; Davies et al., 2017; Schröter and Schroeder, 2017).

Is precise and robust estimation of effects for differences amongst groups for the differences in how they use psycholinguistic variables only the province of megastudies, then? To think and behave in such a way would be to lose some of the diversity and variability that is observed in the estimated effects, which give so much promise for generalisation of effects (Westfall, 2016). A systematic approach to new

[8]Westfall et al., 2014 provide a website by which to calculate sample sizes for effects in study designs used in psycholinguistic experiments: https://jakewestfall.shinyapps.io/crossedpower/; Even though advised that http://jakewestfall.org/power/ is the stable URL, that address gives an Error 404 message.

data gathering that borrows regression analysis techniques from the megastudies and capitalises on the multiple study approach is needed. We suggest a consortial approach coupled with matrix sampling of items as an efficient method by which to gather new data.

Three studies within the meta-analysis adopt a matrix sampling approach to ensure coverage of a large sample of items while participants see only a selection (Balota et al., 2007; McKoon and Ratcliff, 2016; Schröter and Schroeder, 2017). This approach is incredibly flexible, evidenced by its use with adult participants (Balota et al., 2007; McKoon and Ratcliff, 2016) and child participants (Schröter and Schroeder, 2017, see also Hsiao and Nation (2018)). We advocate the 3-form design (Graham et al., 2006). Samples are constructed of a base set, X, and three (or more) further sets. Set X contains items critical to the research hypotheses, while other sets contain items that support the hypotheses. Every participant sees set X and a number of other sets in a systematic pattern. Using this design supports the estimation of means, variances and covariances between variables. Further, by constructing samples by way of a base set of items plus an extended set of items, researchers can choose their own sets of items in the extended set but shared research goals are maintained by set X. Common items between studies will allow for the updating of effects' estimates present herein and also robust estimation of summary effects for those that are missing.

Constructing a base set that measures a selection of variables for which we need a boost of information (for instance, FE estimated effects in this meta-analysis) or to begin to estimate effect sizes (missing estimates) would accelerate the accrual of data to improve the state of our knowledge. Constructing extended sets would allow researchers to additionally focus on the variables in which they are particularly interested. Coupling design with a regression approach for analysis would allow for simultaneous estimation of study-level effects across a number of variables, improving estimation of the effects due to the increased accuracy of the estimates' standard errors.

A consortial approach would also provide for a coherent approach to the

operationalisations of experience or age or ability group construction. Ability contrasts at the group level in children are measured by > 20 different tests. Where available, and given the host language of the study, it may be preferable to reduce this set to a smaller quantity of tests. Alternatively, data sharing and explicit reporting routines would allow for standardisation for comparison across samples to be more efficiently achieved.

With the advent of powerful desktop computing resources, mixed-effects models for these conventional repeated measure designs can now be the default method. The sensitivity of these methods and their regularisation abilities could potentially improve estimation magnitude and precision, compared to aggregated values computed in ANOVA (Baayen et al., 2008; Barr, 2013; Masterson et al., 2007; Matuschek et al., 2017). The flexibility of the generalised linear model also allows modelling of accuracy outcome data within the constraints of its statistical distribution. Given that we have low power across many of the samples, adopting statistical analysis methods that maintain more of the information in the data by design can help.

A multiple regression approach to analysis benefits the field in three ways. First, regression methods conserve more of the information within a data set, and consequently statistical power, because the data is not aggregated as it is in analysis of variance methods. Second, each psycholinguistic variable recovered in the systematic search is claiming to have influence on word recognition processes. Consequently, it is a stronger and more honest test that the effect of each psycholinguistic variable be estimated alongside each other. Partialling out the influence of covariates will give a more precise estimate for the primary variable.

We are sure that psycholinguistic variables' effect sizes will be attenuated under these conditions, however we think it represents a refinement of current practice that will give us greater confidence in our findings for the field as a whole. Finally, by including all the psycholinguistic variables, we are systematically and regularly gathering data across the variable set, thereby strengthening the evidence base for the (un)reliability of each candidate psycholinguistic variable.

Matrix designs with base sets / extended sets allows large amounts of data to be collected in the small controlled study setting that seems preferred by the researchers featured here. Like minded researchers could be running many small studies in parallel while working towards a common goal. With the communication networks and computational infrastructure that exist today, acceleration of knowledge production is well within our reach.

4.4.1 Future Directions

We fervently hope that the availability of the meta-analysis and its findings at the project OSF repository provides a resource and reference point for researchers in the field for planning future studies. The missingness of the ability estimates, particularly for adult-ability, looks to be an intriguing vein of research. A further area for consideration could be a systematic plan to include samples of older children. Only a handful of studies included child samples between 12 -18 years of age, so few that we could not realistically include them as a group here.

There are a couple of immediate recommendations that are cost-free and easy to implement and would make estimation of study-level effects for the purpose of meta-analysis much more reliable. First, reporting condition level means *and standard deviations* in descriptive statistics sections of research reports. With these sufficient statistics in hand, constructing effect sizes and updating these estimates is trivial. Second, we ask that all variables and all findings are explicitly reported in the research report. We can see that our decision to impute missing variable results at the level of $p = .1$ could have introduced a systematic bias into findings. Where variable findings are reported in full, irrespective of whether or not it was a null finding, there is no need for this practice.

In the short to medium term, we think that the issue of missingness can be efficiently redressed by asking people to perform multiple regression analyses that include all the psycholinguistic variables recovered in this meta-analysis. There is a slight cost to the researcher here in resourcing the additional variables and engaging

in regression methods rather than choosing ANOVA, but the gains to the field of research as a whole would be worth it.

In the medium term to long term, we advocate a consortial approach as detailed above, where researchers network their research design, materials and data analysis. Working together in this way is being used very successfully by projects such as the Psychological Accelerator and BabyDev. The promise of a networked approach together with matrix sampling techniques and standard operating procedures for task administration and analysis, all of which can reduce the signal to noise ratio, maximises power to produce estimates with greater precision and build confidence in future results.

4.4.2 Limitations

We imputed hypothesised effects that were not described or completely missing from results sections of eligible studies at a uniform value of $p = .1$. We are aware that this may have introduced a source of bias but chose completeness for this first look at the state of the field. Very few of the Egger's Tests or Rank Correlation Tests were significant as a result of this. Finally, research reports needed to be available in the English language, because of the first author's monolingual status and available resources. The opportunity to include studies presented in languages other than English is not lost, however, as we envision updating the meta-analysis. Researchers who know of such studies are eagerly invited to contribute and participate in the collation of those data.

4.5 Conclusion

We presented the results of a wide ranging meta-analysis detailing the differences in psycholinguistic variable effects on single word recognition task outcomes for groups of different ages and ability, for which the data, code and complementary reports are available at the project OSF repository. We have found that the variables frequency,

length and consistency are the most populated estimates and are more likely to yield non-zero and reliable differences between groups, and that more recent and less studied psycholinguistic variables are yet to achieve results in which we can have confidence. Adults are more likely to be contrasted by age and experience and children are more frequently explored as a function of ability differences. Adult-ability groups show some strong estimates, however, that suggests they are worthy of further attention and research. We have also found that studies, on the whole, lack statistical power at the level of the group interaction effect for which the studies are explicitly designed. We have suggested immediate and longer term approaches to methods in the field of word recognition research to raise confidence in estimates and strengthen our capacity for inference making over typical and atypical reading processes.

5 Longitudinal Study

The meta-analysis collected study-level effects across eight psycholinguistic variables for five types of sub-groups. The resultant summary effects gives us an idea of what we may expect in exploratory work with adult-learners.

To summarise from an adult-learner perspective: we may expect to find that adult-learners show word frequency, consistency and length effects for word naming reaction time and frequency and consistency on accuracy measures, that may differ from both their skilled reading adult peers but also younger readers.

In lexical decision accuracy, we may expect the word frequency effect to be more like younger readers. For the length variable, we may expect to find differences on reaction time measures. They may perform similarly to younger atypical readers. Adult-learners may resemble other readers for the impact of length on accuracy measures.

We may see small effects of N-size on lexical decision reaction time, however it is also possible that we will observe no difference, since the summary effect estimate was unreliable. For all other predictors that featured in the meta-analysis there is either no differences detected between subgroups or missing data.

5.1 The Present Study

Research using individual differences (ID) measures on adult-learners (chapter 2) suggests a tendency for adult-learners to be a) different in their reading-related skills than their adult peers and similar in skill level to younger, typical readers, b) stronger in their semantic skills than their orthographical or phonological knowledge. We know very little about how psycholinguistic variables may vary as a function of reading-related skills in adult-learners. Very few studies have measured individual

differences in the context of psycholinguistic variables (McKoon and Ratcliff, 2016).

Additionally, very few studies report participant samples in the 12–18 years of age range. An assertion that adult-learners read like 11-year-olds may be tinged with confirmation bias, since there is no explicit test of how they compare to readers in this extended age range. Adult-learners could compare more favourably with typically-reading adolescent participants.

We conducted a longitudinal study to answer this question: are there qualitative or quantitative differences in psycholinguistic predictor effects between adult-learners and other readers that are similar or different in their reading skills? Our first research aim was to estimate the variation in the impact of a range of psycholinguistic variables in the context of individual differences across a range of groups. We included six groups, with our primary focus being an adult-learner group (hereafter atypically-reading adults). We compared them to typically-reading adults as a peer group comparison. We also recruited two groups of 11-12-year-old participants as the research literature suggests that this is a reading-age match group of the atypically-reading adults. Finally, we also invited two groups of 16-17-year-olds to take part. The typical-readers of this group are a reference group as they represent a level of reading skill to which the atypically-reading adults aspire. We asked participants to complete a battery of tasks to assess person-level, reading-related skills.

We also asked participants to complete four experimental tasks. The tasks were chosen to represent a progression in processing of visual information from sublexical level to sentence level of print recognition. Differences on any of the tasks for any of the groups may locate a source of difficulty or advantage that contributes to an explanation of variability in performance. We briefly describe them here.

First, a letter search task. This task demands fast processing in serial or parallel fashion and good letter recognition skills to be able to respond quickly and accurately. With the use of unpronounceable nonwords, the task can also indicate levels of processing of letters without lexical activation (Ziegler et al., 2008). Slow reaction times or inaccurate responding may indicate that knowledge and learning at a letter level is under-developed (Mason, 1978). Inefficient letter processing may

indicate a reduced capacity for assimilation of statistical information on the distribution of letters within a language, thus slowing or truncating orthographic knowledge development (Kirby et al., 2010; Perfetti and Hart, 2002).

Allen et al. (1991) showed that skilled-reading, older adults were slower to identify letters within words than younger readers. Chetail (2017) manipulated letter search in the context of high and low frequency bigrams and found that reaction times were facilitated for high frequency bigrams with no effect on accuracy in skilled, young adult readers. Horn and Manis (1985) and Ziegler et al. (2008) showed that letter search accuracy was lower for less-skilled child readers than typical child readers with no difference in speed. A weaker performance may occur in lower-skilled readers, or older readers with higher frequency words improving response time.

In contrast, strong performance may indicate that processing is highly tuned for sublexical information, which may mean processing in parallel at a word level may be underdeveloped.

A lexical decision and a word naming task were also completed. Differences between the two tasks are useful for exploring locations of effects. For instance, lexical decision explicitly examines lexical and sublexical components of word recognition with its mixture of word and nonword items (Balota and Chumbley, 1984). Word naming tasks may not need sublexical processing, where a word is known to the participant, the orthographic form of the word may be sufficient for recognition. Lexical decision may not need the identification of the entire word for a response to be made, while word naming needs the specific word item for its correct pronunciation (Andrews, 1997). Finally, lexical decision has no requirement for a pronunciation, while word naming does. Consequently, the estimates for phonological variables required across the tasks may differ. Balota et al. (2004) found a phonological component across both tasks for skilled readers, and interpreted this as phonological recoding processes being engaged for all types of items in the two tasks.

The fourth task that participants completed was a cloze, sentence reading task (Bruck, 1990). Inclusion of a sentence reading task using cloze procedure allows us to measure, at a crude level of detail, how word reading speed and accuracy fares

in the context of other words. We manipulated the context of the surrounding words, to be meaningful, neutral or isolated (i.e. single word recognition with no surrounding words). For readers who are able to use meaningful sentential contexts, we may expect word recognition to be faster and accurate. For readers who depend upon semantic information to a greater extent (as in the division of labour hypothesis, Plaut, 1996), we would expect variables of the semantic domain to play a greater role in word recognition for this task.

We were also interested to see if a rate of change in ID measure' scores for atypically-reading adult participants was a) detectable and b) commensurate with other readers across the duration of their course. To this end, we used a longitudinal design and collected data across three time points within a calendar year. The study is exploratory in nature. To the best of our knowledge, this is the first study to explore atypically-reading adults' word recognition processes using individual difference measures and multiple psycholinguistic variables.

5.2 Method

5.2.1 Participants

The study collected data from participants over three data collection sessions, with the number of days varying between data collection times and participants. For simplicity, we refer to the three data collection points as T1, T2 and T3. In statistical analyses, we model time using the number of days passed per data collection session at the participant level (H. Goldstein, personal communication, June 19, 2019). Also, for simplicity, we refer to reading groups who are demonstrating age-appropriate reading skill as "typical" and those that are slightly below age-appropriate as "atypical" in either 11-12- or 16-17-years old or adults.

At each data collection session, participants completed a battery of ID measures and four experimental tasks. Recruitment and testing procedures were approved by Lancaster University Research Ethics Committee. Typically-reading

adults were recruited from the local community and atypically-reading adults from FE college GCSE English classes. Participants in the atypically-reading 16-17-year-old group were recruited from GCSE English classes running in the same FE college. Typically-reading 16-17-year-old readers were recruited from a local sixth-form provision within a secondary school. Typically- and atypically-reading 11-12-years-olds were recruited from two local secondary schools. Permission to conduct the study was obtained from institutions. Duty of care was adopted by one school due to the nature of the tasks being very similar to school activities with parents asked to explicitly opt *out* of the study. The other school asked parents to opt in by giving explicit informed consent. All participants gave informed consent at the beginning of each data collection session.

There were 218 participants tested at T1. At T2, 191 participants returned and at T3, 173 participants. All participants reported no history of learning disorders during their secondary school experiences, although some of the atypically-reading adults reported anecdotally they experienced some difficulties as they completed the tasks. All participants reported normal or corrected-to-normal vision.

5.2.1.1 *Eleven-Twelve-Year-Olds*

Eighty-three 11-12-year-old readers (39 females) took part. The average age of participants was 11.8 years (SD = 0.3, range 11.2-13.2 years). Each school sent letters to parents for children who scored between 90 and 110 on the GL Assessment CAT4 Verbal Reasoning test. Individual students who scored between 90-99 were assigned to the atypically-reading 11-12-year-old group ($n = 40$). The remainder were assigned to the typically-reading 11-12-year-old group ($n = 43$). As a thank you for taking part, participants in this group were given a raffle ticket at each time point of data collection. The raffle was drawn at the end of data collection with 10 prizes of £20 vouchers awarded.

5.2.1.2 *Sixteen-Seventeen-Year-Olds*

At T1, 69 16-17-year-old participants took part (33 females). The average age of participants was 17.1 years (SD = 0.8, range 16.2 - 20.2 years). Atypical readers (n = 43) were approached through their English GCSE classes. Enrolment in such classes is due to achieving less than a pass at the first attempt of their English GCSE. Typically-reading 16-17-year-old readers who achieved a level 4 in their English GCSE ($n = 26$) were recruited through a college welcome day and through a year group assembly. Any interested student was given an information sheet and participant consent form. At the outset of the study, there were age differences between the groups. The atypically-reading 16-17-year-olds were slightly older than the typically-reading 16-17-year-olds ($M_{typ} = 16.8$; $M_{atyp} = 17.4$, $t(63) = 3.65$, $p < .001$). Students who took part in the study were entered into a raffle for a £200 prize voucher.

5.2.1.3 *Adults*

Sixty-six adults, aged 20 years and above (44 females) took part. The average age of participants was 43.0 years (SD = 15.8, range 19.7-78.6 years). Atypically-reading adults were approached through their GCSE English classes ($n = 38$). Typically-reading adults ($n = 28$) were recruited through word of mouth, local newspaper adverts, and two days recruitment in local shopping centres. As such they represent a self-selected group of people. At the outset of the study, there were age differences between the groups. The atypically-reading adults were younger than the typically-reading adults ($M_{typ} = 56.4$; $M_{atyp} = 33.1$, t (48) = -8.28, $p < .001$). The differences in age was carried forward to the third data collection point between groups for those present ($M_{typ} = 58.1$; $M_{atyp} = 34.9$, t (44) = -7.36, $p < .001$). Adults who took part in the study were entered into a raffle for a £200 prize voucher.

5.2.2 Data Collection

Data was collected between October 2017 and October 2018. Data collection sessions for school and FE students took place in either the school library or an empty classroom. Data collection for adult participants took place either at college during lesson time or at the participant's home, whichever was most convenient for the individual. Each session took approximately 50 minutes. ID measures were completed in the same order at each time point. Experimental task order was counterbalanced within group and participants across time by way of standard latin square design. After the third data collection session, participants were thanked for their participation and debriefed.

5.2.3 Measures

Six ID measures were used, repeated across data collection sessions. Four experimental tasks were used with different items within tasks at each time point.

5.2.3.1 *Individual Difference Measures*

Word Reading. Participants read Form A of the Test of Word Reading Efficiency Sight Word Efficiency Test (SWE, Torgesen et al., 2012). Over 45 seconds, the individual reads aloud as many of the 104 test items as accurately as possible. Words at the beginning of the test are higher in frequency than words towards the end of the test. The measure is the number of words read correctly in 45 seconds. Faster readers who complete the 104 test items have their actual time recorded. Standard scores for the SWE are only available for 6:0 - 24:11 year olds, so we used raw scores as measures. A word reading skill measure is constructed by dividing the number of words read correctly by the time taken.

Time sampling error rates for same form administration are available for 8-18-year olds with a resting period of two weeks. For a sample of 8-12–year-olds,

test-retest reliability was alpha = .90; for a sample of 13–18-year-olds, test-retest reliability was alpha = .84.

Nanda et al. (2010) tested 108 native English speakers and 88 English-as-a-second-language speakers who were engaged in adult education classes. Test-retest reliability correlation with an approximate delay of four months between testing sessions was .84 for the English speaking atypically-reading adults.

Nonword Reading Skill. Participants read Form A of the TOWRE-2 Phonemic Decoding Efficiency test (PDE, Torgesen et al., 2012). Over 45 seconds, an individual reads out loud as a many of the 63 nonword test items as accurately as possible. The length and complexity of phonemic structure of the test items increases through the test. The test score is the number of items read accurately in 45 seconds. As with the SWE, we used raw scores. Readers who complete the 63 test items have their actual time recorded. A nonword reading skill measure is constructed by dividing the number of nonwords read correctly by the time taken.

Time sampling error rates for same form administration are available for 8-18-year olds with a resting period of two weeks. For a sample of 8-12–year-olds, test-retest reliability was alpha = .91; for a sample of 13 – 18 year olds, test-retest reliability was alpha = .90.

Nanda et al. (2010) also completed test-retest reliability for the PDE on the same sample as listed above and found an alpha coefficient of .78.

The SWE and the PDE have a correlation of .83 for performance in samples for which it is designed, indicating that they may be measuring the same underlying construct or ability.

Phonological Awareness Skill. Participants completed the Phoneme Isolation (PI) subtest from the Comprehensive Test of Phonological Processing - second edition (CTOPP-2, Wagner et al., 2013). Over 32 items, an individual listens to a whole word and then identifies a target sound within that word. This is an untimed test. The items at the beginning of the test are CVC items and highly consistent in

sound-spelling relationships. More complex words are introduced as the items progress. Early in the test, individuals identify beginning and end sounds in words with middle sounds introduced later. The most difficult items have more letters than sounds. Individuals cannot rely on a visual strategy for these words and must engage with some parsing of phonemes to identify the correct target sound. All items were administered. The measure is the number of items answered correctly.

Time sampling error rates are available with periods between testing varying from 1 to 2 weeks. For a sample of 12 – 18 year-olds, test-retest reliability was alpha = .67.

This subtest has a phonological composite correlation score with the SWE of .41 for a sample of typical readers of ages 11-20 years ($n = 384$). The phonological composite correlation with the PDE for the same sample is .25 (Wagner et al., 2013). These low correlations indicate that the PI test may be capturing a different skill than the word and nonword reading tasks.

Processing Speed. We use the Rapid Object Naming (RON) test from CTOPP-2 (Wagner et al., 2013) to capture any general processing speed differences that may underlie reading differences within the participant sample. The *object* naming test is chosen as opposed to letter or digit naming task to measure general processing speed rather than verbal processing speed. The object naming test was preferred to the colour version of the test because people see colour differently, which could introduce a confound (Kirby et al., 2010).

Pictures of six objects are randomly repeated across four rows of nine objects. Participants name the objects in succession, beginning on the top left hand corner, reading each row as quickly and accurately as possible until reaching the last item in the bottom right hand corner. The measure is the total time taken in seconds to finish naming all 36 items. We constructed a skill measure by dividing the number of objects correctly answered by the time taken to give an object / second measure. Test-retest reliability alpha coefficient, estimating error due to time sampling for this subtest is alpha = .86.

Vocabulary Knowledge. The Shipley-2 Vocabulary Scale is a 40 item test (Shipley, 1940). It is designed to measure crystallised knowledge of vocabulary for individuals between the age of 7:0 – 89 years within a 10 minute time period. Early test items are higher in frequency than later test items. Reading skills of 10-years of age are assumed for independent completion, however in the present study, if participants asked for any words to be pronounced for them, they were.

Each of 40 target words has four possible answers listed against them. Individuals must circle one of the four possible answers that shares a similar meaning with the target word. The measure is the number of correctly identified words within the time limit. Although there are standard scores available for ages 7-89 years, we use raw scores as measures to align with other task measures.

Time sampling error rates are available with periods between testing varying from 1 to 2 weeks. For a sample of teens to adults, test-retest reliability was alpha = .89.

The Shipley-2 Vocabulary Scale is known to correlate well with other reading-related measures. The correlation with the Wechsler Individual Achievement Test - Second Edition (WIAT-II) word reading subtest is .79. These are moderate correlations which suggest that word reading and vocabulary may be measuring aspects of the same underlying construct.

Spelling Knowledge. Participants completed the WIAT-II spelling subtest (Wechsler, 2001), items 27 - 53. Item 27 is the recommended basal level for 11-12-year-olds. Items increase in spelling complexity, sampling from a wide range of orthographic patterns. Participants hear the target word, hear the target word in a sentence and hear the word once more before recording their responses. The measure taken is the number of correct answers with a maximum score of 27. All participants had the opportunity to answer all items. At T1, group administration was possible due to the consistent starting point. After T1, administration to the 16-17-year-old and adult participants was individual, while administration to the 11-12-year-olds remained in groups.

Group Contrasts. We created a planned contrast variable such that specific group differences of primary interest were estimated rather than estimates comparing all groups to a reference group (Schad et al., 2020). We planned the following:

- Atypically-reading adults vs typically-reading 16-17-year-olds
- Atypically-reading adults vs atypically-reading 16-17-year-olds
- Atypically-reading adults vs typical 11-12-year-olds
- Atypically-reading adults vs typical adults
- Atypical 11-12-year-olds vs typical 11-12-year-olds

The first two contrasts reflect the gaps in our knowledge about how atypically-reading adults may compare to 16-17-year-old readers. The third contrast seeks to confirm findings of prior studies where atypically-reading adults have shown reading-related skills similar to those of typically-reading 11-year-olds. The fourth contrast is a check that atypically-reading adults are different from typically-reading adults.

We chose to contrast the two 11-12-year-old groups for two reasons. First, if the atypically-reading adults were the same as the typical 11-12-year-olds, and the two 11-12-year-old groups were found to have no discernible differences, we could also assume that there was a low probability of difference between atypically-reading adults and atypical 11-12-year-olds. Second, explicitly testing for differences between the two 11-12-year-old groups may prove useful for the institutions from which the samples were taken, given that there are some more useful years by which an influence through teaching could be explored.

Measures of Time. Time between data collection points varied within and across participants. For instance, for one participant, T1 $\rightarrow$ T2 = 81 days and T2 $\rightarrow$ T3 = 86 days, giving 167 days from T1 $\rightarrow$ T3. This was the shortest data collection period in the data set. The longest data collection period was T1 $\rightarrow$ T2 = 214 days and T2 $\rightarrow$ T3 = 135, giving 349 days from T1 $\rightarrow$ T3. The variability within the time measure could confound any estimates of change, so needs to be included in the model (H.

Goldstein, personal communication, June 19, 2019). The first data collection session for each participant was designated as day 0. The variable rate at which data was collected is captured by two variables: 'Days' as the number of day since day 0 and also 'Age' calculated in years (to 2 decimal places).

5.2.3.2 *Psycholinguistic Measures*

Item Sampling. We selected items across high and low frequency values from SUBTLEX-UK (Van Heuven et al., 2014), collecting measures for multiple psycholinguistic variables at the same time. We also collected measures from N-Watch (Davis, 2005). Four lists were constructed from the item sample of which three were eventually used in the study.

Properties of items for each experimental task are detailed within relevant results sections. List presentation was counterbalanced within group and participants across time by way of standard latin square design, such that by the end of three data collection sessions, a participant would have seen all items for all tasks within the sample. The sampling process for construction of three lists for all tasks is detailed in Appendix D. Items are listed by task in Appendix E.

AoA. Taken from Kuperman et al. (2012), these AoA ratings capture the age in years at which a sample of American participants reported that they first remember understanding the word when somebody used it in their presence. AoA and frequency values tend to show a strong, negative correlation with each other. Across the four tasks, the range of correlations between AoA and frequency is letter search r = -.82, lexical decision and word naming r = -.74, sentence reading r = – .60, all $ps < .001$.

Arousal, Dominance, Valence. Ratings for affect are taken from Warriner et al. (2013). These ratings are collected on a 1 to 9 scale. A rating of 1 indicates high levels of positive valence and arousal and low levels of being controlled or dominated. A rating of 9 indicates high levels of negative valence, low levels of arousal (e.g. 'sleepy'

or 'sluggish') and high levels of controlling behaviour. Citron et al. (2014) reports that many studies that explore affective variables find effects as interactions with other item-level variables. Since the models are not planned to include interactions between item-level variables at this early stage of exploration, they are included here as part of the model as emotion words have been shown to influence word recognition in lexical decision (Estes and Adelman, 2008; Kuperman et al., 2014).

Bigram Frequency. Bigram frequency is a measure of orthographic structure that occurs at the sublexical level (Gernsbacher, 1984). A presence of a bigram frequency effect could support an interpretation that readers process words above the letter level but beneath the whole-word level (Hofmann et al., 2007). The occurrence of adjacent pairs of letters within words are counted within a corpus by type (how many types of words) and token (how many occurrences of each type of word). We collected measures of both.

Bigram frequency (type) is an average of the number of words that share a bigram in the same position with the target word. The number of words for each bigram at each position is added together and divided by the number of bigrams within the word.

Bigram frequency (token) is the summed frequencies of the bigrams across the letter string, divided by the number of bigrams.

We also collected values for Mean log bigram frequency. This is an average value made by summing the logarithmic frequencies of words that share bigrams with the target word and dividing the sum by the number of bigrams in the word.

We extracted type and token values from SUBTLEX-UK and mean log bigram frequency values from the CELEX database via the N-watch program (Davis, 2005). Muncer et al. (2014) found that summed bigram frequency measures were probably the best predictive measure of bigrams for lexical decision and word naming tasks.

Concreteness. The extent to which a person considers the word to refer to some perceptible concept that can be experienced by one of the five human senses is measured by concreteness ratings. A rating of 1 indicates an abstract word; a rating of 5 indicates a word that can be experienced directly through senses or actions. In this study we use values taken from Brysbaert et al. (2014).

Consistency. We constructed a rime consistency ratio using the *friends / friends + enemies* ratio used in Adelman and Brown (2007). We used the 'RegExp' function of the online SUBTLEX-UK search facility (Van Heuven et al., 2014) to count the number of friends for an item: monosyllabic words that have identically spelled rimes to the target words *and* maintained the vowel sound; the number of enemies: monosyllabic words that share the same rime spelling with different pronunciations. Number of friends was then divided by the sum of friends and enemies to produce a rime consistency rating that ranges between 0 and 1. A higher value indicates that the target word has more friends than enemies. A higher value indicates a greater degree of consistency of pronunciation for that rime spelling, conditional upon the SUBTLEX-UK lexicon.

Word-Frequency. We collected words that represented high and low frequency values from the CELEX database (Baayen et al., 1995), SUBTLEX-UK Zipf frequency scale (hereafter Zipf scale, Van Heuven et al., 2014) and the Contextual Diversity scale (CD, Van Heuven et al., 2014), settling on the Zipf frequency values as the best measure (see Appendix D for our evaluation of the four measures).

During a variance inflation factor process (see section 5.2.5.5) CELEX written and spoken frequencies were identified as having very high VIF values and were recommended for removal. CD and Zipf values showed similar VIF values. While CD values have been demonstrated to explain more variance than raw frequency counts (Adelman et al., 2006; Brysbaert et al., 2016), values for items across different lists gave significantly different standard deviation values. The Zipf scale showed equivalent variance values across the lists. For this reason we settled on the Zipf scale

as our frequency rating.

The Zipf scale is a log scale of frequency. The scale ranges between 1 and 7 with low values representing words of very low frequency and high values representing words of high frequency.

Imageability. Imageability values capture the extent to which a word will conjure a mental image. Values are taken from Cortese and Fugett (2004). Undergraduate participants were asked to rate the ease (7: high level of easiness) or difficulty (1: low level of easiness) with which a mental image came to mind on presentation of a word.

Length. Serial reading processes may be in evidence if length of letters or phonemes is influential in the models. We might expect serial reading processes to be in evidence where items are unfamiliar, challenging or in readers that are younger or lower in reading skill. We collected length as number of letters and also number of phonemes. The correlation between the two measures in word naming and lexical decision (same items) $r = .61$, $p < .001$, and for sentence reading $r = .75$, $p < .001$. During the VIF process, length was identified as having a high inflation value. Consequently, number of phonemes was used as a proxy for length in the statistical models (Morrison et al., 2003).

N-size. We took measures of Coltheart's N (Coltheart et al., 1977) from N-Watch (Davis, 2005). We also included orthographic (OLD) and phonological Levenshtein distance values (PLD, Yarkoni et al., 2008) from the English Lexicon Project (Balota et al., 2007). This is the mean number of substitutions, deletions or insertions of letters or phonemes needed to turn the target word into one of its 20 nearest neighbours. Following Yap et al. (2015) we constructed a ratio of OLD and PLD values to make a Levenshtein phonological consistency (LPC) measure. LPC values closer to 1 indicated greater consistency for the spelling to sound mapping of a word.

Sensory Experience Ratings. Taken from Juhasz et al. (2011). The sensory experience rating (SER) variable is designed to measure how much a word evokes sensory or perceptual experience. Undergraduate participants were asked to rate words between values of 1 (no sensory experience) to 7 (strong sensory experience). SER may be a complement to imageability as it has been shown to correlate at a moderate level ($r = .46$).

Polysemy. Words with many meanings are considered to incur a processing cost as multiple potential meanings are activated that must be reconciled, compared to words with fewer or a single meaning. We use two measures:

Number of Word Meanings. Following de Vaan et al. (2007), items were entered into the WordNet search engine (Miller, 1995). Counts for different senses of each item across nouns, verbs, adjectives and adverbs were recorded and summed. This represented a word meaning set. The distribution of the numbers of meaning for items within the data set was skewed by a couple having very high totals. Two items did not have WordNet entries. Consequently, a constant value of 1 was added to each item's sum value and a log transformation across the item set performed to reduce skewness.

Semantic Diversity. Semantic diversity is a continuous measure, assuming that meanings vary with use, dependent upon the language by which it is surrounded and the context (Hoffman and Woollams, 2015). A large value indicates that the item is found across a broad range of contexts. A word that occurs in fewer contexts is represented by a smaller value. It is inferred that the number of definitions for a word of higher semantic diversity will be greater for a word of lower semantic diversity.

Phonetic Onsets. To statistically adjust for systematic bias in our models introduced by initial phonemes and phonetic onsets (Kessler et al., 2002), we made several dummy variables that indicated the presence or absence for placement

categories of initial letters. There are nine categories: alveolars, bilabials, fricative, glottals, liquid, nasal, palatals, velars and voice.

5.2.4 Experimental Tasks: Procedure

All tasks were administered on a Windows 10 laptop with a 17" screen. Items were presented in black 28-point Times New Roman font on a grey background. Each task was administered with DMDX software (Forster K. and Forster J., 2003). Voice responses for the word naming and sentence reading tasks were captured by a Microsoft LifeChat X-3000 headset with integrated microphone and saved for offline processing with the CheckVocal software (Protopapas, 2007). Button presses in the letter search and lexical decision trials were captured using the left (yes trials) and right (no trails) trigger buttons on a Logitech Gamepad F310. Participants sat approximately 50 cm from the screen. Outcome measures for all tasks are accuracy and reaction time. Analyses are conducted on words only (except for word superiority models), and reaction time analyses are always for correct trials only.

5.2.4.1 *Letter Search*

Items. Following Ziegler et al. (2008), items were pairs of 72 words and 72 unpronounceable nonwords of five letters. Unpronounceable nonwords were used rather than pronounceable nonwords to reduce the use of implicit knowledge of transitional probabilities between letters that we assume would be higher for a typical, skilled reader and may confer an advantage for accurate, speeded responding. Words were balanced for low and high frequency of occurrence. Target letter identity and position were balanced across six letter positions: not present, and first position through to fifth, and matched across word and nonword pairs (e.g. "O" in "w*o*uld vs. "v*o*cbs").

The nonwords were created using the WUGGY Pseudoword Generator Software (Keuleers and Brysbaert, 2010). We generated 10 unpronounceable nonwords for each word. Selection of nonwords from each set of 10 was by random

number generation. The researcher scanned the nonword list to check for pronounceability of items. If present, a new random number was generated for the set and a new nonword sampled. This process was repeated until there were 72 unpronounceable nonwords. The 72 word-nonword pairs were divided into three lists of 24 pairs. Four pairs were used for practice trials leaving 20 pairs per list for data collection. In each list, the target letter was absent for 10 pairs of items, and present for two pairs of items at each letter position.

Procedure. The letter search task always preceded the lexical decision task, with a short break in between tasks. A trial started with the presentation of a target letter in upper case (e.g., "O") in the centre of the screen for 500 milliseconds (ms) followed by the item. Initial presentation of the isolated target letter was presented in upper case and the letter strings in lower case to decrease the potential use of visual matching strategies for making decisions. The item remained on the screen until a response was detected or 2000 ms passed at which point the next trial began. Participants responded by way of a key press for 'yes' or 'no' as to whether the target letter was present. No feedback was given. Participants saw all items in a single block with trial presentation randomised between participants. There were eight practice trials (4 words and 4 nonwords) before beginning the task.

A task specific variable for this task was 'position' with six levels (none; 1-5). The reference level is "none".

5.2.4.2 *Lexical Decision and Word Naming*

Items. We selected 150 words for use in lexical decision and word naming tasks. There were an additional 60 words from the sentence reading task isolation condition (described below). Three lists of 50 + 20 words were made with a balance of low and high frequency values and word length. To avoid practice effects, a participant saw different lists for word naming and lexical decision within the same session.

For lexical decision nonwords, we generated 150 nonwords paired with the 150

words using the WUGGY Pseudoword Generator Software (Keuleers and Brysbaert, 2010). We generated 10 pronounceable nonwords for each word matched by initial phoneme and OLD measures. Selection of nonwords from the set of 10 was by random number generation. Words and nonwords were mixed together in the lexical decision task, however blocks within the task ensured that paired words and nonwords were never encountered in the same block.

This gave a total of 120 trials for lexical decision tasks and 70 trials for word naming tasks per session. Lexical decision trials were presented in three blocks of 40 trials. Word naming trials were presented in two blocks of 35 trials. Presentation of trials within blocks was randomised and order of block presentation randomised between participants by the DMDX software. The letter search task was presented as the first block of trials in the lexical decision task at each time point to each participant. Participants were invited to take self-paced breaks in between blocks.

Procedure. Presentation of items for both tasks was the same. Each trial began with a fixation point (*) displayed in the centre of the screen for 500 ms followed by presentation of the item. In word naming trials, the participant was asked to name the item as quickly and accurately as possible. In lexical decision trials, the participant was asked to press one of two keys to respond either "yes" if they believed the item to be a word or "no" if they did not believe the item was a word. In both tasks, the item remained on screen until either a response was detected or 4000 ms passed, at which point the next trial began. No feedback was given. There were 10 practice items in lexical decision tasks (5 words and 5 nonwords) and 5 practice items in the word naming task.

5.2.4.3 *Sentence Reading*

Items. There were 60 words altogether, 20 per session, repeated across three conditions, to make 60 trials per session. Words were balanced for low and high frequency values. The conditions were: isolation, meaningful and neutral contexts.

Words in the isolation condition were presented in the word naming task. The "meaningful" and "neutral" contexts had associated stem sentences. For example, for the target word "school", a meaningful stem sentence could be "I used to play for my football team at…" while the neutral context was always "The next word in this sentence is…". The final word appeared on the screen on its own and participants were asked to name the final word as quickly and accurately as possible.

Procedure. Each trial consisted of a presentation of the stem sentence for 2500 ms, followed by the presentation of the target word to be named. The target word remained on the screen until a response was detected or 4000 ms had passed at which point the target word was cleared from the screen and the next trial began. No feedback was given. There were six practice trials before the task began.

Trials were presented in two blocks with a self-paced break halfway through the task. A paired meaningful and neutral trial occurred in separate blocks. The order of trial presentation was randomised within each block and the order of blocks was randomised between each participant by the DMDX software. A predictor variable specific to sentence reading was 'Task' with three levels (isolation, meaningful and neutral).

5.2.5 Data Analysis

5.2.5.1 *Attrition and Missing Data.*

Attrition. It is important to consider the ID measure profiles of participants who withdrew from the study after T1. If the profiles of participants who withdrew are significantly different from the continuing participants, this constrains the ability to generalise finding around how groups may differ as a function of individual differences at T1 and also from experimental task findings. It may also imply that withdrawing participants left the study because of their reading-related skills which is important when we consider the mechanisms for missing data.

We separated the T1 data of participants who withdrew at either T2 or T3 from the T1 participant data who remained. We then conducted a series of *t*-tests for these subsets to determine if the participants who subsequently withdrew were significantly different on ID measure scores than the participants who remained.

Missing Data. The data analysis strategy involves model comparison techniques. Consequently, within the same analysis strand, we need each of the models to be based upon the same data. This mandates a complete case analysis for the longitudinal data (Long, 2011). Complete cases in this context means, no missing data at the predictor level rather than no missing data at the task outcome level. For instance, some participants partially completed ID measures within a data collection session due to fire alarms or sports day activities.

In preparing the data, we calculated the rate and potential mechanisms for the missing data on ID measures. We used data imputation techniques to estimate missing data values (Gelman et al., 2021; van Buuren, 2018). If a participant was missing data for outcome measures on an experimental task, they were not included in the analysis for that task. Full details for our missing data process is in Appendix F.

5.2.5.2 *Individual Difference Measures*

We use the participants' raw data from T1 ($n = 218$) and conduct tests to detect statistically significant differences between group averages on ID measures. Our null hypothesis is that there are no differences on test scores between groups. For each measure, we performed a one way ANOVA with the ID measure as the dependent variable and group as the independent variable to detect differences in scores across the six groups. We reject the null hypothesis where the p value $< .05$ and move onto post hoc tests to determine the pairs of groups that show significant differences.

We test ANOVA model assumptions with Levene's test for homogeneity of variance (`car` package, Fox and Weisburg, 2019) and the Shapiro-Wilk test for normal distribution of model residuals. Where either the Levene's test or the Shapiro-Wilk's

p values $< .05$, we repeat the analysis using Kruskal-Wallis tests with a pairwise Wilcox test (`rstatix` package, Kassambara, 2023) for post hoc, multiple comparison analyses, using the Holm's method to correct the family wise error rate.

We also describe performance for the ID measures over time. We use spaghetti line plots (Figures 6.5 -6.6) to display the variability in participant performance within groups. We also describe whether differences identified at T1 hold over time.

5.2.5.3 *Cluster Analysis*

Although nominally, there are six labelled groups in the data set, this does not necessarily mean that their performance on the ID measures is different. There may be statistically non-significant differences between group performance as they are currently labelled. The observed data may contain a correlated structure that suggests a different arrangement of participants as groups. A test of this is a cluster analysis, a data-driven approach that will confirm the best solution for numbers of groups and their composition.

We performed a cluster analysis in two stages. In the first stage we used Hartigans' Dip Test (Hartigan and Hartigan, 1985) for unimodality from the `clusterability` package (Adolfsson et al., 2019). The null hypothesis is that the sample under investigation arises from a unimodal distribution. In this simple test, a p value $< .05$ suggests a departure from unimodality and that more than one sample distribution may be present in the data.

In the second stage, we sought to confirm the findings suggested by the Dip Test using the partitioning around mediods (PAM) method. The PAM method identifies observations around which a substantial quantity of other observations are gathered with minimal distance. The identified observation is the *mediod* around which the other data form a cluster. In contrast to other cluster analysis approaches, such as k-means, the PAM method identifies actual observations from the data set as the representative mediods. As PAM is not dependent upon using means' values for

calculations, it has the advantage of being more robust to outliers and can model continuous and non-continuous variables.

As a measure of agreement between the distribution of the participants across groups in the data set and that suggested by PAM, we calculated the randIndex value (in `flexclust` package, Leisch, 2006). The randIndex ranges from -1 to 1 on a scale of no agreement to perfect agreement, respectively. A high value would therefore indicate that we have good agreement between labeling of participants in groups in the data set and identifiable clusters of observations that match those numbers.

We segregated data by data collection point, inputting standardised values for word and nonword reading skill, spelling and vocabulary scores. We did not include phonological awareness nor processing speed measures as the analyses (section 6.2.3 and 6.2.4) suggested that all groups performed to the same level on these measures. We investigated whether the data supported six clusters (to mirror the group labels) or three clusters (to reflect age bands).

5.2.5.4 *Experimental Task Data*

We expect differences within and across groups for the ID and psycholinguistic measures. Across individuals and across data collection session, there may be different rates of change for which we need to take account. There is also structure within the outcome measures data arising from task design. Within data collection sessions, we have repeated observations within participants between items and within items between participants, representing a crossed effects design (Long, 2011). These repeated observations introduce a dependency within a participant's or an item's outcome measure data (Baayen et al., 2008). Additionally, due to the longitudinal design, we have potential for within participant variation across time. We need to adjust estimates for the variation within participant and item to be able to infer that the fixed effect coefficients are not inflated due to repeated measures of the data. Additionally, the participants and items within the study represent a subset of a much larger sample, to which we may like to generalise the findings. Using mixed-effects

models for our analysis strategy will help us account for all of the above.

Mixed Effects Models. We estimate coefficients for independent and interaction terms of fixed effects for ID and psycholinguistic measures. We account for variance between participants within items and between items within participants by including random intercepts and slopes for participants and items. The inclusion of random effects terms will give greater precision to the estimation of coefficients for the fixed effects (Baayen et al., 2008).

We account for random sampling variation over time by also putting ID measures on the participants random effects term. Specifying the ID measures as random slopes on participants will account for the correlation within individual participants over time for any variation on these measures (Long, 2011).

Specifying a time-related variable (for example 'age' or 'days' as a measure of days passed) as a fixed effect will capture any independent effects of time. Adding a time-related variable as a random slope on participants is also warranted within any statistical model to account for the "time-unstructured" nature of the data set (Singer et al., 2003). As the period of time between successive data collection sessions varied between each occasion and each participant, adding a predictor to capture these values will assist for any differential rates of change between participants.

With multiple predictors for both participants and items and with crossed random effects estimating variance at the participant and item level, the quantity of model parameters for estimation is massive. To complicate matters further, some of the variables correlate moderately or strongly with each other, creating potential estimation difficulties due to multicollinearity. This can pose a problem for estimation of effects and convergence of models in the frequentist paradigm of statistical models. Bayesian inference methods offer a complementary approach to the frequentist method.

Practically, Bayesian models can be more robust to convergence difficulties than frequentist models, given that they implement Monte Carlo Markov Chain (MCMC) sampling methods to return a posterior distribution (McElreath, 2020).

Consequently, we estimate effects using frequentist and Bayesian inference methods. There are some differences between the two approaches which we discuss briefly next.

Differences in Interpretation for Frequentist and Bayesian Models. Estimates from Bayesian and frequentist models arise from different modelling assumptions that affect interpretation. In frequentist analysis, the current data set is but one of many data sets, past and future with the model coefficients representing point estimates of those imagined models (McElreath, 2020). The associated standard error for each coefficient represents measurement error.

Bayesian data analysis assumes only the sample data, in the context of explicit prior information and the statistical model. Rather than returning only point estimates, the coefficient is a value within a range of plausible values, conditioned upon the prior information, the data and the statistical model (Gelman et al., 2013).

A posterior distribution is calculated which describes a range of plausible values for the estimate, with the point estimate representing the most plausible value, occurring at the 'peak' of the distribution. Values that are closer to the peak of the distribution are more plausible than values that are far away from the peak, however all are plausible, given the model, the prior information and the sample data. The *width* of posterior distribution indicates a level of certainty for the range of values.

Bayesian Models and Prior Information. We described above how Bayesian inference models are dependent upon the model and the sample data. The third necessary component is prior information. Prior information is the explicit coding of knowledge that is already known into the model. There are levels of prior information – uniform priors, weakly informative priors and strongly informative priors. In writing that Bayesian inference approaches are complementary to frequentist approaches, this is true with respect to uniform priors. A null hypothesis frequentist model is equivalent to a Bayesian inference model conditioned upon a uniform, flat prior.

By specifying informative priors, a researcher can input a range of values that suggest a more plausible parameter space for exploration by the MCMC algorithm.

These can be left as default values, or effect sizes taken from prior research that represent the state of the current knowledge. Where possible, we use the summary effect sizes from the meta-analysis as priors.

Prior information is programmed into the model and the MCMC sampler to have reference to this distribution during the warm up phase of sampling (Gelman et al., 2013). The posterior distribution is then a combination of the prior information, the posterior likelihood and the data. In large data samples, it is likely that the strength of the observed data will minimise the influence of the prior information on the posterior distribution.

We programmed two kinds of prior information in our Bayesian inference models: strong and weak. Weakly informative priors are taken from the Stan manual (Stan Development Team, 2022) and strongly informative priors are taken from the meta-analysis project.

5.2.5.5 *Modelling Strategy*

Analyses are conducted for word items only. Reaction time data for correct trials is modeled using general linear mixed effects models. Accuracy data is modeled using generalised linear mixed effects models, specifying the `logit` link function and, for Bayesian inference models, the Bernoulli distribution.

For each task, outcome measure and for Bayesian inference models, type of prior information, we build six models[1], using a nested modelling strategy. To be clear, this means that the larger models contain the smaller models and are estimated on identical data (see Appendix H for an aide and information criterion values). We describe the general structure of the models below:

1. Base Random Intercepts Model (Base-RI): This model includes fixed effects terms for number of days passed, group, ID measures and age. In letter search there is a predictor for position of letter. In lexical decision there is a predictor

[1] in both frequentist and Bayesian methods.

for word status. In sentence reading there is a predictor for context condition. Random intercept terms are modelled for participants and items.

2. Base Random Intercepts and Slopes Model (Base-RIS): Identical to the Base-RI model with the addition of ID measures as random slopes on both participant and item random effect terms.

3. Additive Random Intercepts Model (Additive-RI): This is the Base-RI model with the addition of fixed effects terms for the psycholinguistic variables.

4. Additive Random Intercepts and Slopes Model (Additive-RIS): This is the Base-RIS model with the addition of fixed effects terms for the psycholinguistic variables and the addition of psycholinguistic variables as random slopes on participants. This model is the design implied model, with all predictors and all random effects terms included.

5. Interactions and Random Intercepts Model (Interaction-RI): This is the Additive-RI model except that the fixed effects interact with each other, giving four way interactions between days passed, group, ID measures and psycholinguistic variables.

6. Interactions and Random Intercepts and Slopes Model (Interaction-RIS): This is the Additive-RIS model except that the fixed effects interact with each other, giving four way interactions between days passed, group, ID measures and psycholinguistic variables. Random slopes terms remain at the level of independent predictors. Interaction terms are not entered as random slopes.

These models include a group contrast predictor, to reflect the quasi-experimental study design that is based upon a sample containing six groups. Yet the data-driven approach of the cluster analysis indicated that the data do not support distinct groups. Consequently, a further set of models was constructed, omitting the group contrast predictor.

Model Selection. Subject to successful convergence, we inspect information criterion values for each model. For frequentist models we use the Akaike-information-criterion (AIC, Burnham and Anderson, 2004) and for Bayesian models we use the Pareto smoothed importance sampling leave-one-out-information-criterion values (LOOIC, Vehtari et al., 2017). The model with the smallest value is presented as the preferred, most compatible model for this data and sample.

Models in the frequentist paradigm are built using the `lme4` package (Bates et al., 2014); Bayesian inference models are built using `brms` (Bürkner, 2017). All data cleaning, wrangling, modelling and plotting are conducted in the R statistical computing environment (version 4.1.0; R Core Team, 2022). Models were run on the High End Computing Cluster facility at Lancaster University. Scripts and data for each tested model are available from the author.

Data Cleaning and Transformations.

Reaction Time Data. We removed observations that occurred below 200 ms (102 observations across all tasks) and above 4000 ms (5 observations). A response below 200 ms is considered too fast to be a valid response and is likely an error. A response that registers above 4,000 ms is a clear malfunction of the equipment since the timeout value was set for 4,000 ms.

We log10 transformed reaction time observations to ameliorate skew in the raw RT distribution and to assist with the linear model assumptions of normally distributed residual values in the frequentist models.

We also created a standardised reaction time variable using the typically-reading 16-17-year-old group as our reference group (Long, 2011). We calculated the mean and standard deviation of reaction time data for the typically-reading 16-17-year-olds for each condition in each task at T1. We used these values to standardise reaction times across task, time and condition for all groups. Consequently, the intercept for reaction time models represents the mean time taken

to respond to items for an average typically-reading 16-17-year-old participant at T1 for that task and outcome and reference levels for other categorical predictors.

Accuracy Data. Every observation of an accuracy outcome arises from a trial where a response can either be correct (coded as 1) or incorrect (coded as 0). This binary response scale motivates the use of logistic regression models. Within the modelling process, the binary response outcome data (0, 1) is transformed and coefficients are expressed on a continuous log-odds scale. Essentially, coefficient estimates will express higher and lower odds of a response being correct with a negative coefficient indicating a lower probability than 50% of a response being correct and correspondingly, a positive coefficient indicating a higher probability than 50% of the response being correct. We present log-odds coefficients, transforming onto a probability scale for verbal description and interpretation.

Predictor Variables. We applied a log10 transformation to continuous predictors of AoA, bigram frequency, days, N-size, and number of word meanings, principally because a few values within predictor values were very high, thus creating a skewed distribution.

It is often advised that entering continuous variables in standardised form helps with interpretation of interaction terms (Baguley, 2012). Standardising variables will give the intercept coefficient a meaningful definition of continuous predictors at their average value. Standardising will also facilitate comparison between coefficients.

Just as with the reaction time data, we used the typically-reading 16-17-year-old mean and standard deviation values by which to standardise each ID measure. Anchoring the average in this way maintains differences (if any) between groups within a time point but also between groups across time points (Long, 2011). For item-level measures, we standardised predictors using the mean and standard deviations per predictor for the item sample set within the task.

We were conscious of problems of estimation arising from multiple collinearity. We ran an automated variance inflation function (using the `vif()` from

the `car` package, Fox and Weisburg, 2019) to check which predictors may be very influential. Setting our threshold at 7.5 (midway between 5-10, the values recommended by Hair Jr. et al., 2017), the CELEX range of predictors, length, mean log bigram frequency, the Zipf and the SUBTLEX-UK contextual distinctiveness scale all showed VIF values > 7.5.

Given this information, we chose the Zipf scale above the CELEX and CD frequency measures as it showed the lowest VIF value and was balanced across all lists of items for equivalent mean and standard deviation values (see Appendix D). We removed length as a predictor variable, choosing to use number of phonemes as a proxy measure for length (Morrison et al., 2003). Bigram frequency (token) and Mean log bigram frequency were also indicated as very high and were removed, leaving bigram frequency (type) as the measure for this construct.

5.2.5.6 *Sensitivity Analyses*

Complete Cases. We repeated the analysis for the preferred and design implied models for each task using a reduced data set comprising those participants who returned for all three data collection sessions. Participants who left the study or where a specific set of experimental task data was missing from one or more data collection points were removed. The number of complete cases for the letter search and lexical decision task is 161; for word naming it is 165; for sentence reading it is 169.

Outlier Analyses. We checked estimates for preferred models for the influence of outliers. We calculated the inter-quartile range of raw reaction time for each participant, multiplied the first and the third quartile range by 1.5 and subtracted / added that value to the first and third quartile for each participant (Baguley, 2012). Reaction times that lay outside these lower and upper boundaries were classed as outliers and removed from the data set. We performed this data reduction on both the full sample and the complete cases, re-running the preferred models for each task outcome measure on these trimmed data.

In the next two chapters, we present our findings. In Chapter 6, we present the descriptive statistics of the sample. At the outset of the chapter we detail the rate of study attrition and missing data. We present the results of the cluster analysis. We explore differences between groups across ID measures at T1. These analyses are motivated by the studies reviewed in Chapter 2 for the adult-learner population. We also visualise participant variability within group data across time to appreciate what type of function may be appropriate for modelling the task data.

To anticipate the individual difference findings, atypically-reading adults show greater similarity with atypically-reading 16-17-year-old readers and typical 11-12-year-old readers, than with typically-reading adult and 16-17-year-olds and atypically-reading 11-12-year-old readers, except for the phonological and rapid naming tasks.

In Chapter 7, we present the results of the preferred models from the four experimental tasks.

6 Results: Descriptive Statistics

6.1 Attrition and Missing Data

6.1.1 Attrition

Figure 6.1 displays rate of attrition at the group level for each time point. At T1 there were 218 participants. This decreased to 191 participants at T2 and further decreased to 173 participants at T3. Typically-reading 16-17-year-old and adults are relatively stable over time. The highest rate of attrition is for the atypically-reading 16-17-year-olds and adults between T1 and T2.

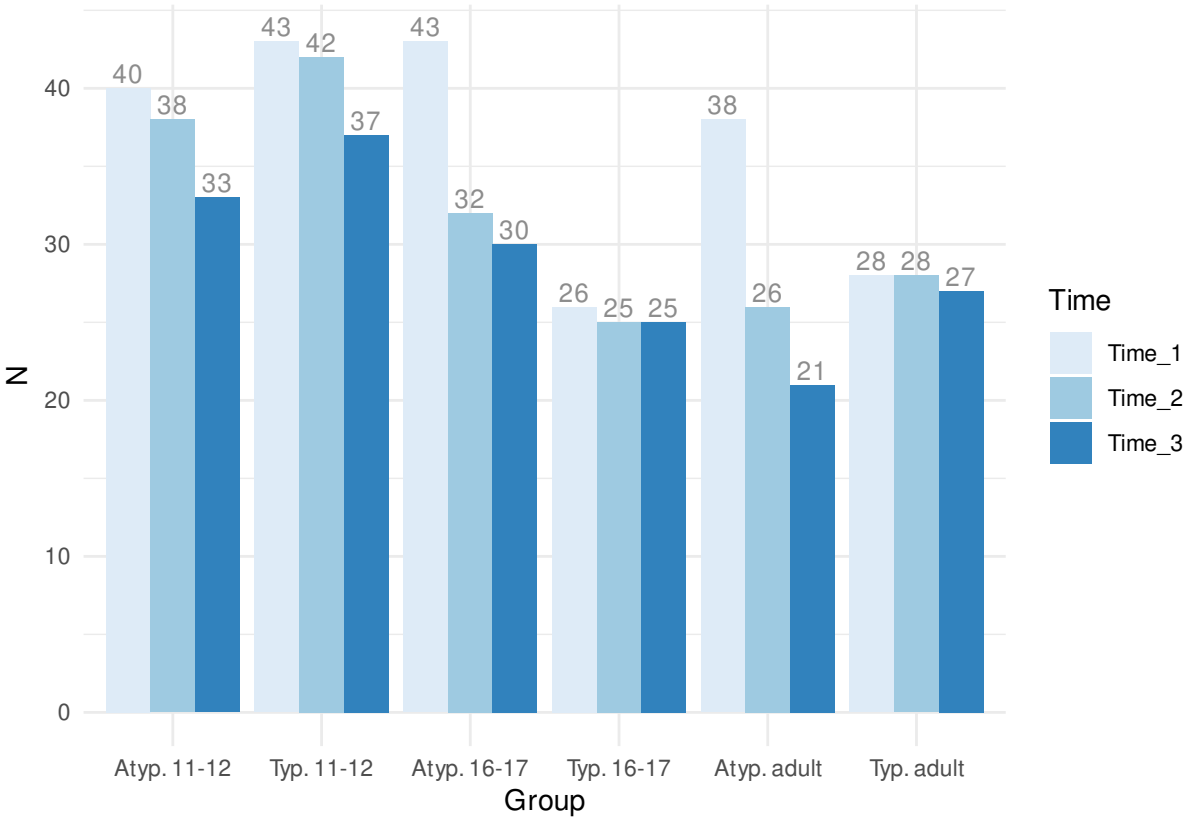
In the 11-12-year-old groups, participants who withdrew were evenly balanced across typical- and atypical-readers. There were no differences in age between continuing and withdrawing participants. In both groups, withdrawing participants performed similarly to their peers at T1 (all ps $> .05$), with the exception for rapid naming skill (RON) in the atypical-readers group. The participants who withdrew showed lower RON skill scores than the participants who continued (mean of 1.2 vs 1.4, $t(10.79) = 2.73$, $p = .02$).

In the 16-17-year-olds, 13 of the 14 participants who withdrew from the study were from the atypically-reading group. There were no mean differences in age between withdrawing and continuing participants at T3. The average performance over the ID measures of those that withdrew from the atypically-reading group at T2 or T3 was not significantly different from the atypically-reading participants that remained (all ps $> .16$).

In the adult reader group, 17 / 18 participants who withdrew from the study were in the atypically-reading group. There were no mean differences in ages between atypically-reading withdrawing adults and continuing atypically-reading adults at T3

Figure 6.1

Barplot of Participant Attrition by Group From T1 - T3



($p = 0.277$). Across all ID measures at T1, the performance of withdrawing participants were not statistically significant from those that remained (all ps $> .07$).

6.1.1.1 *Summary*

But for the significant finding of lower RON skill scores in the 11-12-year-old withdrawn participants for the atypically-reading group, all other groups and ID measures are similar between those participants who completed the study and those that withdrew after T1. We can be confident that any differences we see between groups or influence for ID measures on experimental tasks are not due to the change in group membership scores across time points. The absence of significant differences described above suggests that the risk of bias in model inferences due to attrition is minimal.

6.1.2 Missing Data

Some participants completed all experimental tasks but, due to unforeseen circumstances at the time, are missing data on some ID measures. Where experimental task data are missing, the participant is omitted from that task analysis. Full details for the missing data process are listed in Appendix F.

At T1, missing data is present for spelling ($n = 22$ / 218; 10%) and vocabulary scores ($n = 27$ / 218; 12%). This is due to an error on the administration of the tests to classes which included a subset of participants. We used random regression imputation (Gelman et al., 2021) to impute values for the missing data.

At T2, one participant (1 / 191; 0.5% per ID measure) completed experimental tasks but not ID measures due to a fire drill during the session. A further participant did not complete the spelling test on the day of testing due to absence from school.

At T3, one participant (1 / 173; 0.5%) did not complete ID measures at T3 due to timing difficulties on the day of testing. Five participants (5 / 173; 2.8%) did not complete the spelling measure and four participants (4 / 173; 2.3%) did not

complete the vocabulary measure. This is due to absence from class on the day that the measures were administered to the 11-12-year-old participants.

At T2 and T3 the missing data rates per ID measure are all below 5%. Due to the low rate of missingness and the mechanism of missingness categorised as missing at random, we use single value random sampling to impute values for these participants (Gelman et al., 2021).

After imputation, we tested for differences between the data with missing values and the data with imputed values, there were no statistically significant differences between the data sets (all ps $> .7$). We present visualisations and results from the imputed data set from this point forward.

6.2 Differences Between Groups

We next describe differences between group performance. Mean and standard deviation values for the ID measures by time and group are in Table 6.1[1]. The distribution of scores in each group for each measure at T1 is presented in Figure 6.2. Since our primary focus is the atypically-reading adults, we label significant differences between that group and the other groups on the box plots and describe other significant differences in the text.

We know that the number of days between data collection sessions varies between individuals, but even with this variability, scores within group did not significantly change over time. Consequently, for descriptive purposes for ID measures, collapsing the number of days to single time points has a low risk of introducing bias into the means or any inferences. We inspect between group differences in means in the full sample at T1 ($n = 218$), and check to see if those relationships hold over T2 and T3. We repeat these checks for the complete case data set ($n = 173$).

[1]We report means and sd values here but the majority of tests conducted moved to non-parametric methods when assumptions were not met for the ANOVA.

Table 6.1

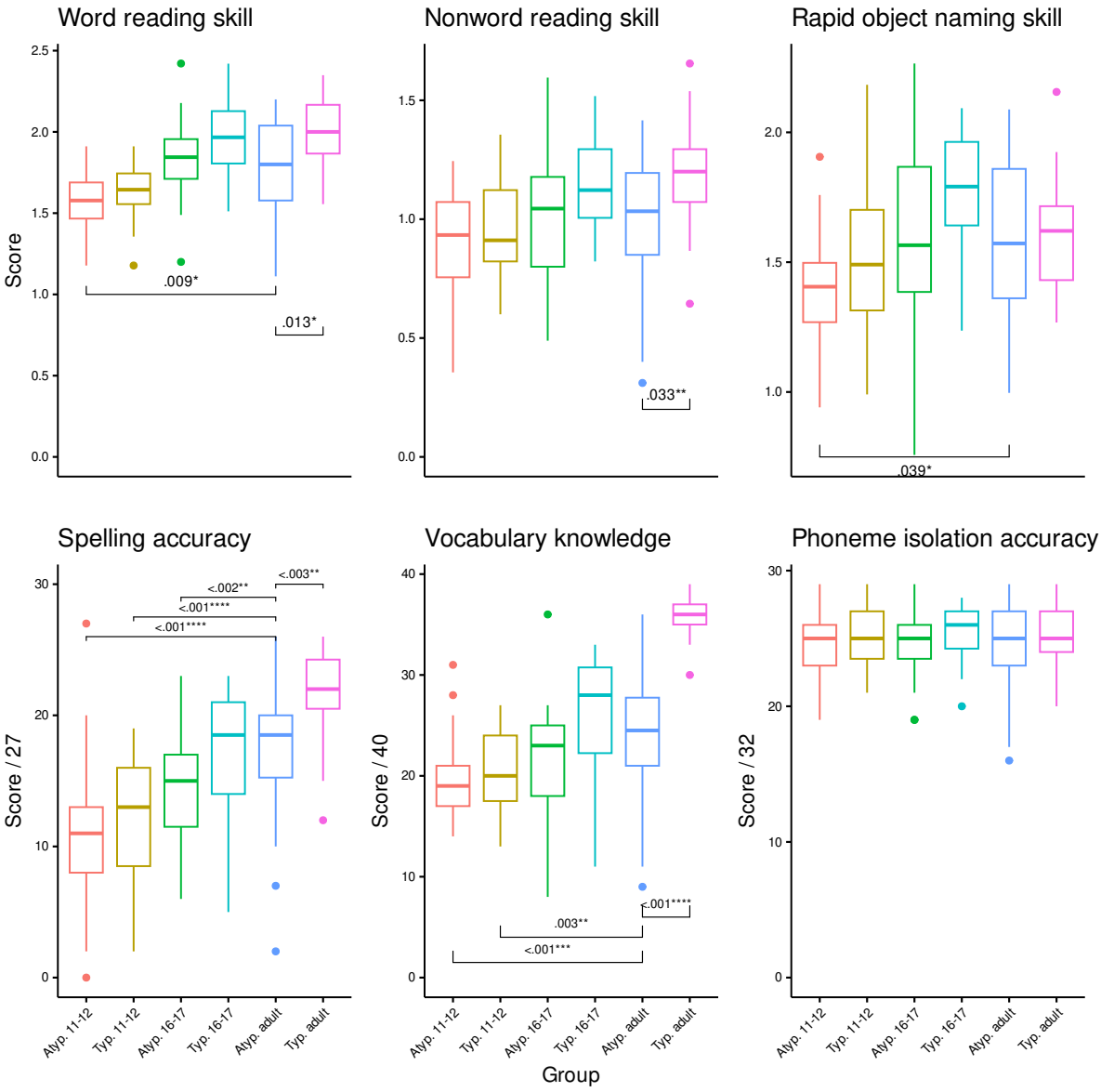
Number of Participants, Means (SD) for Age and ID Measures for Each Group and Data Collection Point.

Time	n	Age (Yrs)	Word accuracy	Nonword accuracy	Word skill	Nonword skill	PI	RON	Spell	Voc
Atypical 11-12										
1	40	11.9 (0.5)	70.7 (8.3)	40.5 (9.6)	1.6 (0.2)	0.9 (0.2)	24.5 (2.5)	1.4 (0.2)	10.9 (5.1)	19.6 (3.7)
2	38	12.1 (0.5)	73.1 (8.6)	43.5 (9.6)	1.6 (0.2)	1 (0.2)	24.9 (2.1)	1.5 (0.2)	11.8 (4.3)	21.2 (3.3)
3	33	12.3 (0.5)	76.8 (8.7)	44.8 (10)	1.7 (0.2)	1 (0.2)	25.1 (2.4)	1.5 (0.2)	13.1 (4.4)	22.4 (4.3)
Typical 11-12										
1	43	11.8 (0.3)	73.7 (6.7)	43.1 (8.3)	1.6 (0.1)	1 (0.2)	25.1 (2.4)	1.5 (0.3)	12.1 (4.7)	20.4 (3.8)
2	42	12 (0.3)	76 (7.2)	44.9 (9)	1.7 (0.2)	1 (0.2)	25.6 (2.2)	1.5 (0.3)	13.6 (4)	22 (3.1)
3	37	12.3 (0.3)	79.5 (8.3)	46.8 (8.3)	1.8 (0.2)	1 (0.2)	25.1 (2)	1.6 (0.3)	14 (4.7)	22.5 (3)
Atypical 16-17										
1	43	17.4 (1)	81.6 (9.3)	45 (10.3)	1.8 (0.2)	1 (0.2)	24.8 (2.4)	1.6 (0.4)	14.3 (3.6)	21.3 (5.3)
2	32	17.4 (0.8)	83.7 (9.3)	47.3 (9)	1.9 (0.2)	1.1 (0.3)	24.6 (2.3)	1.7 (0.3)	15.3 (3.7)	24 (5.1)
3	30	17.7 (0.7)	85.3 (9.9)	48.6 (8.5)	1.9 (0.2)	1.1 (0.3)	25.3 (2.2)	1.7 (0.3)	15.5 (3.8)	23.8 (5.2)
Typical 16-17										
1	26	16.7 (0.4)	87.3 (9.9)	50.5 (7.7)	1.9 (0.2)	1.1 (0.2)	25.5 (1.9)	1.8 (0.2)	17.2 (4.6)	26.2 (5.6)
2	25	17 (0.3)	88.6 (9.1)	53.3 (6.4)	2 (0.2)	1.2 (0.2)	26.1 (1.5)	1.8 (0.2)	17.8 (3.8)	28.1 (4.9)
3	25	17.3 (0.3)	90.2 (8.5)	53.7 (6.6)	2 (0.2)	1.3 (0.3)	25.8 (1.4)	1.8 (0.3)	18 (4.1)	27.8 (3.8)
Atypical adult										
1	38	33.1 (9.4)	79.6 (13.1)	44.5 (11.8)	1.8 (0.3)	1 (0.3)	24.6 (2.9)	1.6 (0.3)	17.7 (5.1)	24.2 (6)
2	26	34 (10.1)	84.3 (11.1)	48.1 (12.8)	1.9 (0.2)	1.1 (0.3)	24.3 (2.5)	1.7 (0.3)	17.3 (4.9)	26.1 (6.2)
3	21	34.9 (10.4)	81.5 (14.9)	47.5 (13.4)	1.8 (0.3)	1.1 (0.3)	25 (2.4)	1.7 (0.3)	16.6 (5.4)	26.1 (7.2)
Typical adult										
1	28	56.4 (12.5)	89.8 (8.6)	52.1 (7.9)	2 (0.2)	1.2 (0.2)	25.3 (2.1)	1.6 (0.2)	21.6 (3.6)	36 (1.9)
2	28	56.8 (12.5)	90.9 (9.4)	52.8 (6.8)	2.1 (0.3)	1.2 (0.2)	26.1 (2.7)	1.6 (0.2)	22.1 (4)	35.8 (2)
3	27	58.1 (11.4)	91.7 (10)	53 (7.7)	2.1 (0.3)	1.2 (0.3)	26.1 (2.4)	1.6 (0.3)	22.2 (3.7)	36.1 (2.5)

Note:
n = Sample size. PI = Phonological awareness skill. RON = Rapid naming skill. Spell = Spelling knowledge. Voc = Vocabulary knowledge.

Figure 6.2

Boxplots of Distribution of ID Measure Scores by Group at T1 (n = 218)



6.2.1 Word Reading Skill

A Kruskal-Wallis test for group on word reading skill scores was significant ($H(5) = 74.39$, $p < .001$). The pairwise Wilcox test showed that atypically-reading adults read more words correctly and faster (mean = 1.8, sd = 0.3) than the atypically-reading 11-12-year-olds (mean = 1.6 sd = 0.1; *adj.p* = .009) at T1 and T2 but not T3. This

difference is not present in the complete cases data set. They are closer in their word reading skill mean which results in a non-significant difference ($adj.p = .424$).

Atypically-reading adults read less words correctly than the typically-reading adults (mean = 2 sd = 0.2; $adj.p = .013$) at T1 and T3. This difference remains in the complete case data set.

The raw accuracy measures show the same pattern. Atypically-reading adults showed no statistically significant differences in word reading accuracy scores from the typically-reading 11-12-year-olds ($adj.p = .079$) or the 16-17-year-olds ($adj.p_{atypical} = .954$; $adj.p_{typical} = .107$). There were no significant differences between the two 11-12-year-old groups nor the 16-17-year-old groups. This pattern is observed across the complete cases data set.

In summary, the word reading skill scores for atypically-reading adults are more like typically-reading 11-12-year-old and both groups of 16-17-year-old readers than their adult peers. The atypically-reading adults are stronger than the atypically-reading 11-12-year-old group however this difference has closed by the end of the study. The pattern is identical for accuracy scores. From this point forward, we report only word reading skill measures.

6.2.2 Nonword Reading Skill

The ANOVA test for nonword reading skill scores showed significant effects of group on nonword reading skill ($F(5, 212) = 7.64$, $p < .001$). The Levene's and Shapiro-Wilk tests returned p values $> .05$ however this was the only ANOVA that did not violate any of the model assumptions. For ease of comparison, we conducted the Wilcox test and report those results here.

The Wilcox test showed that the atypically-reading adults (mean = 1.0, sd = 0.3) differed only from the typically-reading adults on nonword reading skill (mean = 1.2, sd = 0.2, $p = .033$). The difference was only apparent at T1, not at T2 or T3. Atypically-reading adults perform equivalently to the younger participants. There were no differences between the two groups in 11-12- and 16-17-year-olds.

The difference between the adult readers is present in the complete cases data set. The atypically-reading adults show a lower nonword skill score in the complete cases data set compared to the full data set (mean = 0.96, sd = 0.3). They remain significantly different from the typically-reading adults (mean = 1.2, sd = 0.2, $adj.p$ = .016).

In summary, the atypically-reading adults show no statistically significant differences with the younger groups on the nonword reading skill measure. The atypically-reading adults do show differences with the typically-reading adults, however the conditions for a significant finding are unstable. The pattern is identical for accuracy scores. From this point forward, we report only nonword reading skill measures.

6.2.3 Phonological Awareness Skills

A Kruskal-Wallis test showed no significant differences between groups in phonological awareness skills ($H(5) = 4.20$, $p = .520$) at T1. This pattern remains for T2 and T3. There are no differences between groups in the complete cases data set. The range within which each group mean falls is incredibly narrow: 24.5 - 26.1 out of 32 items. The majority of participants in each group successfully completed two-thirds of the items.

6.2.4 Processing Speed

A Kruskal-Wallis test showed significant effects of group on rapid object naming skill (RON; $H(5) = 29.14$, $p < .001$). A pairwise Wilcox test showed that the atypically-reading adults (mean = 1.6, sd = 0.3) differed significantly from the atypically-reading 11-12-year-olds (mean = 1.4, sd = 0.2, $adj.p = .039$). There were no statistically significant differences between the atypically-reading adults and any of the other groups.

The atypically-reading 11-12-year-olds were significantly different from all of

the older reading groups (all *adj.p*s $< .039$). The typically-reading 11-12-year-old group (mean = 1.5, sd = 0.3) was significantly different from the typically-reading 16-17-year-olds (mean = 1.8, sd = 0.2, $adj.p = .003$). There were no differences between groups within the same age bands for 11-12- and 16-17-year-olds. This patterns remains for T2 and T3.

In the complete cases data set, the observed difference between the atypically-reading adults and the atypically-reading 11-12-year-olds was no longer present. Atypically-reading adults also showed no statistically significant difference with any of the other groups.

In summary, RON scores for atypically-reading adults differ only between the scores for the atypically-reading 11-12-year-olds. We identified that for those 11-12-year-old readers who withdrew from the study after T1, there was a lower mean score, which probably explains why this difference is not found in the complete cases data set.

6.2.5 Spelling Knowledge

A Kruskal-Wallis test showed significant effects of group on spelling scores ($H(5) = 85.35$, $p < .001$). A pairwise Wilcox test showed that atypically-reading adults spelled more words correctly (mean = 17.7, sd = 5.1) than both 11-12-year-old groups (atypical: mean = 10.9, sd = 5.1, $adj.p < .001$; typical: mean = 12.1, sd = 4.7, $adj.p < .001$).

The atypically-reading adults also showed a significant difference from the spelling scores of the atypically-reading 16-17-year-old group (mean = 14.3, sd = 3.6, $adj.p = .002$). On average, atypically-reading adults spelled more words correctly than the atypically-reading 16-17-year-olds. These differences were present for T1 but no longer apparent at T2 or T3. There was no statistically significant difference between the atypically-reading adults and typically-reading 16-17-year-old readers.

The atypically-reading adults did not spell as many words correctly as the typically-reading adults (mean = 21.6, sd = 3.6, $adj.p = .003$) at T1. This pattern

holds across T2 and T3.

There were no differences between groups within 11-12-year-olds. There was a significant difference between 16-17-year-old groups (atypical mean = 14.3, sd = 3.6; typical mean = 17.2, sd = 4.6; $adj.p = .003$).

The pattern of differences is the same between the atypically-reading adults and both 11-12-year-old groups in the complete cases data set at T1 to T3. Atypically-reading adults showed slightly higher spelling means (mean = 17.8, sd = 5.4) and remained significantly different from the atypically-reading 11-12-year-old group (mean = 11.2, sd = 5; $adj.p < .001$) which is also a slightly higher score. The typically-reading 11-12-year-olds also increased their score (mean = 12.4, sd = 4.7, $adj.p < .002$) and this difference remained statistically significant.

The atypically-reading adults and the atypically-reading 16-17-year-old group (mean = 14.8, sd = 3.7) remained significantly different from each other ($adj.p = .047$). The atypically-reading adults continued to score significantly lower than the typically-reading adults (mean = 21.6, sd = 3.6; $adj.p = .047$).

We explored the spelling errors of the sample to understand whether atypically-reading adults were varied in the type or quantity of errors for the spelling task. The full error analysis is listed in Appendix G with a summary of findings reported here for brevity.

Atypically-reading adults had higher odds of omitting an answer than typically-reading 16-17-year-olds. They were just as likely to write a real word as a substitute for a target word that was not a homophone. The odds of supplying a real word when the target word *was* a homophone increased, as it did with typically-reading 16-17-year-olds and the difference in error rates here was not significant.

Fewer of the errors are likely to be a plausible sound-match to the target word in atypically-reading adults and atypically-reading 16-17-year-old groups. In a sound-match type of spelling error, these two groups resemble the younger reading groups. The atypically-reading adults were inconsistent in the spelling errors for the same target word on separate occasions, both orthographically and phonologically.

In summary, in terms of absolute scores for correct answers, atypically-reading adults show stronger spelling skills than the youngest readers, and the atypically-reading 16-17-year-olds. We must interpret these findings with caution, however. It is expected that the youngest readers in the sample may not have experienced some of the word items in the spelling test so it is appropriate that they are lower in skill at this time.

When we look at spelling errors, atypically-reading adults look very similar in their approach and types of errors to the younger readers. In relation to their sources of knowledge, atypically-reading adults appear inconsistent in applying their phonological knowledge to spelling. They are both less frequent and more varied in their attempts at providing a plausible sound match to the target word.

6.2.6 Vocabulary Knowledge

A Kruskal-Wallis test showed significant differences between groups for vocabulary knowledge ($H(5) = 100.22$, $p < .001$). A pairwise Wilcox test showed that atypically-reading adults identified more synonyms correctly (mean = 24.2, sd = 6.0) than both 11-12-year-old groups (atypical: mean = 19.6, sd = 3.7, $adj.p < .001$; typical: mean = 20.4, sd = 3.8, $adj.p = .003$) at T1 and T2 but not at T3. There were no statistically significant differences for vocabulary scores between atypically-reading adults and either of the 16-17-year-old groups. Atypically-reading adults did not know as many synonyms as the typically-reading adults (mean = 36.0, sd = 1.9, $adj.p < .001$).

There were no differences between the two 11-12-year-old groups. There was a significant difference between the two 16-17-year-old groups. The atypically-reading 16-17-year-olds (mean = 21.3, sd = 5.3) knew fewer synonyms than the typically-reading 16-17-year-old group (mean = 26.2, sd = 5.6, $adj.p = .005$).

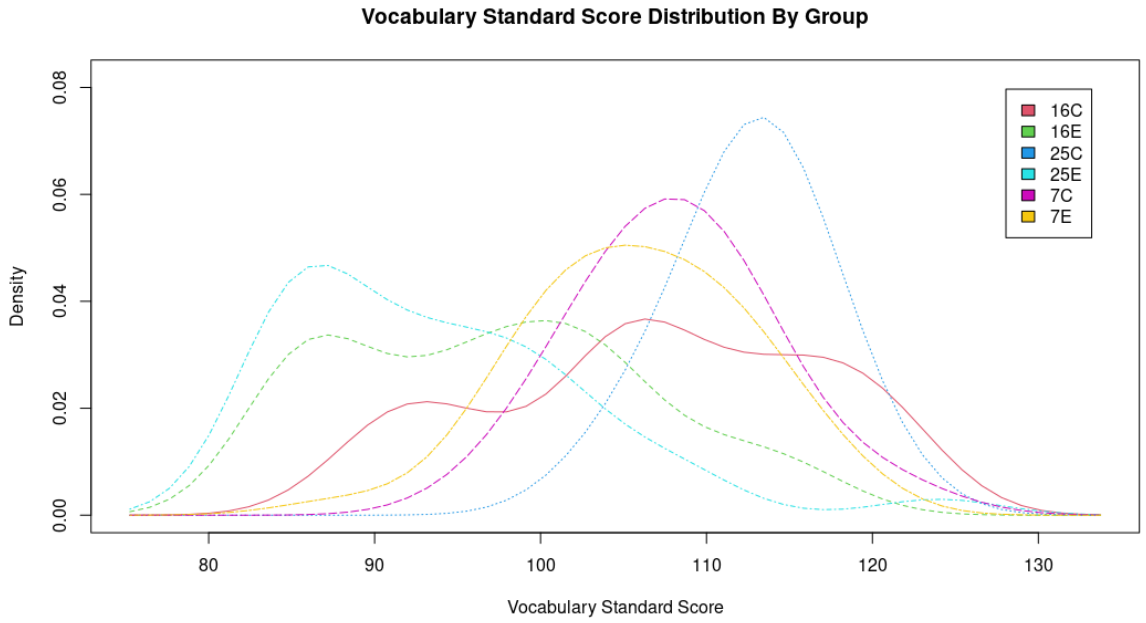
In summary, the atypically-reading adults appear to have similar vocabulary knowledge as 16-17-year-old readers. They have significantly weaker vocabulary knowledge compared to their typically-reading adult peers. The finding that they

differ from the 11-12-year-old readers is unsurprising, given that the 11-12-year-old readers are much younger and so much less experienced in everyday, language exposure. The loss of this significant difference at T3 is due to stability of scores in the atypically-reading adults and continuing increase in the scores of the younger groups. The pattern of change over time was mirrored in the complete cases data set.

The above describes vocabulary knowledge as a function of raw scores from this sample. The Shipley Vocabulary Scale (Shipley, 1940) is designed for use between the ages of 7-89 years and has standard scores available across the age range. For interest, we calculated the standard scores for vocabulary for the sample (Figure 6.3). The youngest reader groups display age-appropriate levels of vocabulary with the mass of their distribution being centred around 100-110. The bulk of the atypically-reading 16-17-year-olds and adults' distributions are located towards the lower half of the plot. Both groups show low standard scores for vocabulary.

Figure 6.3

Density Plot of Distribution of Vocabulary Standard Scores by Group at T1 (n = 218)



6.2.7 Cluster Analysis

For each data collection point, the Dip Test was non-significant (T1: *Dn*: 0.02, p = .587; T2: *Dn*: 0.02, p = .755; T3: *Dn*: 0.02, p = .783). Consequently, we fail to reject the null hypothesis and assume that the sample distribution is unimodal at each data collection point.

The randIndex value for a six cluster solution was T1 = 0.11, T2 = 0.16, T3 = 0.12. The randIndex value for a three cluster solution was T1 = 0.11, T2 = 0.13, T3 = 0.14. These low numbers indicate a lack of agreement between the group labeling and arrangement of observations as identified by the PAM method.

The findings suggest that that data do not support six or three distinct groups as the labels or ages of the groups may suggest, Rather, to echo the analysis within ID measures across groups, the 11-12-, 16-17-year-olds and adult readers form one unimodal distribution.

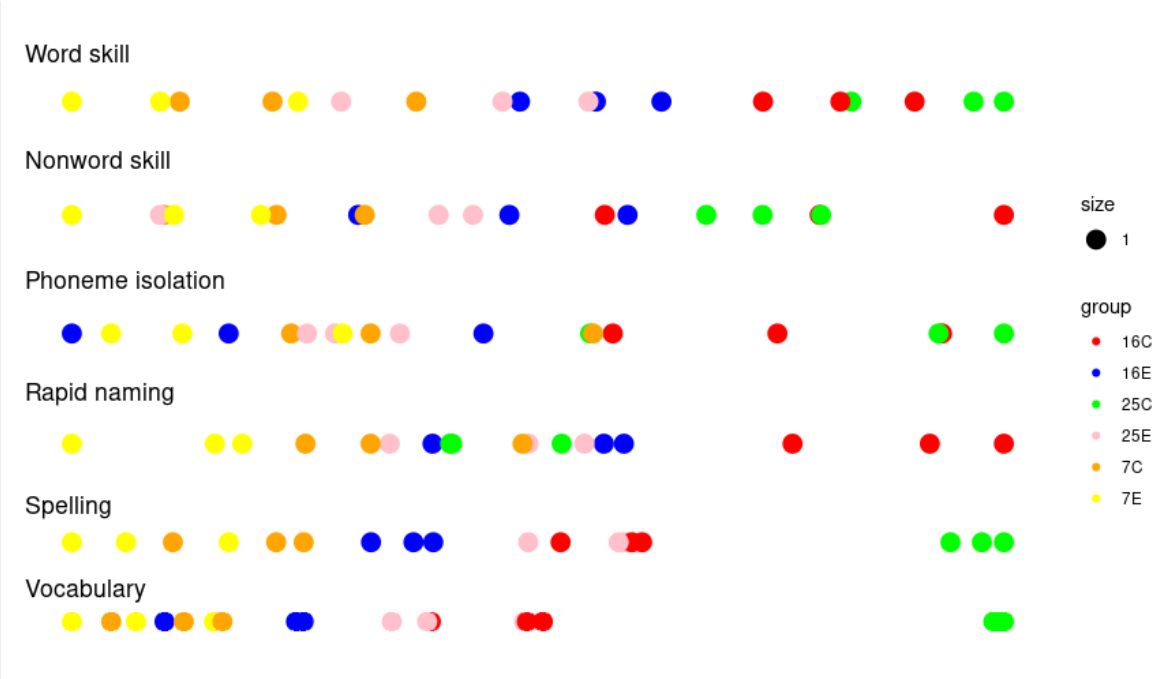
We plotted group average scores for each time point and arranged them in ascending order along a number line (Figure 6.4). The atypically-reading 11-12-year-olds are generally at the lowest point across all skills and the typically-reading 16-17-year-olds or adults towards the highest. Groups overlap or are contiguous in their scores, but notice the mingling of the atypically-reading adults (pink dots) with the younger reading group means.

6.3 Individual Variation for Skills Over Time

The cluster analysis suggests that the data collected at each session comes from a unimodal distribution and that the study's nominal grouping structure is not in evidence in the data. Inspecting differences between groups supports this with differences mainly occurring for the atypically-reading adults and the youngest, atypical readers and the typically-reading adults. The overarching trend suggested by mean differences is that, in ID measures at least, atypically-reading adults' skills are more similar to typical 11-12- and 16-17-year-old readers.

Figure 6.4

Line Plots of Group Means per ID Measure, T1 - T3, in Ascending Order



The findings so far have been based upon group means with a focus of between group differences. We can also visualise how individual participant performance varies from their group mean. Figures 6.5 and 6.6 display spaghetti plots for each measure by group. The plots are arranged from younger to older participants from left to right in each row. Individual participant trajectories (grey lines) are visible over which is drawn a smoothed line of best fit to visualise the group average for easier comparison to participants within groups and also between groups.

The plots serve two purposes. The first is to visualise the spread of participant scores within groups. Not only can we compare by relative lows and highs of scores but also by time. The x-axis now represents the variable number of days between data collection sessions rather than labels of T1, T2 and T3. The second purpose is to visualise trends over time to assist with a function for modelling the longitudinal data (Long, 2012).

First of all, the length of grey lines are longer in the older reader groups,

Figure 6.5

Spaghetti Plots Showing Individual Variation By Group in Performance for Word Reading Skill (Top), Nonword Reading Skill (Middle) and Phonological Awareness Skill (Bottom) Across Time. Grey Lines Represent Individual Participant Curves. Blue Lines Represent the Group Average (LOESS Estimate).

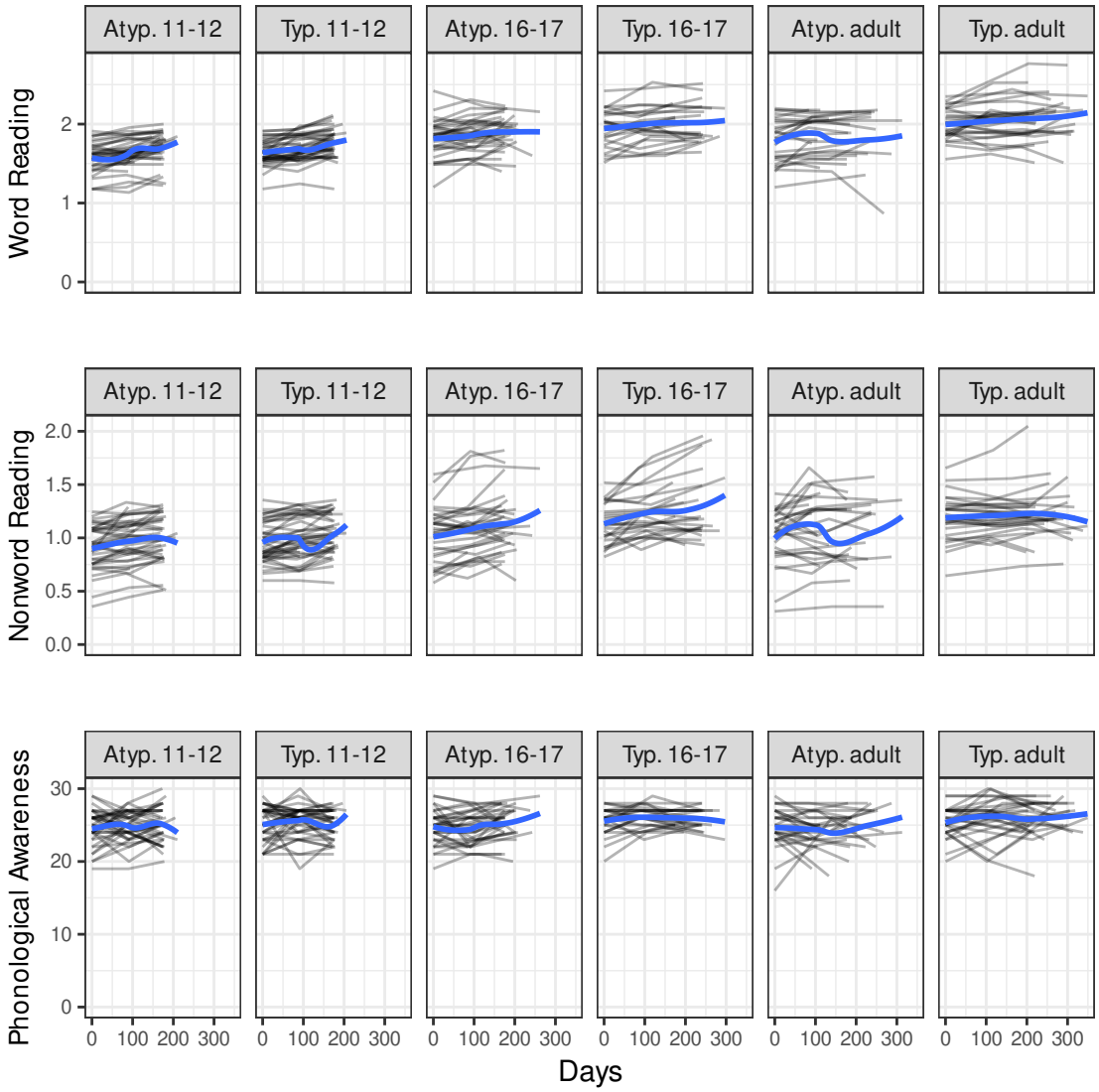
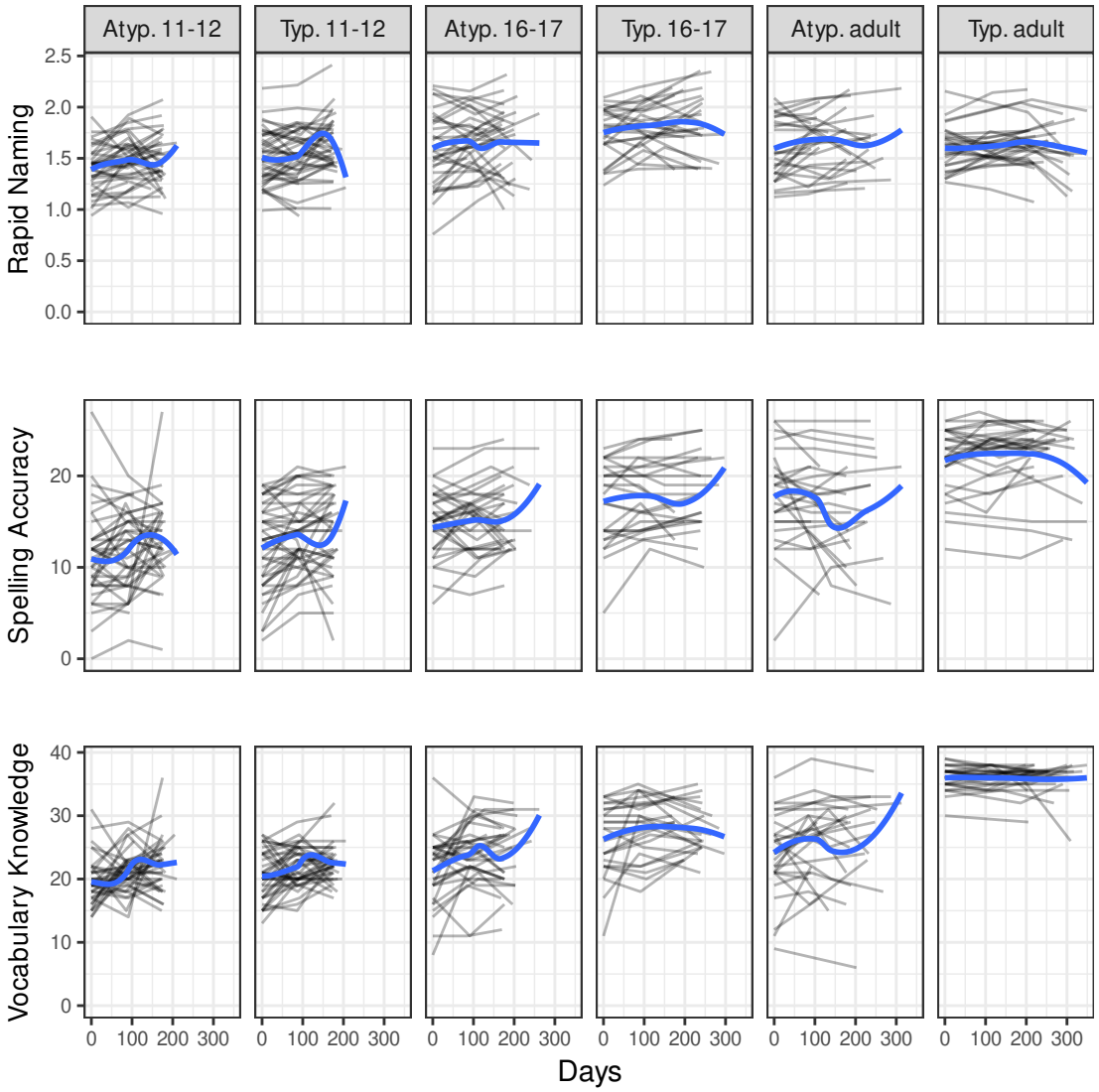


Figure 6.6

Spaghetti Plots Showing Individual Variation By Group in Performance for Rapid Naming Skill (Top) Spelling Accuracy (Middle) and Vocabulary Knowledge (Bottom) Across Time. Grey Lines Represent Individual Participant Curves. Blue Lines Represent the Group Average (LOESS Estimate).



illustrating the fact that they tended to have a higher number of days between data collection points. Second, it is clear that as participant age increases from group to group, performance on each measure rises, with two exceptions. The first exception to this is for phonological awareness skill (bottom row Figure 6.5). Performance on this measure for all at day 0 is very similar, very high and performance for each group remains flat over time (supported by inferential test results in section 6.2.3). The second exception is RON skill (c.f. section 6.2.4). Although subtle, performance appears to improve over time for 11-12- and 16-17-year-olds but the rate of change between the typically-reading 16-17-year-olds and the typically-reading adults is less.

Three other patterns emerge from the plots. First, variability within the atypically-reading adults tends to be greater or equivalent to younger group performance, across all measures. Second, the spelling measure appears to show the greatest variability both within groups and across time. Third, there is a marked difference in the spread of vocabulary scores between the 11-12-, 16-17-years-olds and atypically-reading adults, and the typically-reading adults. Both the youngest readers and typically-reading adults show much narrower spread of vocabulary scores than the 16-17- and atypically-reading adults.

There is clear evidence of variation within groups for patterning of scores between individuals, evidenced by the individual grey lines. Yet the blue lines that represent the general trend in group data tend to show a positive linear function, suggesting that the modelling strategy of models with independent and interaction terms is appropriate.

6.4 Bivariate Correlations Between ID Measures

Table 6.2 lists Pearson's r correlations between the ID measures, collapsed over time and group. Since Pearson's r is well known for returning significant p values < 0.5 when the number of observations upon which it is based is large, it is more useful to discuss the size and direction of the correlations.

The highest correlation value is that between the skill measures of word and

nonword reading. They show a very large, positive correlation ($r = .74$). This is slightly lower than the reported correlation of .83 from the test manual (Torgesen et al., 2012). Intuitively this makes sense since both measures involve the decoding of letter strings and developmentally, all words are nonwords when experienced for the first time.

The next highest correlation is between spelling knowledge and nonword reading ($r = .66$). We interpret this as the two skills tapping an underlying sublexical knowledge of letters and how they are arranged to form words for decoding (nonword reading) and encoding (spelling) printed letters. We note that spelling and word reading skill also share a large sized correlation ($r = .62$), and that the difference in size between the two correlations is unlikely to be statistically significant.

Age and vocabulary knowledge are also positively correlated with each other ($r = .65$). This large sized relationship is not surprising given the superior knowledge of the typically-reading adults, also clear from Figure 6.2 and the spaghetti plot for vocabulary in Figure 6.9. The large size of the relationship between age and vocabulary compared to the small sized correlation values for age and word / nonword reading suggest that vocabulary knowledge continues to accrue over a range of age while word and nonword reading skill does not to the same extent (Keuleers and Balota, 2015). This makes sense if we consider that mastery of letter-sound relationships may constitute a finite set of knowledge and skills of word and nonword reading while the combinatorial possibilities of those letters to make words for vocabulary learning are numerous. Given the strength of the relationship between age and vocabulary, age needs to be an independent predictor in the experimental task models.

Phonological awareness skill shows small and positive relationships with all the ID measures. Even in a sample that has relatively good reading skills (i.e., none are beginners), the early skill of being able to isolate a phoneme within a string of phonemes shares a modicum of variance with the higher order skills.

RON shows a medium positive correlation ($r = .44$) with nonword reading skill and a large positive correlation with word reading skill ($r = .55$). Each of these

Table 6.2

Summary of the Bivariate Correlations Between ID Measures

	W-Skill	NW-Skill	PI	RON	Spell	Voc
W-Skill						
NW-Skill	0.74***					
PI	0.19***	0.25***				
RON	0.55***	0.44***	0.10*			
Spell	0.62***	0.66***	0.29***	0.28***		
Voc	0.52***	0.41***	0.21***	0.20***	0.62***	
Age	0.32***	0.18***	0.05	0.04	0.48***	0.65***

Note:
W-Skill = Word reading skill. NW-Skill = Nonword reading skill. PI = Phonological awareness skill. RON = Rapid naming skill. Spell = Spelling accuracy. Voc = Vocabulary knowledge.
[1] p <.05; ** p <.01; *** p <.001.

relationships are for speeded naming measures. Correlations are lower between RON and the untimed measures of spelling ($r = .28$), vocabulary ($r = .21$).

6.5 Discussion

We discuss the findings with a particular focus on the differences of the atypically-reading adults with other groups.

The cluster analysis showed no significant support for distinct groups, suggesting that the range of skills present within the data are similar to each other. This is further supported by the visualisations in Figures 6.3 to 6.6. The crude arrangement of group means shows a continuum of skill while the spaghetti plots clearly show increases in intercepts as age increases. The variation within each group overlaps with other groups. In terms of similarity, this sample of atypically-reading adults lies between the typically-reading 11-12- and 16-17-year-olds.

The lack of significant change within groups in scores across time can be interpreted in different ways. First, each of the scores contain a true score plus

measurement error. At each data capture, the observed scores may be higher or lower, but they hover around a true score that is stable and robust. When we aggregate the small deviations, they cancel each other out so that over time, a finding of no change is observed. Second, the measurement error is minimal but increments of change are too small to detect in the available time frame. This is an area for consideration for future research design.

There were no group differences in scores for the phonological isolation measure. We do not discuss this measure any further. Atypically-reading adults showed stronger RON scores than the atypically-reading 11-12-year-olds (only at T1) but no differences with any of the other groups. This suggests that any differences that may be observed in the experimental tasks are not underpinned by slower processing in the atypically-reading adults (Kirby et al., 2010).

In nonword reading skill, the only difference was with the typically-reading adults, where the atypically-reading adults showed weaker skills. In word reading, atypically-reading adults show similar word reading scores as the typically-reading 11-12- and 16-17-year-olds but were significantly stronger than atypically-reading 11-12-year-olds and significantly weaker than typically-reading adults.

In vocabulary knowledge, there were no significant differences with the 16-17-year-old readers, but atypically-reading adults were stronger than 11-12-year-old readers and weaker than typically-reading adults in raw scores. When we looked at standard scores, we found that the greater proportion of scores were lower than expected for age for atypically-reading 16-17-year-olds and adults.

Any differences that were present at T1 between atypically-reading adults and the younger groups of readers, did not hold over time. The atypically-reading adults showed difference with every group in spelling scores except the typically-reading 16-17-year-olds. They were stronger in skill than 11-12-year-olds and the atypically-reading 16-17-year-olds and weaker than typically-reading adults. This suggests that spelling is a relative strength in adult readers.

How do these findings comport with previous studies on atypically-reading adults? The lack of difference in phonological skills may align with Greenberg et al.

(1997), in that they found phonological awareness skills equivalent to that of 11-year-old readers in their adult-learner sample. Yet, the lack of difference here also includes a group of typically-reading adults. These findings do not converge with Braze et al. (2007) who found that phonological awareness could be a skill that could be used for discerning groups – all our groups were similar to each other.

Not only were scores similar but it was the final section of the test on which all participants faltered. We suggest that the level of phonological awareness demonstrated in the sample is at a level that is necessary and sufficient to support the development of higher order skills and a lack of difference shows asymptote levels of performance for this task for a population that is characterised as "average" readers.

For nonword reading skill, there begins to be some trace of between group differences. Atypically-reading adults are not significantly different from the younger reading groups but are significantly weaker from their adult peers. This is markedly different from the level of skills as described in Greenberg et al. (1997), Mellard et al. (2010), and Nanda et al. (2010). They all found that their adult-learners showed nonword reading skill below 4th grade readers ($<$ 10 years) however the atypically-reading adults and the typically-reading 11-12-year-olds show nonword reading skills equivalent to a range of readers – US 5th - 10th grade equivalency. The atypically-reading adults in this sample appear to show nonword reading skills that are similar to that of average readers of secondary school leaving age.

Much of the literature describing adult-learners proposed that their orthographic knowledge and skills, while still weak in absolute terms, were a relative strength compared to their phonological skills (Greenberg et al., 1997; Mellard et al., 2010; Tighe and Schatschneider, 2016). In this sample, word and nonword reading skill appear to be equivalent in strength. The difference observed from the atypically-reading 11-12-year-olds at T1 is gone by T3, as the younger readers increase their scores. There are no differences for word reading scores between the 16-17-year-old readers or the typically-reading 11-12-year-olds, a similar pattern to that of the nonword reading skills.

Greenberg et al. (1997) observed correlations between word and nonword

reading in the range of $r = .66$, however here we see a stronger relationship ($r = .74$). Further differences are observed between spelling and word / nonword reading measures. Greenberg et al. (1997) observed larger correlations between word and spelling than nonword and spelling measures across child and adult samples. In this sample, those correlations are almost equivalent in size (word-spelling $r = .62$; nonword-spelling $r = .66$).

In the context of the lexical quality hypothesis, this equivalence of the word and nonword correlation scores may suggest that the two strands of skill are yet to integrate, akin to the three factor solution for the less-skilled readers in Perfetti and Hart (2002). If orthography has not yet become the dominant source of information for word recognition, as Mellard et al. (2012b) suggested in their results, the presentation of a word's orthographic form may not be sufficient to activate the orthographic and phonological information for fast and accurate recognition meaning more sources of information are necessary (Perfetti and Hart, 2002). For the younger readers, this is entirely appropriate, as they have time by which the skills may strengthen and become integrated. However, this may be a site of weakness in a reading-related skills profile for the atypically-reading adults.

A further sign of lack of integration may be the very small correlation relationships between processing speed and the reading-related skills. The correlation here between PI and RON is $r = .10$. Swanson et al. (2003) found that weaker correlations existed for their less-skilled readers, with much stronger relationships for individuals who were stronger readers. While processing speed per se is similar to other groups, as observed in the non-significant differences between groups - this weak correlation between PI and RON may jeopardise the mapping of a letter to a sound but also the learning of adjacent relationships that underlie the body of learning that is orthographic knowledge (Kirby et al., 2010).

The strength of the relationships between RON and word / nonword reading skill echo those summarised in Tighe and Schatschneider (2016) ($r = .53$) and suggest that RON still has an important role to play for fluent word recognition in this sample (Hulslander et al., 2010; Mellard et al., 2012b).

Spelling skill differences between groups are present, supporting the inference that spelling skill is variable even amongst skilled readers (Andrews and Lo, 2012). Atypically-reading adult spelling scores are weaker than typically-reading adults', supporting findings by Eme et al. (2014) and Beidas et al. (2013). The strong correlation between spelling and nonword reading is also observed by Beidas et al. (2013).

Our analysis of spelling errors revealed that it was the atypically-reading adults who were the most likely to choose to omit an answer. We could interpret this as a tendency to approach a spelling trial as if the word is something that is known or unknown and can be recalled rather than built from its constituent sound parts at any time.

The error analysis that looked at real-word substitutions across groups may support this interpretation. Typically-reading 16-17-year-olds and adults were more likely to supply alternative homophonic spellings as errors when the target word was a homophone, however the atypically-reading adults were significantly less likely.

This was echoed when we matched errors with target words by `soundex` code. The atypically-reading adults were equivalent with atypically-reading 16-17-year-olds in their propensity for giving errors that differed in sound from the target word. Although making the same amount of errors as typically-reading 16-17-year-olds, more of the errors are less phonological plausible matches. Replication of this finding is needed. If replication confirms the finding, then this could be an indication that atypically-reading adults are less able to exploit phonological information for spelling / word production than typically-reading 16-17-year-olds, with which they share the same word reading performance scores.

Martin-Chang et al. (2014) suggested that a person could be incorrect for spelling but consistency of that erroneous spelling over time would indicate a high quality lexical representation. That does not seem to be in evidence here. Across time, and at the group level, atypically-reading adults demonstrate more varied choices in their spelling attempts than 16-17-year-olds and the typically-reading adults. This is true for orthographic and phonological similarity measures.

More errors of weak sound-matches that also vary over time suggests low lexical quality. This variability looked very like typically-reading 11-12-year-olds in their propensity for matching the `soundex` code of the target word. Over time, this variability would substantially diminish the opportunity for both orthographic learning and for the assimilation of the true statistical distribution of spelling-sound relationships that is orthographic knowledge.

Atypically-reading adults show higher vocabulary raw scores than the 11-12-year-old readers. They show similar vocabulary knowledge in absolute terms to the 16-17-year-old participants. Braze et al. (2007) posited that vocabulary knowledge in adult-learners is underpinned by spoken forms of words rather than printed forms. If this were true, the atypically-reading adults may have semantic knowledge, however the source of the information is from a phonological code only. While vocabulary scores appear to be a relatively strong source of knowledge compared to other reading-related skills for atypically-reading adults, the quality of knowledge may be of a weaker kind and so its benefit to word reading may be weak also.

In summary, the atypically-reading adults in this sample appear to be stronger in skills than adult-learner samples described in previous studies. They show reading-related skills that are similar to students of late secondary school age. Yet, they may still show some signs that reading-related skills are not sufficiently developed to effect efficient word recognition. Notable of these is the lack of dominance of word reading over nonword reading skill, a low correlation between RON and PI scores and weak vocabulary skills that may suggest that their knowledge is predicated upon spoken language experience, and not supported by knowledge of the corresponding orthographic form of the word. Each of these symptoms has been linked to lower skills and slower development of skills for word reading.

Going into the experimental task analyses, we have a picture of the atypically-reading adults. Our next question is whether the similarities or small differences observed between the groups here manifest as quantitative or qualitative differences in how this sample of readers uses the psycholinguistic information contained in the items of the single word reading tasks. We turn to this next.

7 Results: Experimental Tasks

We present the findings of four experimental tasks across accuracy and reaction time data. Each section lists item properties and average performance over time by group. We present the preferred model, findings of the analyses for full sample, complete case and outlier-removed data sets and model predictions.

Without exception, the information criterion values were lower for Bayesian inference than frequentist models. Consequently, we took the models with the lowest LOOIC values as the most compatible with the data for this sample and present them as the preferred models. Information criterion values for all models are listed in Appendix H. Information around model diagnostic check routines are in Appendix I for each task and outcome measure.

The design of the study includes six groups, however the data of the study suggest one (see section 6.2.7 for cluster analysis findings). We made a decision to use model selection as our decision strategy for which set of findings to present as the "preferred" model, allowing the data to direct us. The design implied model is the Additive-RIS model. Since the study is exploratory in nature, in each section we also present the coefficients for the Additive-RIS model and briefly describe it in an effort to keep a space for possibilities open. There are different decisions that can be made around study design and analysis for future research, and foreclosing on one model early in the process seems premature.

We present accuracy model coefficients on the log-odds scale, where zero is the critical threshold. Values above zero indicate higher odds of a correct response. Values below zero indicate lower odds of a correct response, relative to the average response given in the intercept. In the narrative, we translate the log-odds units into probabilities for ease of interpretation. For interpretation of log-odds estimates in terms of effect size, Rosenthal (1996) suggests the following scale:

- log-odds 0.4 = small or weak
- log-odds 0.9 = medium or moderate
- log-odds 1.38 = large or strong
- log-odds 2.30 = very large or very strong

Additionally, we adopt a "very small" label for log-odds estimates < 0.4.

We analysed reaction time data using linear mixed-effects models. Recall that reaction time data is log10 transformed before being standardised by task and condition using the mean and standard deviation value of the typically-reading 16-17-year-olds at T1. As a result, the intercept represents the mean reaction time for the predictors at an average of 0 and the reference levels of categorical predictors. The reference group is the typically-reading 16-17-year-olds. Positive coefficients for reaction time indicate slower reaction times for a 1 standard deviation increase in a predictor. Negative coefficients indicate faster reaction times for a 1 standard deviation increase in a predictor.

7.1 Letter Search

Details of the items and procedure are in section 5.2.4.1. As a reminder, participants had to identify whether a previously displayed target letter was present in the subsequent presentation of a letter string. The letter string could be either a real word or an unpronounceable nonword. The target letter could either be absent, or present at any one of the five letter positions.

Our research question was whether atypically-reading adults were detecting the presence or absence of a letter at a similar rate and accuracy to the other groups. If the condition of group interacted with any of the ID or psycholinguistic measures, this could indicate a difference in either strategy or knowledge for completing trials. No interactions between the group variable and ID or psycholinguistic measures would suggest that the groups are approaching the task similarly. Ziegler et al. (2008), measuring only individual differences, found deficits on error rates rather than speed

Table 7.1

Descriptive Statistics for Frequency for Three Item Lists in the Letter Search Task

	Mean Frequency (SD)	
List	High	Low
1	5 (0.1)	2.9 (0.1)
2	5 (0.1)	2.9 (0.1)
3	5.1 (0.1)	2.9 (0.1)

between a group of developmental dyslexic young readers (mean age 9:10 years) and typically developing young readers, with no difference between words and nonwords.

7.1.1 Item Properties

Mean scores for low and high frequency values across lists 1 - 3 are in Table 7.1. A two-way ANOVA for effects of frequency category (low and high) and list (1-3) on frequency ratings confirmed a significant main effect of frequency category ($F(1, 54) = 5215.72$, $p < .001$) and a non-significant main effect of list ($F(2, 54) = 0$, $p = .998$). Thus, our design to incorporate low and high frequency words is supported and we can infer no differences for frequency ratings across the lists.

The descriptive statistics for properties of psycholinguistic variables for items are displayed in Table 7.2 with distributions by variable and list displayed in Figure 7.1. Results of a series of ANOVA tests for differences between lists within psycholinguistic variables indicated that none of the variable means differed significantly between lists (all ps $> .18$).

7.1.2 Analyses

The modelling strategy was described in section 5.2.5.5. There is a task specific predictor of "position" for the letter search task. This predictor has six levels (none; first; second; third; fourth; fifth) that indicates the letter position at which a letter occurred for a present trial. The level of "none" is the reference level.

Table 7.2

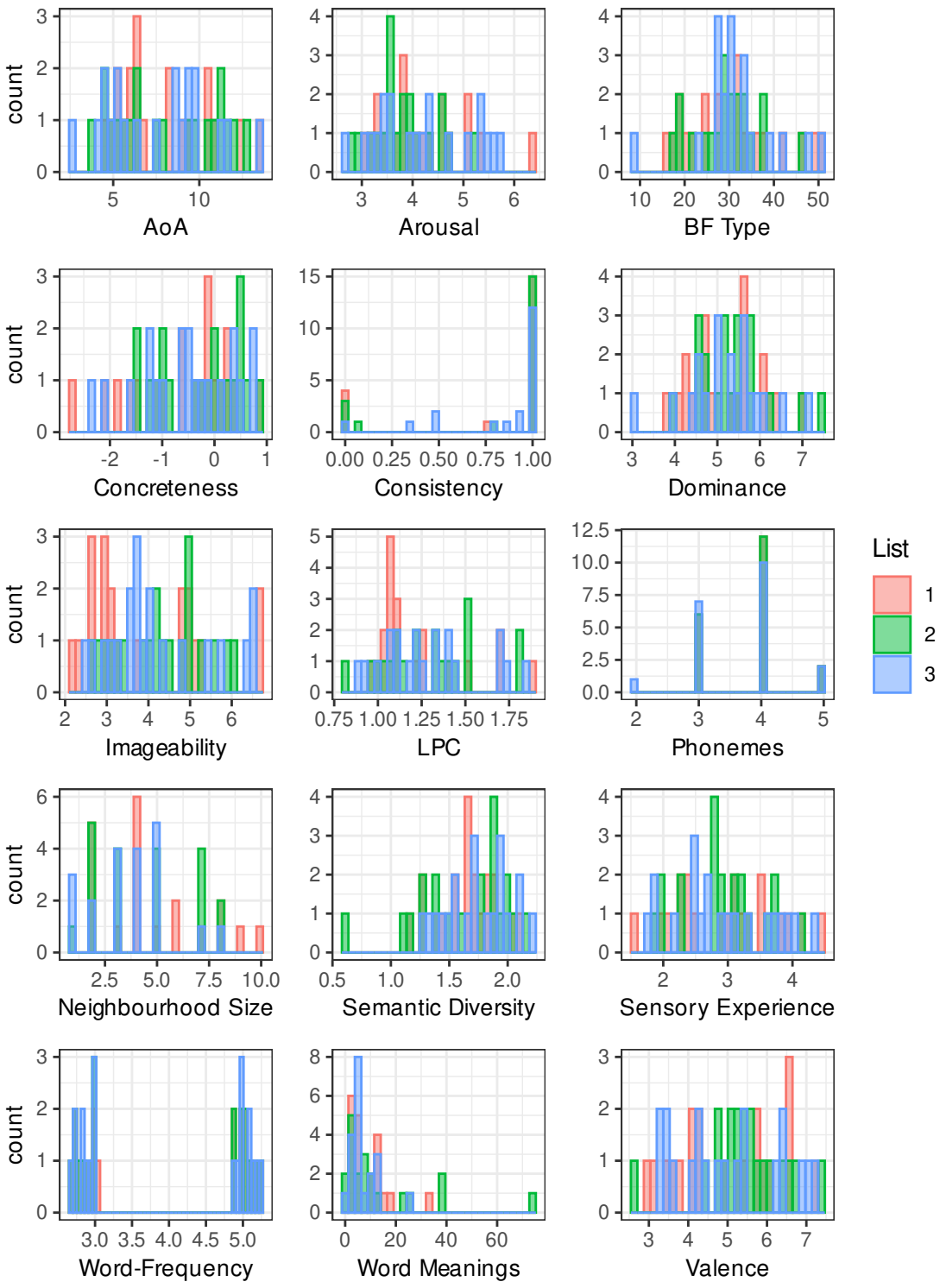
Summary of Psycholinguistic Variable Measures for Letter Search Word Items with F-Ratio and P Values to Signify Differences Between Item Lists

					ANOVA	
Psycholinguistic Variables	Mean	SD	Min	Max	F(2, 57)	p
AoA	7.8	2.8	2.8	13.7	0.17	0.84
Arousal	4.1	0.8	2.6	6.3	1.41	0.252
BF Type	30.5	8.9	8.2	50.2	0.27	0.768
Concreteness	3.6	0.9	1.5	5.0	0.98	0.383
Consistency	0.8	0.4	0.0	1.0	0.14	0.87
Dominance	5.3	0.9	3.1	7.5	0.16	0.851
Imageability	4.1	1.2	2.2	6.7	0.66	0.522
LPC	1.3	0.3	0.8	1.9	0.35	0.709
Phonemes	3.8	0.7	2.0	5.0	0.34	0.711
Neighbourhood size	4.3	2.3	1.0	10.0	1.09	0.342
Semantic diversity	1.7	0.3	0.6	2.2	0.75	0.478
Sensory experience	2.9	0.7	1.5	4.5	0.02	0.983
Word frequency	4.0	1.1	2.7	5.2	0	1
Word meanings	9.9	12.2	1.0	75.0	1.75	0.182
Valence	5.1	1.2	2.7	7.4	1.05	0.357

Note:
AoA = Age of acquisition. BF = Bigram frequency. LPC = Levenshtein Phonological Consistency.

Figure 7.1

Histograms Showing the Distribution of Psycholinguistic Properties of Items for the Letter Search Task, Across Three Lists



Additionally, for this and the lexical decision data, we tested for a word-superiority effect, running a further analysis that included data for both words *and* nonwords plus ID measures. In this model, we included only trials where a target letter was present (across letter position 1 - 5). We expected that the direction of effects would follow the literature, with more accurate letter identification when the letter was presented in a word.

7.1.2.1 *Number of Observations*

Full Sample. We collected 23,280 observations across words and nonwords in the letter search task. We excluded 480 observations for 12 participants as being duplicate items from previous waves of data collection. Nineteen observations that were made < 200 ms were also removed as technical malfunctions. This left 22,781 observations across all letter positions for words and nonwords.

To test the word superiority effect, we removed observations in the "none" position (n = 11,392) giving 11,389 observations.

To measure the impact of ID and psycholinguistic measures on accurate responses, we extracted the word trials (n = 11,388). We removed incorrect responses for reaction time analyses (n = 1413), leaving 9,975 observations.

Complete Case Analysis. The number of participants who completed three data collection sessions across that included the letter search task was 161. There were 18,679 observations for words and nonwords available for a complete case analysis. The analysis was repeated using the preferred model for accuracy (n = 9,356) and reaction time outcomes (n = 8,402).

Outlier Analysis. After removing timed-out observations (n = 668, 2.3%), interquartile ranges per participant were calculated and outliers identified (see section 5.2.5.6) and removed (n = 2,241, 9.8%), leaving 19,872 word and nonword observations. The analysis was repeated using the preferred model for accuracy (n =

10,015) and reaction time outcomes (n = 9,096). In the complete case analyses with no outliers, the number of observations for accuracy analyses was 8,400; for reaction time analyses n = 7,639.

7.1.3 Accuracy Results

7.1.3.1 *Descriptive Statistics*

Accuracy rate across the sample for word and nonword items was 87%. We calculated mean accuracy rates per participant per time point and display them by group for words and nonwords in Figure 7.2; averages across accuracy and reaction time by position, time and group are displayed in Figure 7.3 and 7.4 for words and nonword respectively. Most groups became less accurate between the first and third data collection sessions. Just looking at mean performance by condition and group, there is no clear visual evidence of a word superiority effect.

Word Superiority Effect for Accuracy Responses. The model for a word superiority effect showed that while the coefficient indicated a minute advantage of an accurate response in nonwords (mean = 88.6, SD = 31.7) over words (mean = 87.6, SD = 33), the credible intervals indicated uncertainty around the effect, such that a zero difference was plausible (log-odds = 0.07 [-0.22, 0.34]). We conclude that the data for this sample does not support evidence of a word superiority effect. Participants did not identify letters that were present in words any more accurately than letters that were present in nonwords.

7.1.3.2 *Preferred Model*

When the target letter is not present in the item, the accuracy measure represents a correct response of “no”. When a target letter is present in the item, the accuracy measure represents a correct response of “yes”. For the position predictor variable, position = “none” is the reference level.

Figure 7.2

Histograms Showing the Distribution of Mean Accuracy Rates per Participant by Groups Across Time Points for Words and Nonwords in the Letter Search Task

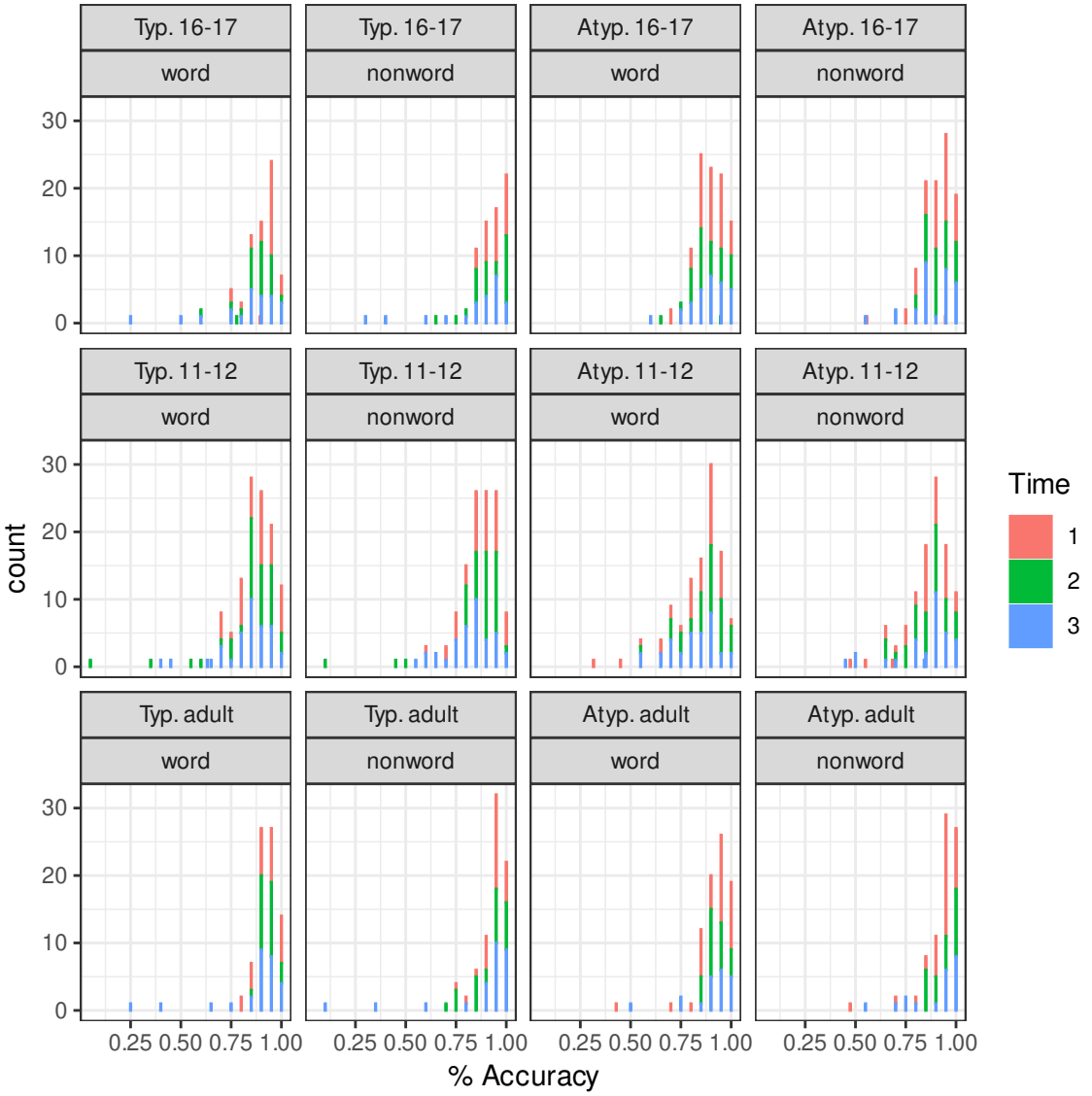


Figure 7.3

Accuracy and Mean RT By Group, Position and Time for Letter Search Words

	Time 1			Time 2			Time 3		
Position	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typically-reading 16-17 years									
none	93.5	930.9	390.4	88.4	968.1	327.7	80.9	957.4	426.5
first	100.0	767.5	243.2	93.9	780.1	258.3	93.2	778.8	276.4
second	86.5	934.6	281.1	80.0	925.0	273.3	79.5	931.9	291.5
third	86.5	849.7	284.9	89.8	901.1	311.4	86.4	783.3	208.9
fourth	88.2	872.3	255.4	88.0	929.4	295.5	86.4	948.3	284.6
fifth	90.4	901.5	366.6	84.0	944.1	312.6	81.8	1008.5	379.6
Atypically-reading 16-17 years									
none	90.2	893.2	363.4	86.6	919.8	327.4	84.5	937.0	343.7
first	97.7	721.5	285.3	89.1	790.9	363.0	96.6	785.4	294.9
second	81.4	935.8	419.9	90.5	955.3	435.3	87.9	920.2	266.8
third	96.5	794.8	319.0	95.3	827.9	281.3	96.6	876.7	391.3
fourth	86.0	856.9	236.9	87.5	924.2	421.3	96.6	871.9	251.3
fifth	82.6	915.8	320.4	87.5	953.9	320.5	87.9	961.5	281.2
Typically-reading 11-12 years									
none	87.9	1128.6	397.7	82.4	1091.8	399.2	80.3	1082.0	391.1
first	96.4	1015.8	475.4	86.9	974.6	284.0	94.6	948.6	411.6
second	85.7	1164.7	371.2	81.0	1067.5	335.8	85.1	1142.9	424.0
third	88.1	1014.4	388.2	89.3	1018.7	330.9	82.4	1043.0	460.3
fourth	85.7	1046.8	444.3	82.1	1122.4	390.0	83.8	1015.3	401.8
fifth	82.1	1069.2	286.1	85.7	1138.8	324.5	79.5	1090.5	332.1
Atypically-reading 11-12 years									
none	84.6	1131.7	387.4	88.9	1085.1	350.8	77.8	1084.0	364.6
first	85.9	939.5	288.8	90.5	1070.5	451.4	92.2	952.0	333.1
second	80.8	1201.0	357.5	81.1	1079.8	347.1	82.8	1183.8	480.3
third	91.0	1084.7	370.7	85.1	1027.2	344.6	84.4	963.3	330.6
fourth	73.1	1096.5	304.7	86.5	1082.4	301.4	84.4	1064.3	302.6
fifth	74.4	1137.7	299.0	82.4	1100.5	347.7	81.2	1051.9	316.8
Typically-reading Adult									
none	91.4	1045.7	347.9	92.3	1063.7	341.2	87.0	1036.7	319.8
first	94.6	809.7	225.5	96.2	948.4	274.1	87.0	892.3	267.9
second	96.4	1058.0	327.5	96.2	1052.4	281.2	87.0	1003.3	246.0
third	92.9	982.0	377.8	90.4	1027.8	373.0	85.2	900.8	250.7
fourth	98.2	971.5	258.3	100.0	1056.5	246.7	88.9	1002.1	273.0
fifth	85.7	916.4	252.0	86.5	983.9	192.8	85.2	966.5	241.0
Atypically-reading Adult									
none	91.8	941.5	330.6	91.2	893.5	304.8	85.5	934.6	320.4
first	93.3	767.7	208.0	98.0	754.4	183.4	97.5	826.3	195.3
second	89.5	948.6	234.8	92.0	966.8	305.9	90.0	965.6	279.0
third	92.1	864.4	282.7	90.0	798.3	227.8	95.0	922.2	380.4
fourth	90.7	876.5	204.5	100.0	867.6	195.9	95.0	1009.2	462.8
fifth	93.2	861.7	250.2	86.0	852.2	279.5	97.5	906.8	292.0

Note:
Acc % = Percentage Accuracy; RT (ms) = Reaction time in milliseconds

Figure 7.4

Accuracy and Mean RT By Group, Position and Time for Letter Search NonWords

	Time 1			Time 2			Time 3		
Position	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typically-reading 16-17 years									
none	97.3	926.0	301.1	95.2	971.6	351.6	85.5	918.1	330.4
first	100.0	773.5	293.1	94.0	781.5	190.8	86.4	788.8	334.3
second	88.5	928.2	300.8	86.0	1046.1	359.4	86.4	977.9	292.9
third	86.5	948.9	358.9	80.0	906.1	267.4	72.7	818.4	288.7
fourth	92.3	968.8	232.9	90.0	956.4	221.1	88.6	984.2	270.9
fifth	90.4	894.1	254.8	88.0	1005.4	280.4	84.1	1018.7	298.8
Atypically-reading 16-17 years									
none	95.1	881.7	296.8	94.4	948.7	395.3	89.0	932.1	334.8
first	93.0	771.7	339.7	95.3	712.4	195.2	91.4	775.3	225.1
second	91.9	918.9	339.3	89.1	941.9	329.4	86.2	960.8	284.8
third	74.4	861.0	295.1	82.8	932.8	369.3	87.9	945.6	516.0
fourth	83.5	1041.5	387.6	85.9	953.0	318.1	89.7	1036.3	291.1
fifth	84.9	886.8	328.0	87.5	1006.6	473.4	86.2	949.5	326.4
Typically-reading 11-12 years									
none	91.2	1152.1	384.7	89.8	1116.5	405.7	84.6	1087.1	382.5
first	96.4	959.9	398.2	89.3	984.5	345.0	90.5	951.7	378.4
second	86.9	1084.1	280.1	84.5	1151.8	327.6	77.0	1121.9	269.6
third	69.0	1136.9	400.8	69.0	1009.1	265.8	81.1	1192.8	475.0
fourth	83.3	1202.9	387.9	75.0	1166.7	365.5	70.3	1224.3	374.2
fifth	82.1	1123.6	302.9	88.1	1143.4	319.9	79.7	1119.7	354.3
Atypically-reading 11-12 years									
none	88.4	1147.8	350.8	90.5	1112.4	359.1	87.1	1120.9	353.6
first	92.3	979.9	349.6	87.8	1003.0	326.8	87.5	981.0	332.7
second	87.2	1141.9	294.1	82.4	1118.9	257.9	79.7	1127.1	345.1
third	69.2	1099.4	480.8	74.3	1086.8	400.2	81.2	1046.9	292.1
fourth	80.8	1193.6	358.2	78.4	1212.7	325.8	78.1	1091.2	358.0
fifth	74.4	1114.5	285.9	82.4	1089.8	274.0	87.5	1194.8	496.7
Typically-reading Adult									
none	97.5	1103.2	356.4	92.7	1118.5	356.8	90.4	1031.4	284.6
first	96.4	891.0	282.7	100.0	924.4	214.5	90.7	922.0	268.8
second	92.9	1068.6	255.9	84.6	1083.2	254.7	88.9	1041.6	327.4
third	80.4	998.0	276.3	73.1	1015.6	296.0	81.5	1050.2	361.4
fourth	89.3	1168.0	356.3	92.3	1182.8	307.7	85.2	1089.0	302.3
fifth	89.3	1061.7	278.6	94.2	1054.8	223.5	88.9	1082.8	283.6
Atypically-reading Adult									
none	97.4	922.0	301.8	95.2	938.2	317.4	89.0	906.6	269.8
first	96.1	751.0	192.1	98.0	801.3	246.4	97.5	890.9	344.9
second	85.5	931.1	210.5	92.0	941.4	242.5	95.0	1035.5	277.7
third	82.7	883.1	262.5	90.0	904.5	282.1	82.5	1002.3	303.4
fourth	88.2	1022.2	248.3	92.0	992.5	265.0	92.5	1046.3	310.2
fifth	86.8	944.2	270.8	90.0	938.1	274.8	95.0	1004.4	421.4

Note:
Acc % = Percentage Accuracy; RT (ms) = Reaction time in milliseconds

The preferred model for the letter search accuracy data was the Base-RIS model. This included predictors for letter position, group contrasts and individual differences as fixed effects with random intercepts and slopes on participants and items. The model satisfied diagnostic checks. The explained variance in the accuracy outcome was $R^2_{bayes} = 20.4\%$ [18.7, 22.1]. The coefficients for the fixed effects in the model, on a log-odds scale with 95% credible intervals, are presented in Table 7.3 and in Figure 7.5. We briefly explain the dot-and-whisker plot to aid interpretation.

The "dot" represents the mean value of the posterior distribution for that predictor variable, i.e. it is the most plausible value given the model, the data and any prior information when all other predictors are held constant at a mean of zero. Recall that predictors are standardised, such that all mean values are 0 with a standard deviation value of 1.

The "whisker" represents the range of alternative yet credible values of the coefficient that lie within the 95% probability mass of the posterior distribution. Probability for lower and higher values decreases as the value moves away from the mean. The wider the "whiskers", the greater the range of plausible values which decreases our certainty of the posterior distribution estimate.

Further, we have drawn a dashed, vertical reference line at '0' on the x-axis. Positive values to the right of the line indicate higher log-odds of an accurate response with a standard deviation increase in the predictor; negative values to the left of the line indicate lower log-odds of an accurate response with a standard deviation increase in the predictor.

Where a *whisker crosses the critical value of 0*, this indicates that the model is uncertain about the direction of the effect, as it includes a range of both positive and negative values within the 95% probability mass. We refer to such estimates as "unreliable". When a whisker does not cross zero, we refer to estimates as "reliable".

Model Inference. The intercept reflects the mean rate of accuracy on the trial condition where no target letter was present, with a probability of making an accurate response of approximately 94% where all predictors are at their mean level and the

Table 7.3

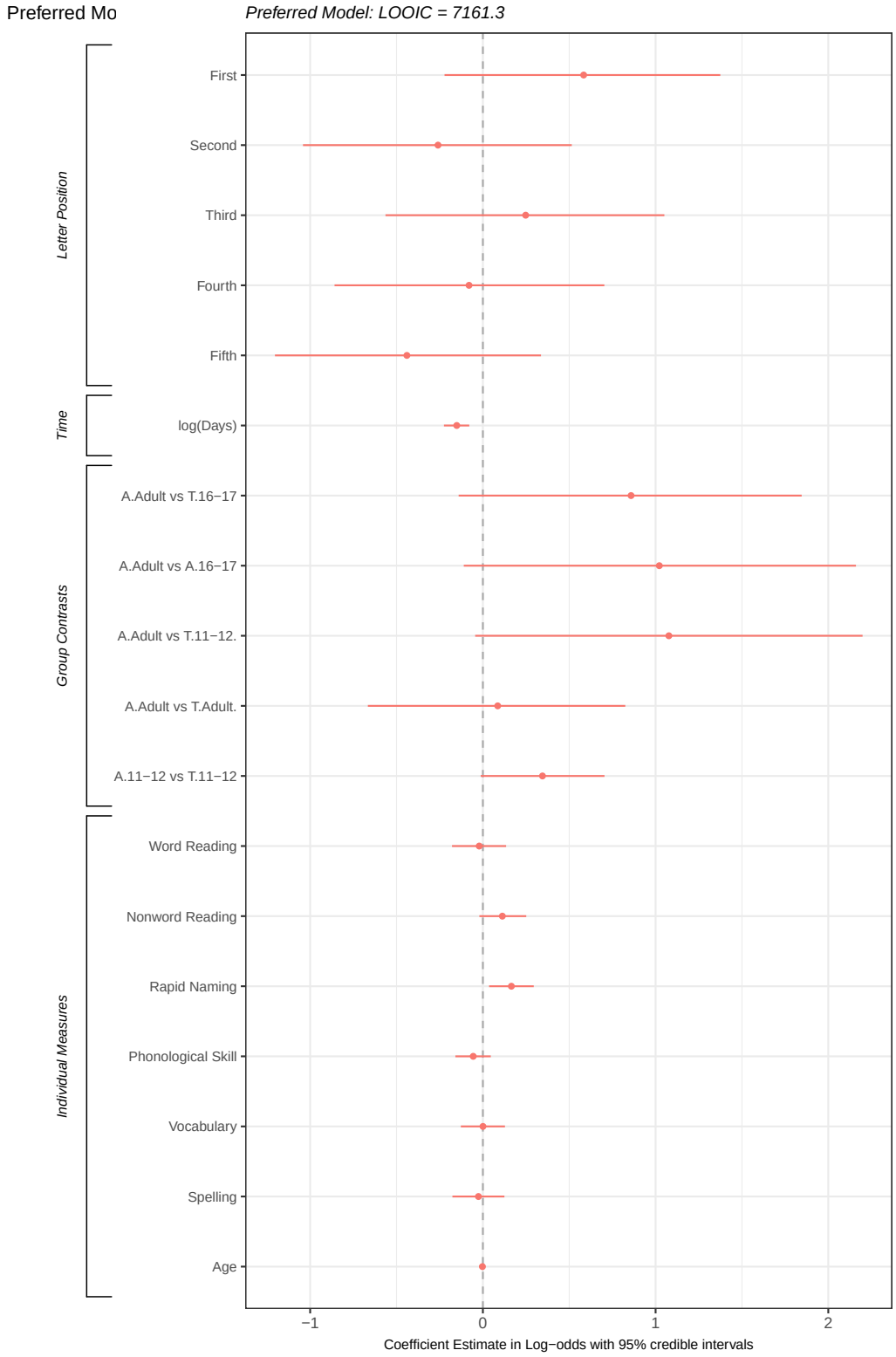
Summary of Standardised Fixed Effects for Letter Search Accuracy

Term	Estimate	SE	Lower CI	Upper CI
Intercept	2.78	0.20	2.39	3.17
Position				
First	0.58	0.41	-0.22	1.37
Second	-0.26	0.39	-1.04	0.51
Third	0.25	0.41	-0.56	1.05
Fourth	-0.08	0.40	-0.86	0.70
Fifth	-0.44	0.40	-1.20	0.34
Time				
(Log) Days	-0.15	0.04	-0.23	-0.08
Group Contrasts				
A. Adult vs T. 16-17	0.86	0.51	-0.14	1.85
A. Adult vs A. 16-17	1.02	0.58	-0.11	2.16
A. Adult vs T. 11-12	1.08	0.57	-0.04	2.20
A. Adult vs T. Adult	0.09	0.38	-0.67	0.82
A. 11-12 vs T. 11-12	0.35	0.18	-0.01	0.70
Individual Differences				
Word reading	-0.02	0.08	-0.18	0.13
Nonword reading	0.11	0.07	-0.02	0.25
Rapid naming	0.17	0.07	0.04	0.29
Phonological skill	-0.06	0.05	-0.16	0.05
Vocabulary	0.00	0.07	-0.13	0.13
Spelling	-0.03	0.08	-0.18	0.12
Age	0.00	0.00	-0.01	0.01

Note:
CI = Credible intervals. A. Adult = Atypically-reading adult; T. 16-17 = Typically-reading 16-17-year-old; A. 16-17 = Atypically-reading 16-17-year-old; T. 11-12 = Typically-reading 11-12-year-old; T. Adult = Typically-reading adult; A. 11-12 = Atypically-reading 11-12-year-old.

Figure 7.5

Estimates from the Posterior Distribution of the Preferred Model for ID and Psycholingustic Predictors on Letter Search Accuracy Data



reader is a typically-reading 16-17-year-old. The first and the third letter position shows a trend for greater accuracy, with second, fourth and fifth trending on lower accuracy. However, the credible intervals all cross zero rendering the estimates unreliable in this model and sample.

The effect of log (days) is estimated with much greater certainty. It is estimated as negative for this model and sample. For a 1 SD increase in number of days from day 0, the probability of an accurate response decreases by approximately 1%. Essentially, participants were reliably less accurate by their third data collection session.

Although the model with the group contrast did show the lowest LOOIC value, none of the group contrast estimates are reliable. The trend in the data is for the atypically-reading adults and the atypical 11-12-year-olds to show greater odds of responding correctly than the contrasted group. The difference is much smaller when compared with typically-reading adults. The mass of probability suggests that the atypically-reading adults are more likely to be correct, however, the results implied by this model and this data are inconclusive.

RON is the single ID measure with credible intervals that do not cross zero (log-odds = 0.17 [0.04, 0.29]). With a 1 SD increase in RON score, the probability of making an accurate response increases by approximately 1%. This is a very small effect.

Each of the remaining predictors have very small values and credible intervals that cross zero. It is likely that there is insufficient data to estimate such small effects with confidence.

Complete Case and Outlier Analyses. Letter position coefficients remained stable with respect to size, direction and uncertainty. Accuracy across time log (days) becomes positive, but unreliable in the complete case and outlier analyses, potentially indicating that under different sample conditions, there is a greater likelihood for more accurate responses with each data collection point. However, for this sample, the credible intervals lie on or cross zero.

Where each coefficient for the group contrasts is positive in the full sample, preferred model, the coefficients for the atypically- and typically-reading adult contrast and the atypically- and typically-reading 11-12-year-old change direction. Credible intervals still cross zero. The 11-12-year-old effect remains positive in the full sample data without outliers but credible intervals are wider, indicating a greater range of plausible values. The effect becomes less certain once outliers are removed.

RON remains stable across the sensitivity analysis models. There is no change in the remaining ID measures, either. Taken together, the sensitivity analyses suggests that, but for RON effects, the effects for letter search accuracy data may be dependent upon the sample herein and generalisation should be done with caution.

Letter Search Accuracy Design Implied Model. The design implied model is displayed in Figure 7.6. As stated at the beginning of the chapter, this is always the Additive-RIS model. For the letter search task, it also includes the position predictor.

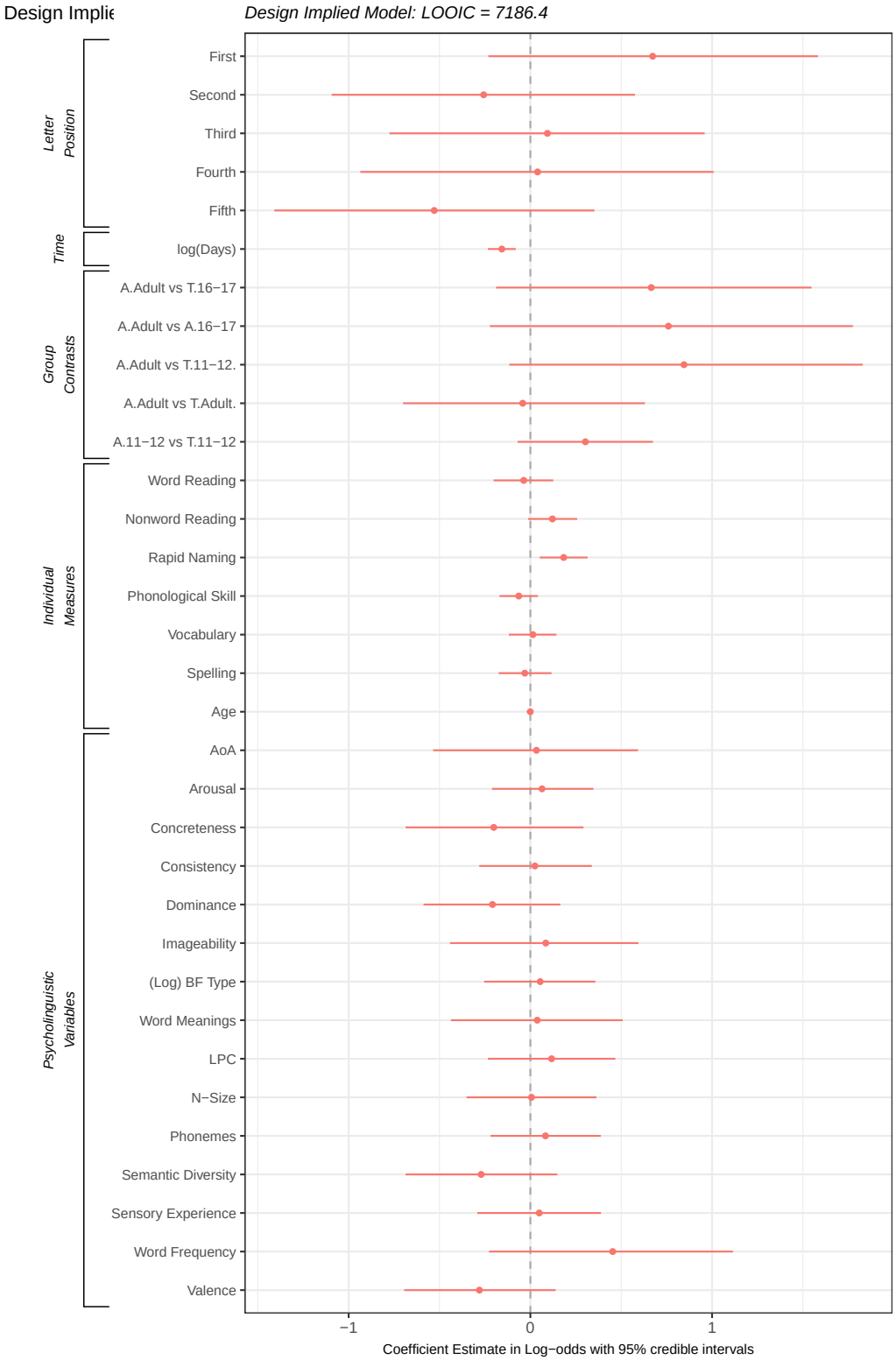
The coefficients for letter position, log (days) and group are comparable across the preferred and design implied models. Notice though that the coefficient for the planned contrast between atypically-reading adults and typically-reading adults is negative in the design implied model, opposite to that of the preferred model, indicating that atypically-reading adults have lower odds of making a correct response. A further note of caution in interpreting the group predictor. This caution is echoed in the much wider credible intervals in the preferred model that describe a high level of uncertainty around the posterior estimates.

In the design implied model, the coefficients for psycholinguistic variables are displayed towards the bottom of the plot. All the credible intervals include zero such that the possibility of no effect and reverse directions of effects are plausible according to estimates from the posterior distribution. Given the manipulation of frequency values in the sample, it is surprising that even the word-frequency effect is unreliable.

Model Predictions. We can use the model to generate predicted variation of the impact of a particular predictor by providing new data and updating the model

Figure 7.6

Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on Letter Search Accuracy Data



implied estimates under those new conditions. Model implied predictions for accuracy data in the letter search task are shown in Figure 7.7. These plots illustrate predictions with credible intervals (blue bands) derived from SD values. Residual error variance values were so large that their inclusion made the plots unreadable.

Since the preferred model for letter search accuracy contains a small number of predictors, we can illustrate model predictions for each of them.

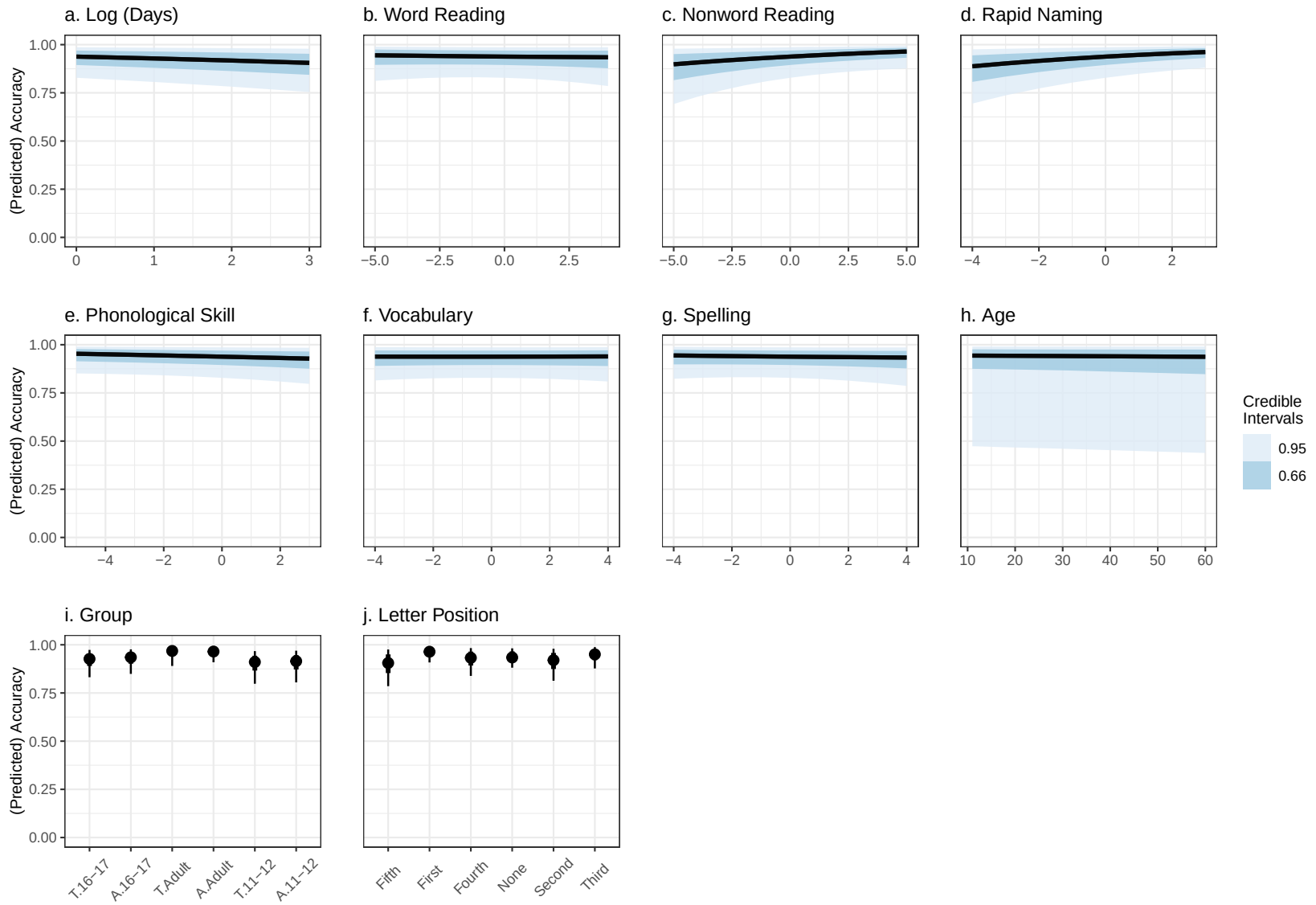
Plots (a) - (h) of Figure 7.7 display variation in the probability of making a correct response for different levels of a predictor. The solid black line represents the mean prediction value while the blue bands represent 66% and 95% credible intervals.

For every plot, the y-axis represents the probability of making a correct response. From left to right on the x-axis, the values in each plot denote low to high levels of skill in the predictor, while keeping all other predictors at their mean value in the model. A rising slope indicates the trend for the accuracy to increase with increasing skill. A falling slope indicates that accuracy rates decrease with increasing skill. Plots (i) and (j) show the model implied predictors for the rates of accuracy for the categorical predictors of group and letter position.

The first thing to notice is that, the expected predicted probability of an accurate response to words in the letter search task never falls much lower than approximately 87%, this is in line with the observed data. Even at the lowest skills levels for nonword reading (plot c) and rapid naming (plot d), the expected probability levels appear to be approximately 87%. The log (days) predictor (plot a) and phonological skill show a weak, negative relationship with the probability of a correct response over varying levels of skill. The model implied inference is that accuracy will slightly decrease over time, and is lower for higher levels of phonological skill in the ability to isolate single phonemes within words.

Figure 7.7

Preferred Model Predictions for the Effects of Individual Differences, Group and Letter Position on Letter Search Accuracy Performance



Nonword reading and rapid naming (plot c and d, respectively) show positive relationships. The credible intervals are wider at lower levels of skill, indicating a greater range of accuracy rates for lower levels of skill than for higher levels of skill. As foreshadowed by the dot and whisker plot, word reading skill (b), vocabulary (f) and spelling (g) show little variation across skill levels on the probability of making an accurate response. The predictions for age correspond to the observed data model, with the credible intervals displaying that while the observed data suggest high rates of accuracy for all age levels in this model and sample, the age predictor is compatible with chance levels of accuracy.

Summary and Discussion. The preferred model for the letter search accuracy data included variables for letter position, group and ID measures only with ID measures as random intercepts and slopes. Even though this model included the group contrast predictor, the effects are unreliable.

Accuracy for correctly identifying that a letter *was not* present in a word was ~ 87% in this sample. The highest rate of accuracy for correctly identifying that a letter was present was for letters at the first position (93.2%). The accuracy level suggests that participants in the sample have the knowledge to perform the later tasks and differences between group or findings are not related to a lack of foundational, letter-level knowledge.

The data and the model were also inconclusive about the presence of a word superiority effect. Mean values show that accuracy levels for word and nonwords were equivalent. This sample was not assisted by letter identification when it was embedded in a word rather than an illegal letter string. This finding converges with the younger child reading sample of Ziegler et al. (2008). This could be an effect of the type of nonwords used. We used unpronounceable nonwords here, to reduce the use of implicit knowledge for transitional probabilities between letters. This type of nonword may have encouraged a serial search strategy because, for at least half the items, the relationships that would normally exist between letters in words or pronounceable nonwords were not present.

Support for this interpretation may be suggested by the small effect of RON and lack of influence coming from word reading skill, vocabulary or any psycholinguistic predictors. The model implied coefficients suggests that the differences in word knowledge that we saw in the descriptive statistics analyses are not relevant to the accuracy outcome and that participants are not using word reading skill or word meanings to help identify words.

A further indication that lexicality may have been made redundant is the lack of a reliable frequency effect (see design implied model Figure 7.6). We may be over-interpreting here, but the lack of influence for any psycholinguistic variables may suggest that words are not being accessed because they are not salient units within the task.

Although with a high level of uncertainty, the values with the highest probability mass in these models predicted that the atypically-reading adults and atypically-reading 11-12-year-olds were more accurate at this task than the groups with which they were contrasted. Ziegler et al. (2008) found deficits for their young readers with dyslexia compared to readers without dyslexia. If individual differences were associated with this finding, we would expect to find no difference as the atypically-reading adults tended to be equivalent with their peers across the ID measures. Looking across the remaining tasks for the same kind of accuracy advantage may help suggest explanations for this finding.

7.1.4 Reaction Time Results

7.1.4.1 *Descriptive Statistics*

Distributions for mean reaction time in milliseconds per participant are displayed at the group level in Figure 7.8 for correct responses to words and nonwords and in Figure 7.3 and 7.4. In the models that follow, when the target letter is not present in the item, the reaction time represents the latency between onset of the item and a correct response of “no”. When a target letter is present in the item, the reaction time

represents the speed of correct responses that answer "yes".

Recall that reaction time data is log10 transformed before being standardised by task and condition using the mean and SD values of the typically-reading 16-17-year-olds at T1 as the reference group (word RT mean = 897.3 ms, SD = 347.1 ms; nonword RT mean = 913.3 ms, SD = 299.0 ms). As a result, '0' on the x-axis represents the mean of the outcome data with reference to a typically-reading 16-17-year-old. Positive values for reaction time indicate slower reaction times for a 1 SD increase in a predictor. Negative values indicate faster reaction times for a 1 SD increase in a predictor.

7.1.4.2 *Preferred Model*

As with the accuracy measures, the preferred model for reaction time data in the letter search task was the Base-RIS model. LOOIC values for models with and without a group predictor were almost equivalent, however the model without a group was favoured. The model satisfied diagnostic checks. The explained variance in the reaction time outcome for this model and data was R^2_{bayes} = 34.0% [32.8, 35.1]. The coefficients for the fixed effects of the model are presented in Table 7.4 and Figure 7.9.

Model Inference. The coefficients for the first, third and fifth position suggest a faster response time than the average response for a target letter being absent. The second and fourth letter positions are estimated as being slower. Credible intervals for the second, fourth and fifth letter position cross zero, meaning the effects are unreliable and a range of both positive and negative relationships are compatible with the model and the data. The first letter is approximately 160 ms faster than a correct response for an absent letter (log odds = -0.46 [-0.64, -0.27]). The third letter is approximately -97 ms faster.

The positive coefficient for log (days) (log-odds = 0.04, [0.02, 0.06]) indicates that reaction time increases over data collection sessions. This is a small, reliable effect, however it is estimated with a high level of certainty. This equates to an

Figure 7.8

Histograms Showing the Distribution of Raw Mean Reaction Time (ms) per Participant by Group for Words and Nonwords in the Letter Search Task

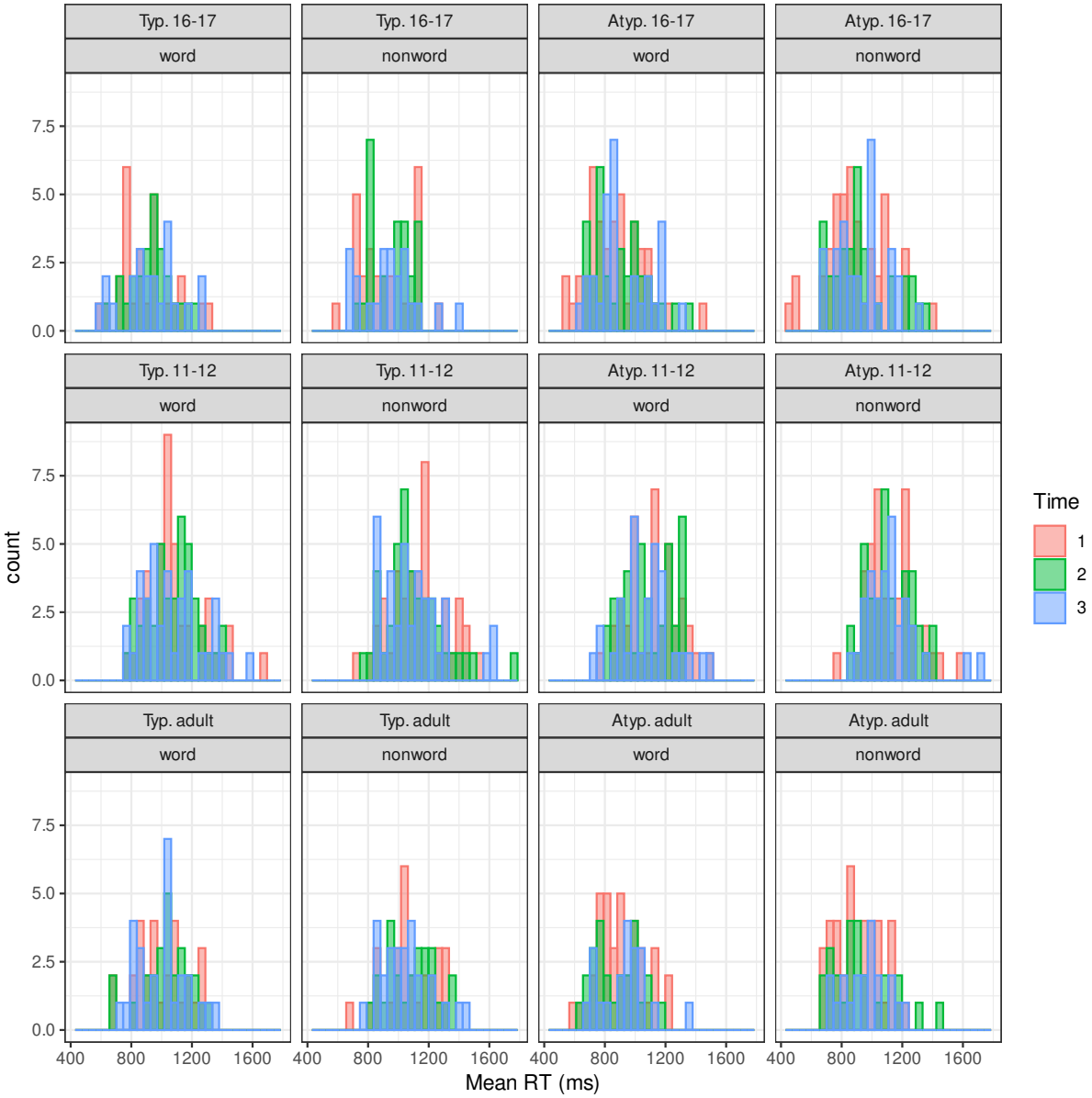


Table 7.4

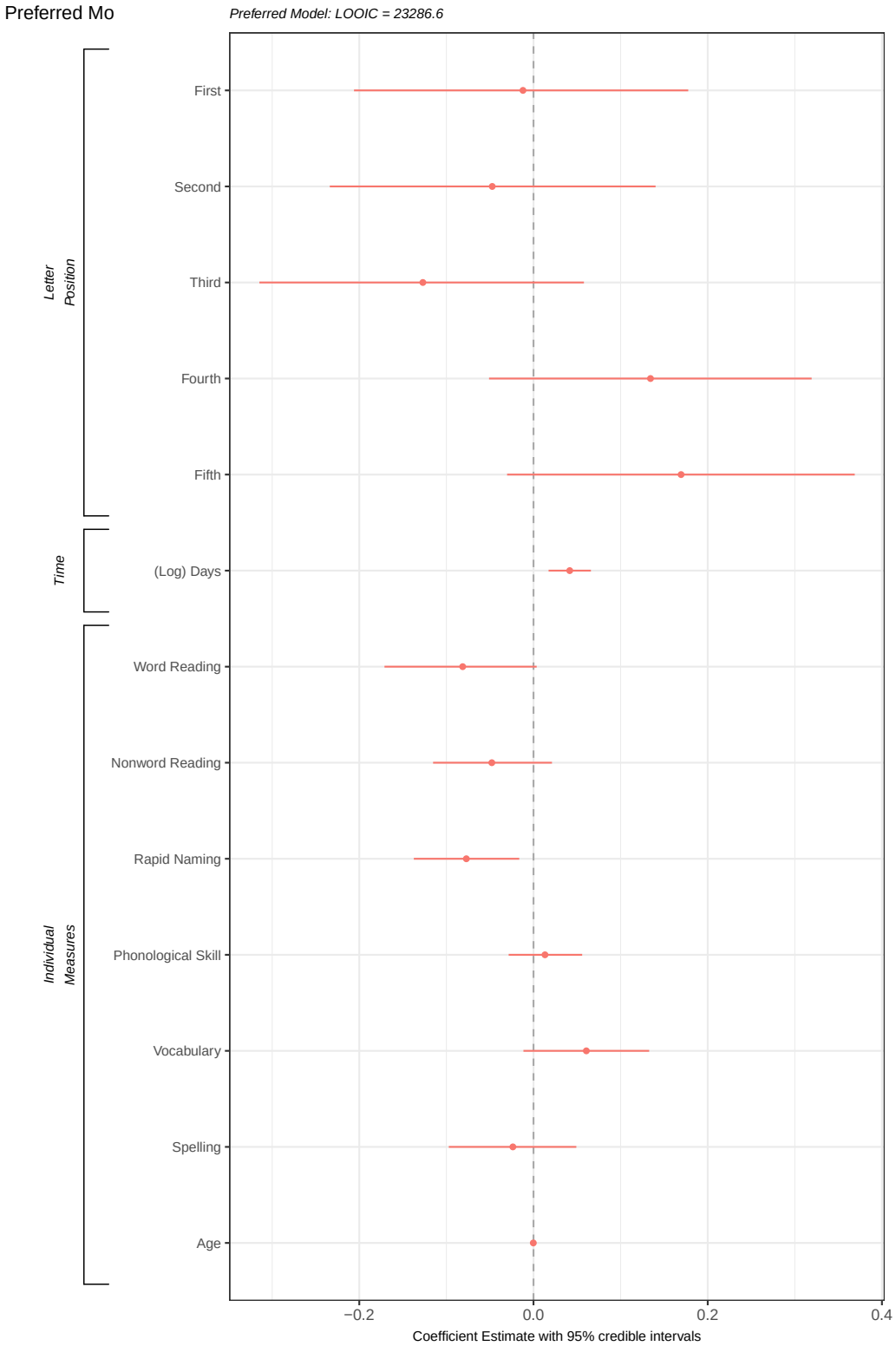
Summary of Standardised Fixed Effects for Letter Search Reaction Time

Term	Estimate	SE	Lower CI	Upper CI
Intercept	0.21	0.06	0.09	0.33
Position				
First	-0.46	0.09	-0.64	-0.27
Second	0.09	0.09	-0.09	0.27
Third	-0.28	0.09	-0.46	-0.10
Fourth	0.01	0.09	-0.17	0.19
Fifth	-0.03	0.10	-0.23	0.16
Time				
(Log) Days	0.04	0.01	0.02	0.06
Individual Differences				
Word reading	-0.08	0.04	-0.16	0.01
Nonword reading	-0.05	0.03	-0.11	0.02
Rapid naming	-0.08	0.03	-0.14	-0.02
Phonological Skill	0.01	0.02	-0.03	0.05
Vocabulary	0.06	0.04	-0.01	0.13
Spelling	-0.03	0.04	-0.10	0.05
Age	0.00	0.00	0.00	0.00

Note:
CI = Credible intervals.

Figure 7.9

Estimates from the Posterior Distribution of the Preferred Model for ID and Psycholingustic Predictors on Letter Search Reaction Time Data



increase of ~13 ms every 95 days.

RON is the only reliable predictor of the ID measures in this model, as it is with accuracy, and is associated with a decrease in reaction times (log-odds = -0.08 [-0.14, -0.02]). An increase in 1 SD in RON skill will decrease reaction times by approximately 28 ms.

The remaining predictors all show coefficients with credible intervals that cross zero, and so the estimates are unreliable and the model is are inconclusive about their effects for this data and this sample.

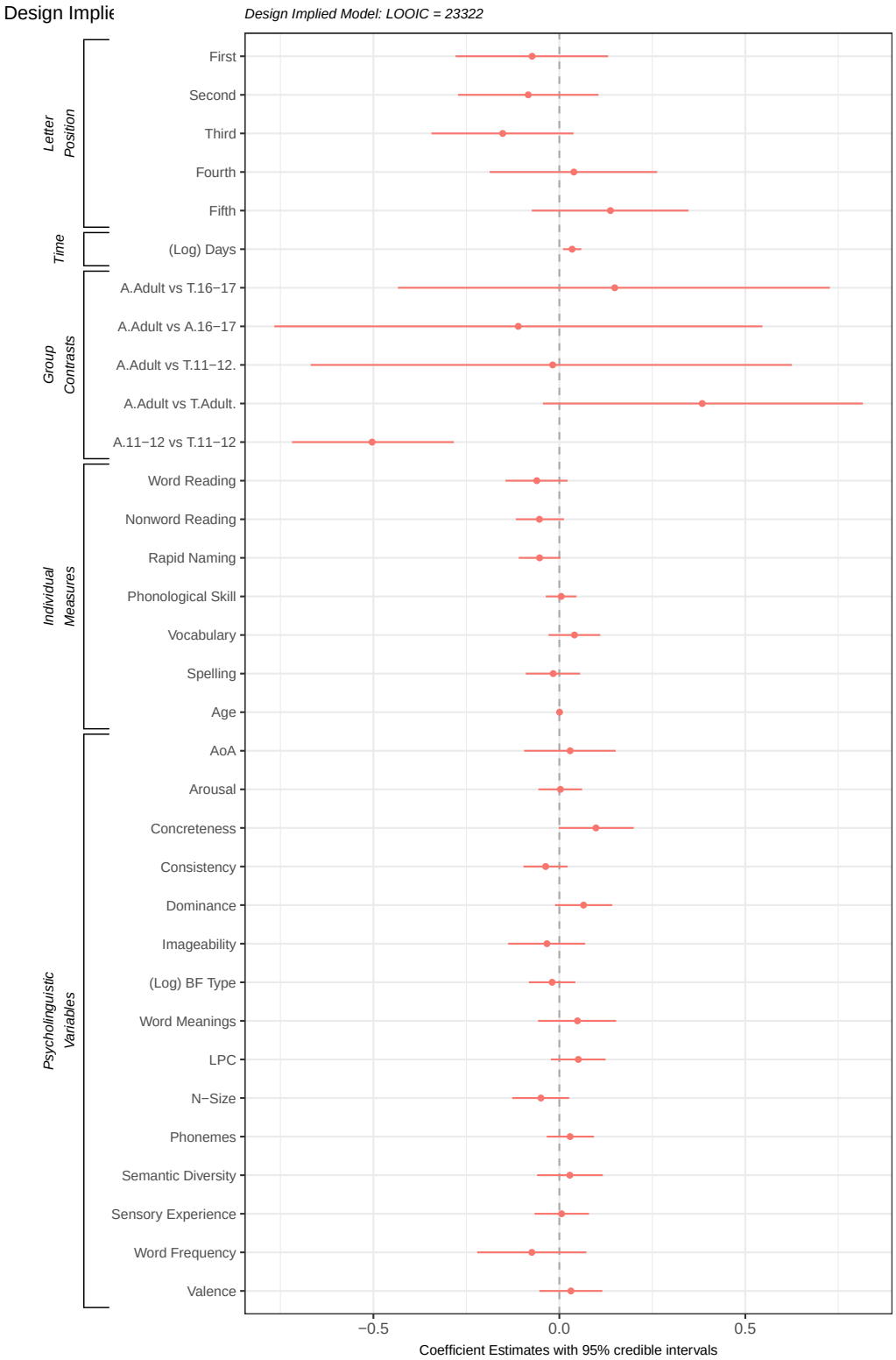
Complete Case and Outlier Analyses. The effect of slowed responses over data collection sessions is also a stable effect. The direction of effect on the first letter position becomes positive in the complete case and outlier analyses. Second to fifth letter position estimates remain stable. RON remains a stable effect but for the complete case analysis. Nonword reading has credible intervals that are just shy of zero. There is a change in the vocabulary coefficient in the complete case and complete case with no outliers model. Vocabulary is estimated with greater certainty such that the model implied coefficient suggests that readers of higher vocabulary knowledge have slower reaction time. The estimates for phonological skill, spelling and age remain inconclusive.

Letter Search Reaction Time Design Implied Model. The design implied model is displayed in Figure 7.10. The coefficient values and credible intervals for letter position, log (days) and ID measures are equivalent across preferred and design-implied models. The model clearly estimates that the atypically-reading 11-12-year-olds are slower than their typically-reading peers (Typ-11: mean = 1094.7, SD = 401.5; Atyp-11: mean 1110.8, SD = 363.9). The data are inconclusive as to the size of any contrast effect for the atypically-reading adult readers with the other reading groups.

The positive coefficient for concreteness suggests that words of higher concreteness ratings slow reaction time responses. None of the other psycholinguistic

Figure 7.10

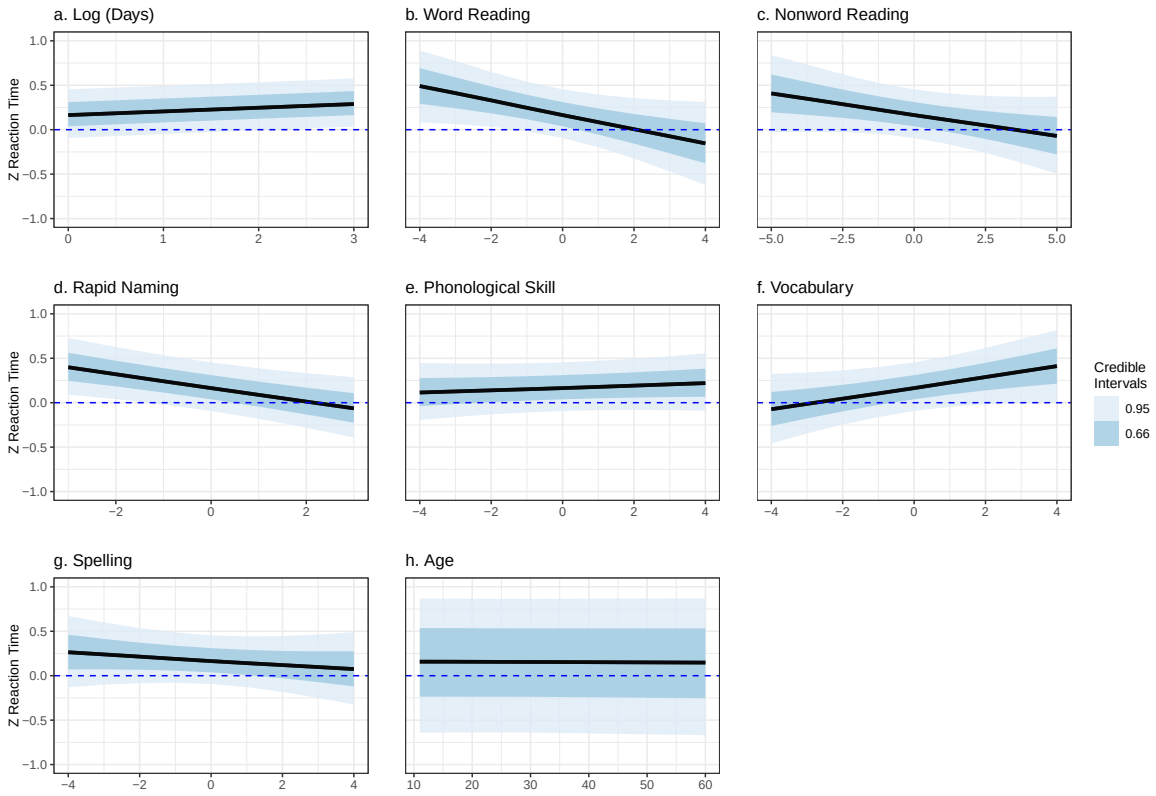
Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on Letter Search Reaction Time Data



variables contribute any reliable influence to the reaction time outcome.

Figure 7.11

Preferred Model Predictions for the Effects of Individual Differences and Group on Letter Search Reaction Time Performance



Model Predictions. We simulate model implied predictions for changes in standardised reaction time at varying levels of ID measures: representing a range of values for a predictor, maintaining the other variables at zero and updating the model. Predictions for the ID measures are displayed in Figure 7.11.

There are two slight differences in reaction time prediction plots from accuracy prediction plots. We intentionally draw the y-axis with the same limits (-1 - 1) on each reaction time prediction plot for each task. This allows a crude, visual comparison of the slope for one predictor with other predictors because the range of the y-axis of each plot is identical. The rate of change across predictors is not

identical, however, since the range of skills for each ID measure, plotted on the x-axis is not identical.

Secondly, we indicate the mean value of the *observed data* by the dashed blue line at '0' on the y-axis. Recall that this is the mean value for a typically-reading 16-17-year-old participant. The mean value of the *predictive distribution* is indicated by the solid black line in the plot. Subsequently, where the intervals and solid black line are above the dashed line, the model implied predictions for reaction times are slower than the observed mean for the level of skill. Where intervals and the solid black line are below the dashed line, the predictions are for faster reaction times than the observed mean. Where the solid black line intercepts the dashed line, the two distribution means correspond with each other.

The model implied predictions indicate that log (days), higher levels of phonological skill and vocabulary (plot a, e and f, respectively) will be associated with slower reaction times. Higher levels of word, nonword and RON and spelling skill will facilitate faster reaction times.

Most notable is that reaction times for out-of-sample predictions will tend to be slower than those within the sample. This is likely because across all letter positions our reference group of the typically-reading 16-17-year-olds were slightly slower than the atypically-reading 16-17-year olds and adults, however the 11-12-year-olds and the typically-reading adults were much slower than them, pulling the predictions away from the centre of the distribution. This may also explain why stronger vocabulary knowledge is predicted to slow responses, since the typically-reading adult readers had by far the highest vocabulary scores and were consistently slower than the reference group.

Summary and Discussion. For letter search reaction time data, the preferred model was a simple model of only letter position, time and ID measures. Although, once more, none of the estimates for letter position were conclusively positive or negative (and unstable in sensitivity analyses), the direction of effects indicated that first and third letter positions were fastest to be responded to, just as they were more

likely to be correctly responded to in the accuracy data. Reaction time slowed across data collection points and higher scores in RON skill were associated with faster reaction times.

What is surprising is the lack of influence from skills such as nonword reading. This seems to suggest that there is an ability to strategise for this task and choose to not decode the information, while the start of the item, as a location rather than a bias for letter identity, may give the initial letter position privileged status (Kessler et al., 2002).

RON alone predicted performance for both accuracy and reaction time measures. No other predictor was reliable. The level of analysis in this task is the letter level and that, plus the mixed presentation of the word items and unpronounceable letter strings, likely nullified the concept of word as salient to successful performance. Ziegler et al. (2008) explicitly included unpronounceable nonwords believing that identification would then occur without lexical access. The absence of any influence of psycholinguistic predictor influence would seem to support this. Consequently, predictors that focus at the word-level may also be reduced in salience relative to the letter search task specific demands: there is no pronunciation so the requirement for phonological category variables is reduced – e.g., phonological skill or number of phonemes. The diminished status of items as words may also be supported by the lack of a word superiority effect.

Slower reaction time for people with higher levels of phonological skill and vocabulary are indicated in the prediction plots. There is no overt pronunciation here, nor is there any overt word identification. The lexical quality hypothesis suggests that people of high skill are more likely to have integrated processes across the three components of orthography, phonology and semantics. Perhaps the strength of their skill means that mere presentation of the orthographic form of the word activates the information which inhibits responses for the letter.

7.2 Lexical Decision

Participants indicated whether a letter-string was either a word or a nonword in the lexical decision task. The items were real words and legal nonwords. Details of list construction and task procedure are reported in section 5.2.4.2.

Our research question was whether atypically-reading adults could discern words from nonwords to the same extent and at the same rate as other groups of readers. If the preferred model included the group contrast predictor, this may indicate that there are differences between the atypically-reading adult readers for accuracy rates or speed of making decisions. If the preferred model included interaction effects between group predictor and any of the ID or psycholinguistic measures, this could indicate a difference in either strategy or knowledge for completing trials. No interactions between the group variable and ID or psycholinguistic measures would suggest that groups are approaching the task similarly.

7.2.1 Item Properties

At each time point, a participant would see 120 items (50 words, 50 nonwords and 20 words for the isolation condition of the sentence reading task). Word items were balanced across lists for word-frequency (high vs low) and length (3-7 letters). Mean frequency scores for high and low ratings and also the mean length of items across lists 1- 3 are in Table 7.5. A two-way ANOVA for frequency (high and low) and list (1-3) on frequency ratings confirmed a significant main effect of frequency ($F(1, 204) = 859.9$, $p < .001$) and a non-significant main effect of list ($F(2, 204) = 0.2$, $p = .782$). A one-way ANOVA for the effect of list on length confirmed a non-significant finding ($F(1, 207) = 0$, $p = .944$). Thus, our design to incorporate low and high frequency words is supported and we can infer no differences for frequency ratings or length of words across the lists.

The properties of other psycholinguistic variables and findings of inferential

Table 7.5

Descriptive Statistics for Frequency and Length for Three Item Lists in the Lexical Decision Task

	Mean Frequency (SD)		
List	High	Low	Length
1	5.2 (0.7)	2.9 (0.4)	4.6 (1.2)
2	5.3 (0.6)	3 (0.5)	4.6 (1.3)
3	5.2 (0.7)	2.9 (0.4)	4.7 (1.3)

tests for differences across the lists are in Table 7.6 and displayed in Figure 7.12. None of the variables differed significantly between lists (all *p*s $> .08$).

7.2.2 Analyses

In lexical decision, a correct response represents correctly identifying that a word is a word and that a nonword is a nonword. An incorrect response represents deciding that a word is a nonword and that a nonword is a word.

Additional to the accuracy and reaction time models, we compared accuracy rates between words and nonwords. We expected that words would have a higher rate of accuracy than nonwords.

7.2.2.1 *Number of Observations*

Full Sample. We collected 69,840 observations across words and nonwords in the lexical decision task. We excluded 1,440 observations from 12 participants as being duplicate items from a previous wave of data collection. We further excluded 28 observations that were < 200 ms, leaving 68,372 observations. After removing the nonword observations ($n = 28{,}488$), 39,884 correct and incorrect word trials were available for accuracy analyses. We further removed incorrect trials on words ($n = 6{,}147$) to leave 33,737 observation of correct trials for words for reaction time analyses.

Table 7.6

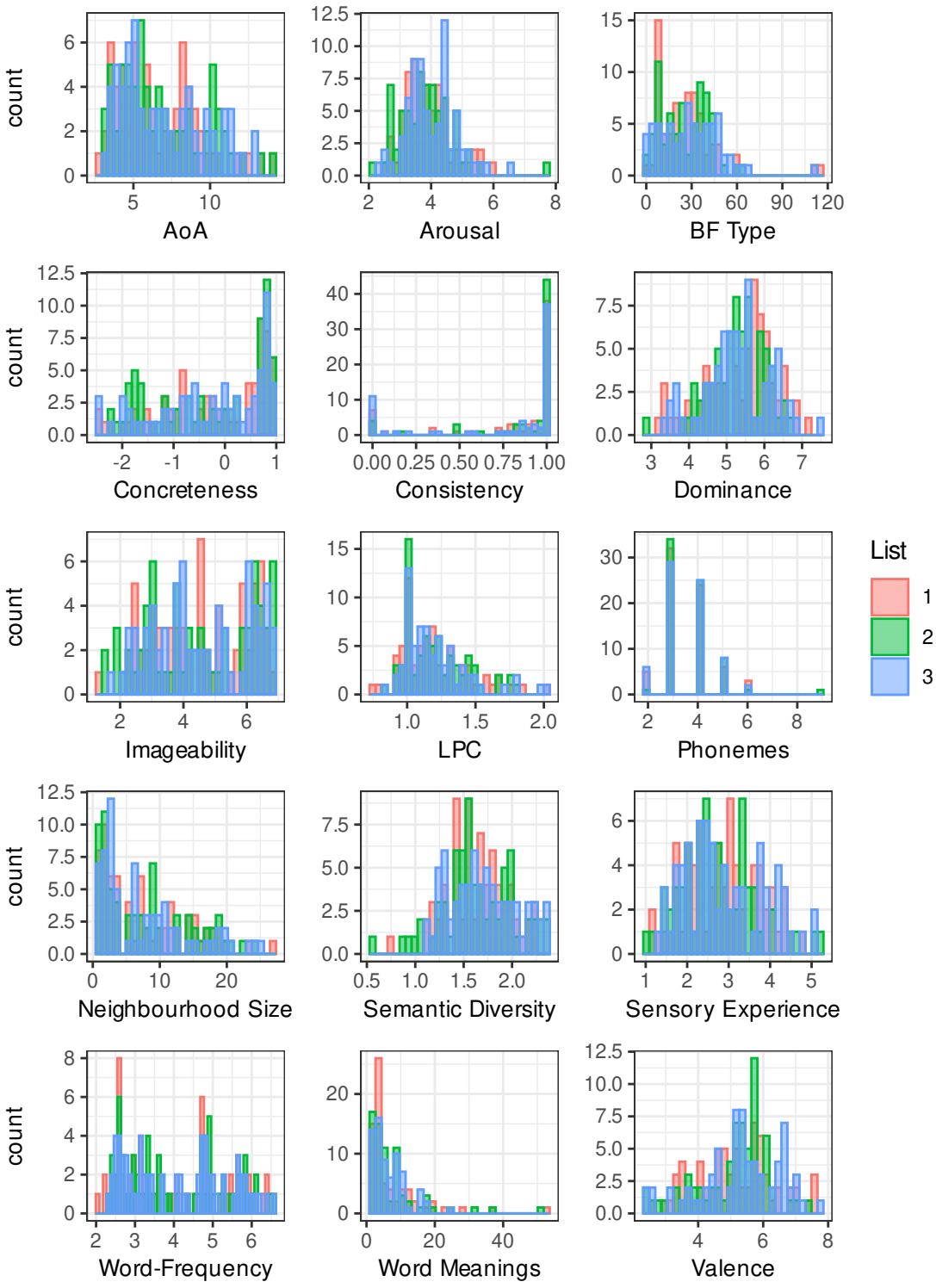
Summary of Psycholinguistic Variable Measures for Lexical Decision Word Items with F-Ratio and P Values to Signify Differences Between Item Lists

					ANOVA	
Psycholinguistic Variables	Mean	SD	Min	Max	F(2, 207)	p
AoA	6.8	2.7	2.9	14.2	0.37	0.69
Arousal	3.9	0.8	2.1	7.7	0.99	0.374
BF Type	27.2	18.7	1.0	116.2	0.96	0.384
Concreteness	3.8	1.0	1.7	5.0	0.19	0.824
Consistency	0.8	0.3	0.0	1.0	2.06	0.13
Dominance	5.3	0.9	2.8	7.4	0.11	0.897
Imageability	4.5	1.5	1.4	6.9	0.01	0.986
LPC	1.2	0.2	0.8	2.0	0.01	0.991
Phonemes	3.6	0.9	2.0	9.0	0.39	0.675
Neighbourhood size	7.9	6.5	1.0	27.0	0.03	0.972
Semantic diversity	1.7	0.3	0.6	2.4	0.77	0.464
Sensory experience	2.8	0.9	1.0	5.2	0.61	0.547
Word frequency	4.1	1.3	2.0	6.6	0.05	0.953
Word meanings	7.2	7.6	1.0	52.0	0.46	0.629
Valence	5.2	1.2	2.3	7.7	0.77	0.462

Note:
AoA = Age of acquisition. BF = Bigram frequency. LPC = Levenshtein Phonological Consistency.

Figure 7.12

Histograms Showing the Distribution of Psycholinguistic Properties of Items for the Lexical Decision Task, Across Three Lists



Complete Case Analysis. Just as with letter search data, 161 participants had three data collection sessions of lexical decision trials. This left 56,731 observations available for a complete case analysis of accuracy data. The analysis was repeated using the preferred model for accuracy (n = 33,018) and reaction time data (n = 28,568).

Outlier Analysis. After removing timed-out observations (n = 1,697, 2.5%), inter-quartile ranges per participant were calculated and outliers identified (see section 5.2.5.6) and removed (n = 5,866, 8.6%), leaving 60,809 word and nonword observations. The analysis was repeated using the preferred model for accuracy (n = 36,180) and reaction time data (n = 31,669). In the complete case analyses with no outliers, the number of observations for accuracy analyses was 30,576; for reaction time analyses n = 26,806.

7.2.3 Accuracy Results

7.2.3.1 *Descriptive Statistics*

We calculated average participant performance for words and nonwords by time. We display the distributions for accuracy and reaction time at the group level in Figures 7.13 and 7.19; averages across accuracy and reaction time by time and group are displayed in Figure 7.14 and 7.15 for words and nonword respectively. For accuracy, as with the letter search task, it appears as if more participants in each group are less accurate at the third data collection session. The spread of the lower accuracy is more pronounced in the nonword graphs.

Figure 7.13

Histograms Showing the Distribution of Mean Accuracy Rates per Participant, by Group and Time for Words and Nonwords in the Lexical Decision Task

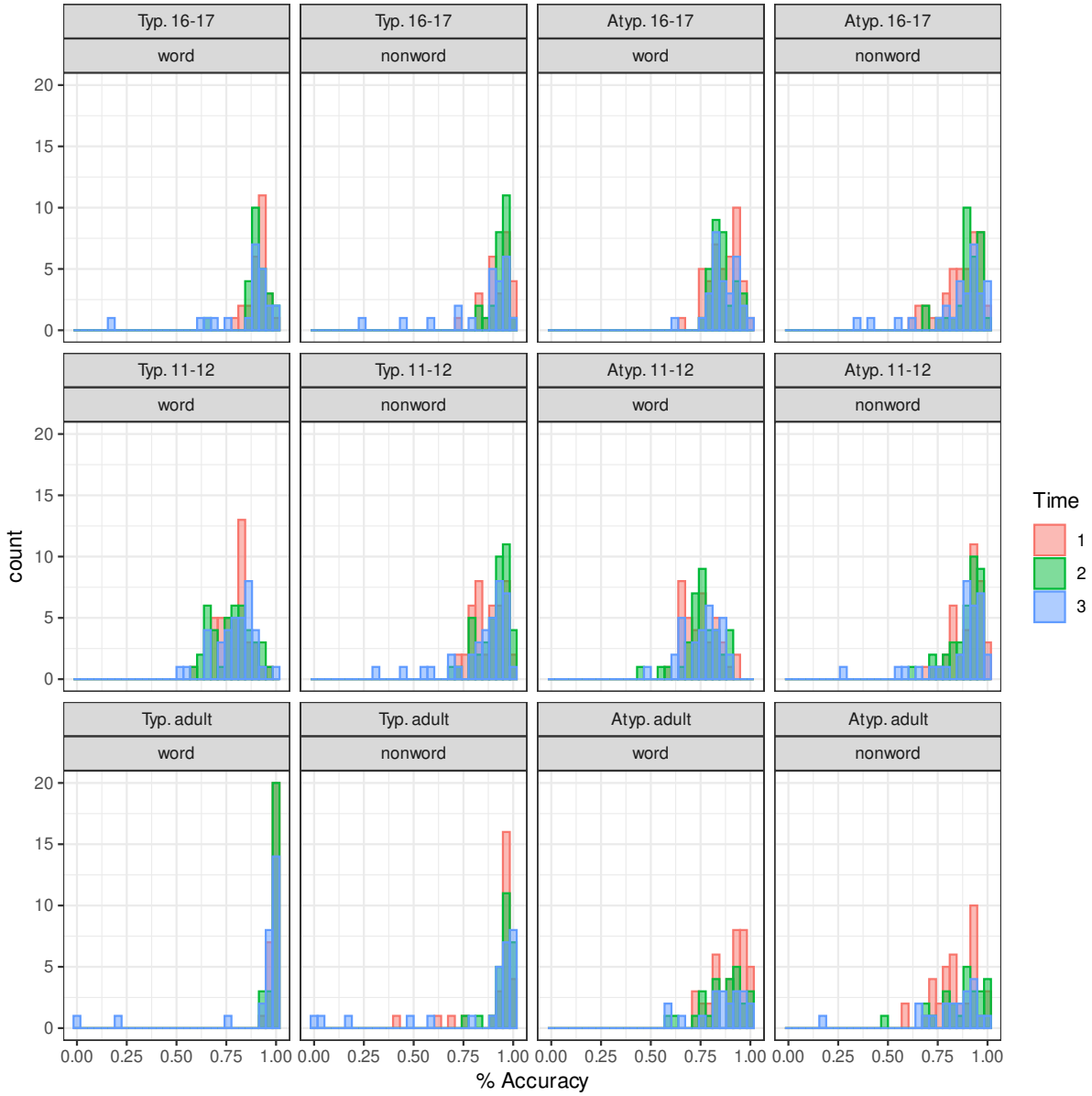


Figure 7.14

Accuracy and Raw Mean RT for Words By Group and Time for Lexical Decision

	Time 1			Time 2			Time 3		
Group	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typ. 16-17	91.2	889.4	415.8	90.3	933.2	502.4	84.5	1181.1	964.9
Atyp. 16-17	86.7	869.5	434.8	85.8	972.6	568.3	86.3	990.3	612.3
Typ. 11-12	79.6	1069.9	482.6	77.8	1136.8	570.8	79.6	1165.8	714.3
Atyp. 11-12	75.5	1127.0	532.2	75.7	1126.0	577.7	76.3	1151.7	630.1
Typ. adult	98.5	750.9	269.0	98.2	820.9	315.2	90.4	1056.0	924.5
Atyp. adult	89.4	869.9	416.9	85.9	910.5	571.9	85.6	981.8	671.4

Note:
Acc % = Percentage Accuracy; RT = Reaction time; Atyp. Adult = Atypically-reading adult; Typ. 16-17 = Typically-reading 16-17-year-old; Atyp. 16-17 = Atypically-reading 16-17-year-old; Typ. 11-12 = Typically-reading 11-12-year-old; Typ. Adult = Typically-reading adult; Atyp. 11-12 = Atypically-reading 11-12-year-old.

Figure 7.15

Accuracy and Raw Mean RT for Nonwords By Group and Time for Lexical Decision

	Time 1			Time 2			Time 3		
Group	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typ. 16-17	92.2	1048.7	444.1	94.0	1013.9	452.6	84.3	1278.5	977.5
Atyp. 16-17	86.4	1042.0	492.8	89.6	1031.8	519.1	84.9	1197.8	788.2
Typ. 11-12	87.5	1201.9	518.1	90.5	1137.4	502.8	84.3	1314.8	879.3
Atyp. 11-12	89.9	1186.8	510.7	88.8	1175.1	609.2	86.0	1223.2	736.6
Typ. adult	91.6	1040.0	407.1	95.2	1036.5	464.8	83.1	1382.2	1086.0
Atyp. adult	84.2	1123.9	484.3	86.9	1095.1	569.8	81.2	1282.9	897.9

Note:
Acc % = Percentage Accuracy; RT = Reaction time; Atyp. Adult = Atypically-reading adult; Typ. 16-17 = Typically-reading 16-17-year-old; Atyp. 16-17 = Atypically-reading 16-17-year-old; Typ. 11-12 = Typically-reading 11-12-year-old; Typ. Adult = Typically-reading adult; Atyp. 11-12 = Atypically-reading 11-12-year-old.

Difference Between Words and Nonwords. We estimated a model to test for a difference between word and nonword items. The model included predictors for log(Days), ID measures, random intercepts and ID measures on random slopes for

participants and words. We did not include psycholinguistic variables since nonwords do not have values for these predictors.

The model estimated an average probability of a word being identified correctly as 96.3%. If the item was a nonword, the odds were lower (log-odds = -1.11 [-1.39, -0.82]). This suggests that nonwords have 6.6% lower probability of being identified correctly in this model and this sample. This is a medium sized effect and supports our prediction that words would be identified with greater accuracy than nonwords.

Worth noting is that the log(Days) showed an attenuated rate of decrease in this model compared to the model on words only (log-odds = -0.08 [-0.13, -0.04]). Nonword reading and vocabulary were estimated unreliably but spelling was reliable (log-odds = 0.18 [0.06, 0.30]). When word and nonword items were modelled together, a 1 SD increase in spelling knowledge (approximately 4.5 words) increased the odds of an accurate response by approximately 0.6%.

7.2.3.2 *Preferred Model*

The preferred model for the lexical decision data was the Additive-RIS model, i.e., the model containing ID and psycholinguistic predictors as independent terms on fixed effects, random intercepts and slopes. The model *without* the group predictor was preferred over the model with the group predictor. This model explained R^2_{bayes} = 40% [39.2, 41.0] of the variance in the accuracy outcome. Figure 7.16 and Table 7.7 display the fixed effects' coefficients for the model.

Model Inference. On average, the probability of making a correct response is 96.3%. The effect of log (days) is reliable and negative (log-odds = -0.15 [-0.21, -0.09]) indicating that with a 1 SD increase in days (~ 95), the probability of making a correct response decreases by about 0.5%. This is a very small effect.

Each of the mean values of estimates for ID measures lie to the right of zero, suggesting that they all show a trend for increasing the odds of an accurate response.

Table 7.7

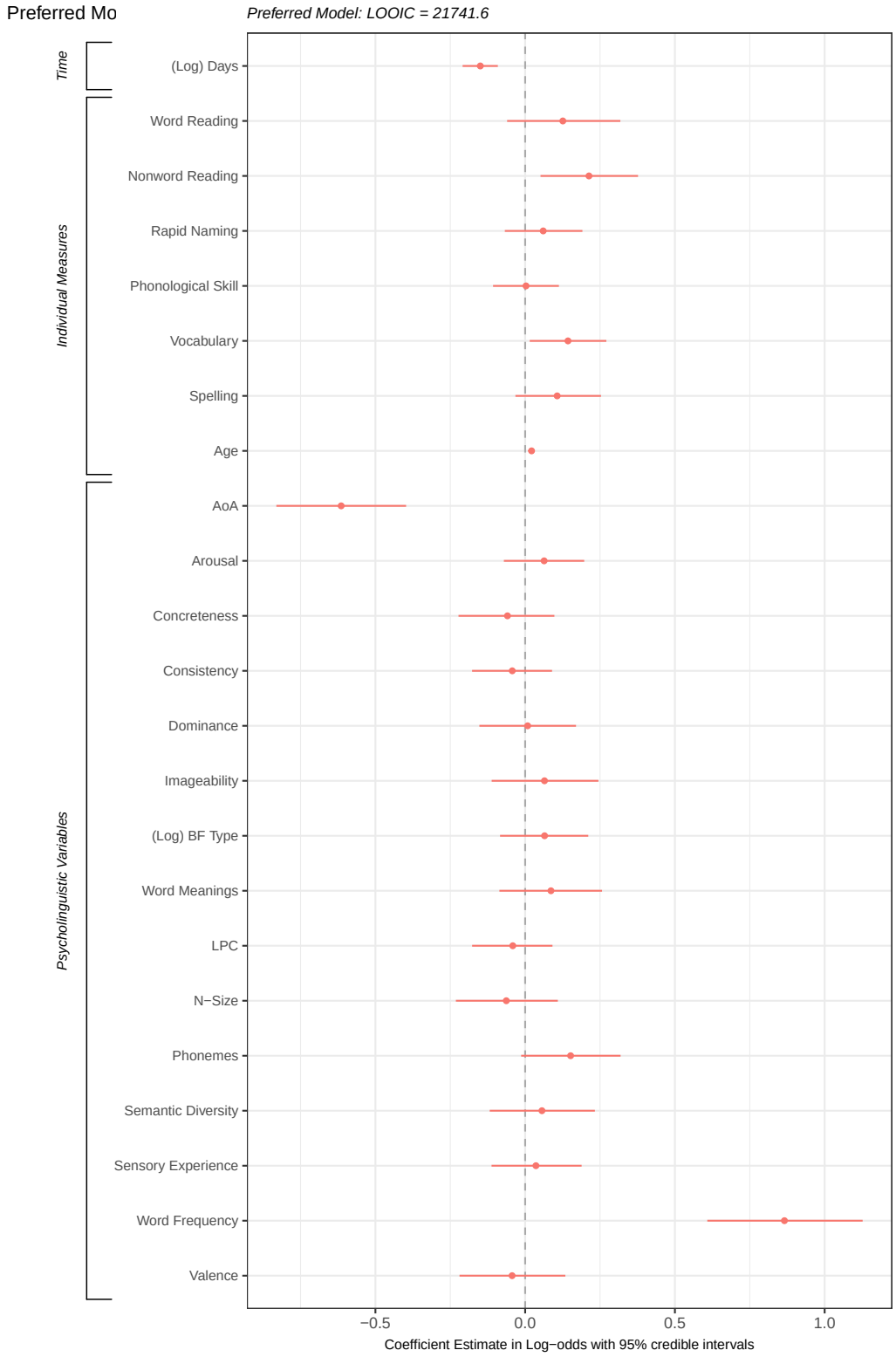
Summary of Standardised Fixed Effects for Lexical Decision Accuracy

Term	Estimate	SE	Lower CI	Upper CI
Intercept	3.26	0.13	3.01	3.51
Time				
(Log) Days	-0.15	0.03	-0.21	-0.09
Individual Differences				
Word reading	0.13	0.10	-0.06	0.32
Nonword reading	0.21	0.08	0.05	0.38
Rapid naming	0.06	0.07	-0.07	0.19
Phonological skill	0.00	0.06	-0.11	0.11
Vocabulary	0.14	0.07	0.02	0.27
Spelling	0.11	0.07	-0.03	0.25
Age	0.02	0.00	0.01	0.03
Psycholinguistic Variables				
AoA	-0.61	0.11	-0.83	-0.40
Arousal	0.06	0.07	-0.07	0.20
Concreteness	-0.06	0.08	-0.22	0.10
Consistency	-0.04	0.07	-0.18	0.09
Dominance	0.01	0.08	-0.15	0.17
Imageability	0.06	0.09	-0.11	0.24
(Log) BF Type	0.06	0.08	-0.08	0.21
Word meanings	0.09	0.09	-0.09	0.26
LPC	-0.04	0.07	-0.18	0.09
Neighbourhood size	-0.06	0.09	-0.23	0.11
Phonemes	0.15	0.09	-0.01	0.32
Semantic diversity	0.06	0.09	-0.12	0.23
Sensory experience	0.04	0.08	-0.11	0.19
Word-Frequency	0.87	0.13	0.61	1.13
Valence	-0.04	0.09	-0.22	0.13

Note:
CI = Credible intervals. AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency.

Figure 7.16

Estimates from the Posterior Distribution of the Preferred Model for ID and Psycholingustic Predictors on Lexical Decision Accuracy Data



Only nonword reading skill, vocabulary and age are estimated with such certainty however, that we can assert that the sign of the effect is positive. Word reading skill, rapid naming skill, and spelling's lower credible intervals include zero. Phonological skill is uninformative as to outcome in this model as it sits squarely on zero.

Reliable or not, these are all very small effects. The coefficient for nonword reading (log-odds = 0.21 [0.05, 0.38]) indicates that with 1 SD increase in nonword reading skill, there is an increase in the probability of responding accurately of 0.7%. The SD of nonword reading for this sample was 7.7, so in raw units of nonword skill and probabilities of making a correct response, this means a person who answered approximately 50 of the 63 nonwords had a 96.3% probability of making a correct response. This increased to 97% for a person who could answer approximately 57 words.

With an increase of 1 SD in vocabulary scores, the log-odds of making a correct response increases by 0.14 [0.02, 0.27]. This is an increase from 96.3% to 96.8% - half a percentage point for knowing six more vocabulary items on the Shipley vocabulary test.

Age is estimated as a very small, positive and reliable effect. Older participants show higher odds of being accurate than younger people (log-odds = 0.02 [0.01, 0.03]). For a 1 SD change in the age variable which equates to 38.6 years, accuracy increases from 96.31% to 96.38% - a difference of 0.07%.

Of the 15 psycholinguistic variables, only AoA and word-frequency show reliable estimates. A range of positive and negative relationships with accuracy are possible according to the model implied coefficients for the remaining 13 predictors, however none are reliable for this model and this sample.

AoA has a small, negative association with lexical decision accuracy (log-odds = -0.61 [-0.83, -0.40]. Later learned words show lower odds of being identified as words than early learned words. As AoA increases by 1 SD, the probability of making a correct decision for a word decreases from 96.3% to 93.3%, approximately 3%. The SD for AoA on the raw scale is 2.7 years. Consequently, the model estimates that words that are learned approximately two and half years apart have significantly

different probabilities of being identified as a word.

Word-frequency shows a small positive relationship (log-odds = 0.87 [0.61, 1.13]) with lexical decision accuracy. Words that carry higher frequency ratings show higher odds of being answered correctly than words with lower frequency ratings. The SD value for the frequency scale is 1.3 in this dataset. A move of 1.3 categories up the Zipf-scale will increase accuracy from 96.3% to 98.4%.

Since this is an exploratory study, we think it worth noting the estimate for phonemes. Recall that the length predictor was identified as having a high VIF rate and so we dropped length from the modelling process, retaining the phonemes predictor as a proxy measure (Morrison et al., 2003). The estimate here is log-odds = 0.15 [-0.01, 0.32]. One SD in the phonemes measure is equivalent to an increase of one sound to a word. While unreliable and very small, the estimate suggests that as the number of phonemes increases by 1, they have a higher odds of being identified correctly as words. The advantage is roughly 0.5%.

Complete Case and Outlier Analyses. The effects of nonword reading and vocabulary remain stable in the complete case analysis. Plus word reading skill is reliable and positive. Participants with higher word reading skill have higher odds of making a correct response. The complete cases analysis also shows that the tendency to be less correct over time is diminished. All other ID measures remain inconclusive. AoA and frequency maintain their status as well defined, certain effects for accuracy on lexical decision trials.

Since AoA and frequency have sometimes been discussed as variant of a frequency type predictor, we created two additional models to scrutinise whether a model with one of them omitted performed better than the model that included both. Neither model performed better from a reading of the subsequent LOOIC values ($LOOIC_{both}$: 21741.6; $LOOIC_{AoA}$ = 21808.3; $LOOIC_{freq}$ = 21746.9).

In the model with AoA removed, the frequency coefficient increases from a log-odds of 0.87 [0.61, 1.13] to 1.22 [0.97, 1.47]. The change is not significant between the two absolute effect sizes, however in terms of interpretation, this means the

frequency is now a moderate effect. The direction of effect remains the same. Without AoA in the model, a 1 SD increase in frequency increase accuracy by approximately 2.5%. The difference in accuracy effects when AoA is included and not included is approximately 0.5% of a probability point.

In the model with frequency removed, the AoA coefficient increases from -0.61 [-0.83,-0.40] to -0.94 [-1.15, -0.75]. The effect size is not significantly different between the two models, however the direction of the effect remains unchanged with the absolute magnitude being larger. This represents an overall decrease in accuracy by approximately 5% for every 1 SD increase in AoA - a difference of 2% when frequency is estimated in the same model. It is clear from these further models that AoA can recover some of the frequency variance, in its absence, where frequency cannot do the same for AoA.

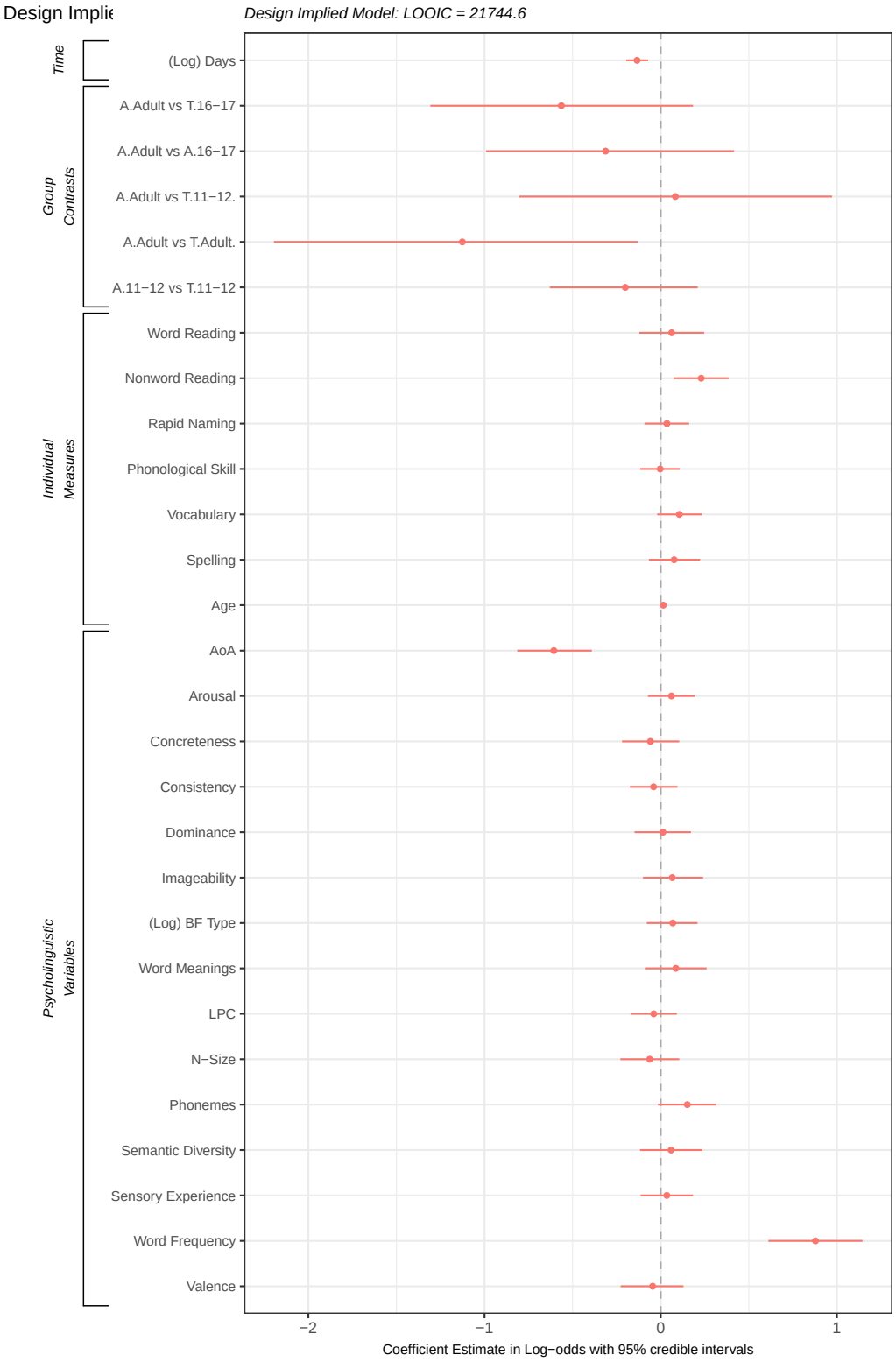
Design Implied Model. The design implied model is in Figure 7.17. The difference between models is only the inclusion of the group contrast predictor. The model implied relationship between atypical and typically-reading adults is reliably negative and of a medium size. Atypically-reading adults have lower odds than typically-reading adults to correctly identify that a letter string is a word. The credible intervals are wide, indicating a wide range of plausible values. This is the only contrast amongst the range of group contrasts where the model is confident that the relationship is negative, with both positive and negative relationships plausible for the remaining contrast effects.

There are changes in the estimates for ID measures. The effects of vocabulary and age are less certain here; their credible intervals cross zero, meaning no difference is a plausible relationship. Nonword reading, AoA and word-frequency remain strong and each follow the same direction of effect as in the preferred model.

Model Predictions. The model implied predictions are shown in Figure 7.18. As before, these plots are drawn on the probability scale with the y-axis ranging from 0 - 1. We present the ID measure predictions (plots a - h) and the model implied

Figure 7.17

Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on the Lexical Decision Accuracy Data



predictions for AoA (plot i) and word-frequency (plot j). Credible intervals denoting 66% and 95% certainty for the predictions surround the curve, here indicated by the solid black line.

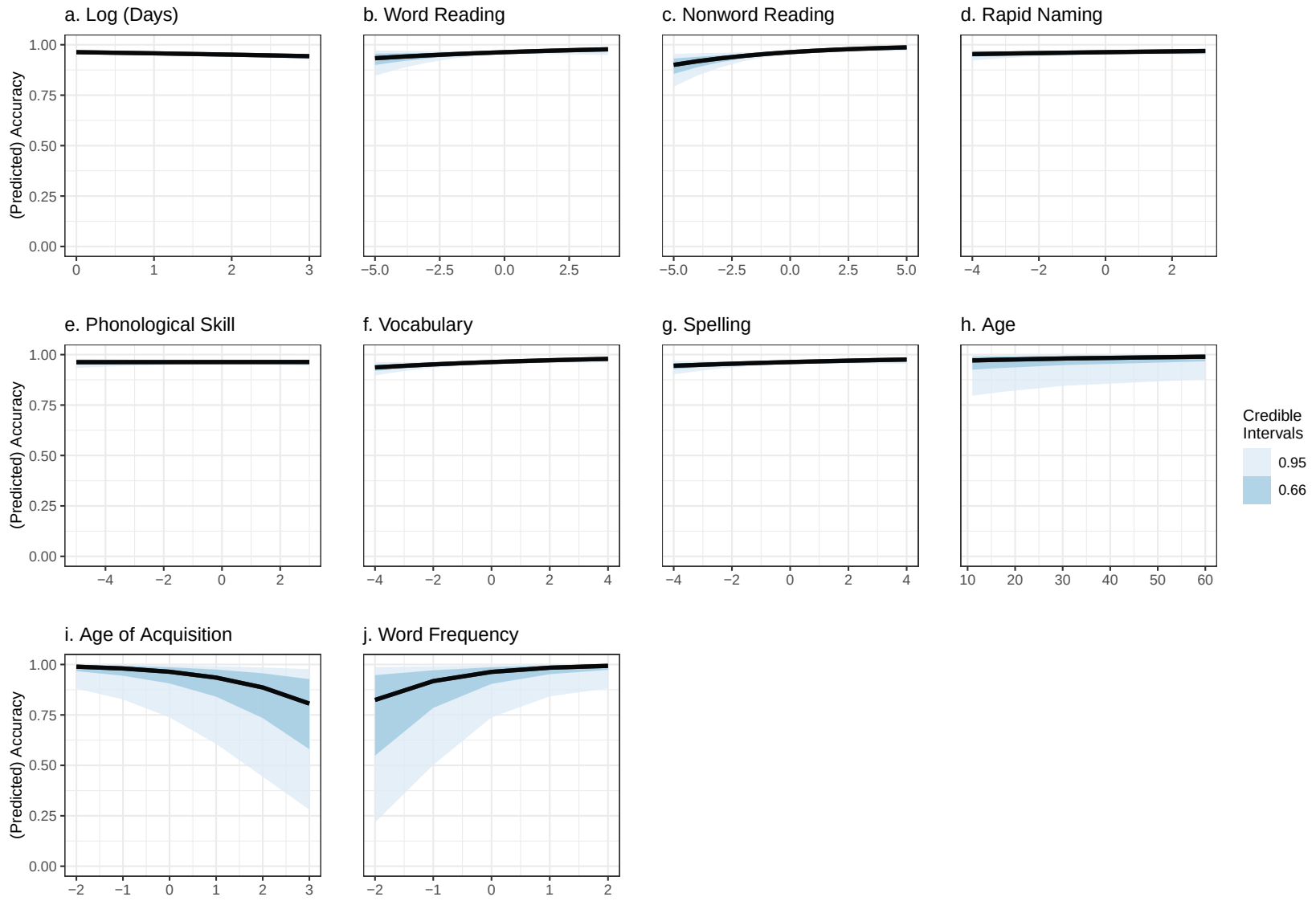
As with the letter search task, at the extreme values of any of the predictors, lexical decision accuracy is high. The predictions for reading skill (plot b) and nonword reading skill (plot c) suggest that even at a very low level of skill, accuracy is not predicted to fall below ~ 80%. The lowest age prediction (10 years) suggests that the lower credible interval falls just above 75% predicted accuracy. The predictions for AoA and word-frequency appear to be mirror images of each other. The model implied probability for correctly identifying a word as a word for the latest learned words and the lowest frequency words is approximately 80%.

Summary and Discussion. Words have a higher probability of being correctly classified than nonwords in this data sample. Over time, responses generally decrease in accuracy, echoing the letter search effect of time. The preferred model for lexical decision accuracy does not include the group predictor. Nor does it include any interaction terms, which suggests that, given this range of models and this data, irrespective of age or literacy status, participants approach lexical decision in similar ways.

Having higher levels of nonword reading skill and vocabulary knowledge, and being older participants means your odds of correctly identifying a word as a word are higher. People of lower nonword reading and vocabulary skill, and younger are less likely to be accurate. It is emphasised that these are *very* small sized effects.

Figure 7.18

Preferred Model Predictions for the Effects of Individual Differences, Age of Acquisition and Frequency on Lexical Decision Accuracy Performance



Nonword reading rather than word reading is the reliable predictor for this model and sample. In the complete case analysis, word reading becomes reliable, alongside nonword reading and vocabulary. We observed that rates of attrition for the atypically-reading adults and 16-17-year-old groups were the highest among the sample (Figure 6.1). We suggest that as the number of participants of atypically-reading status left the study, word reading could register as an influential predictor as the balance between groups remaining in the study had stronger word reading skills. For nonword reading to remain reliable in the complete cases analysis, however, it must be relevant to all readers to some extent. This may be the related to the use of pronounceable nonwords mixed with the word items, promoting decoding skills for all participants for the unfamiliar items.

The presence of vocabulary as a reliable estimate is unsurprising as discerning words from nonwords in a lexical decision is often believed to recruit the use of semantics as a decision principle for words. Perhaps the strength of the nonword estimate - for both the full sample and the complete cases analysis - is a reflection of the low vocabulary we observe for the atypical groups. This interpretation invokes the division of labour hypothesis (Plaut et al., 1996) where phonological and semantic information work together to facilitate word recognition.

We also noted the almost reliable predictor of phonemes. The phoneme predictor is acting as a proxy for an absent length predictor, so it isn't clear if this is a pseudo length effect or if the number of phonemes is the actual effect. If it were, this would suggest that phonological information is being used to make decisions in the lexical decision task, where no overt pronunciation is required. Phonological recoding in the absence of a phonological output supports the strong phonological theory for this sample (Frost, 1998).

The model estimates for AoA and word-frequency show negative and positive relationships, respectively. Both effects are in line with previous studies. However, once more, because the baseline accuracy rate in the lexical decision task for words is high (96.3%), these effects are small, as accuracy approaches 100%, effects become compressed.

7.2.4 Reaction Time Results

7.2.4.1 *Descriptive Statistics*

Distributions for the mean reaction time in milliseconds are displayed at the group level in Figure 7.19, 7.14 and 7.15 for correct responses to words and nonwords. Atypical and typically-reading adult and 16-17-year-old readers appear to be faster for word responses than nonword responses, that trend is not so clear for atypically-reading 16-17-year-old readers or the younger groups.

7.2.4.2 *Preferred Model*

The preferred model for reaction time data in the lexical decision is the Additive-RIS model without the group contrast predictor, as with accuracy data. The model contains predictors for time, ID measures and psycholinguistic variables, with predictors on random intercepts and slopes for participants and items. The explained variance in the reaction time data for the model is $R^2_{bayes} = 35.5\%$ [34.9, 36.2]. The fixed effects of the model are presented in Figure 7.20 and Table 7.8.

Model Inference. Log (days) is positively and reliably associated with reaction time ($\beta = 0.07$ [0.06, 0.09]). With a 1 SD increase in days, reaction time slows by approximately 23 ms. In the first data collection session, the model implied average response time is approximately 825 ms with this increasing to approximately 847 ms 95 days later.

The general trend for the group of ID measures is for them to quicken responses, all but phonological skill being negatively associated with reaction time. Only nonword reading skill and rapid naming reliably estimated, however. The remaining variables show that both positive and negative effects are compatible with the data.

Higher nonword reading scores are associated with faster reaction times ($\beta =$ -0.07 [-0.12, -0.02]). A reader that is 1 SD stronger in nonword reading is

Figure 7.19

Histograms Showing the Distribution of Raw, Mean Reaction Time (ms) By Participant, Group and Time Point for Words and Nonwords in the Lexical Decision Task

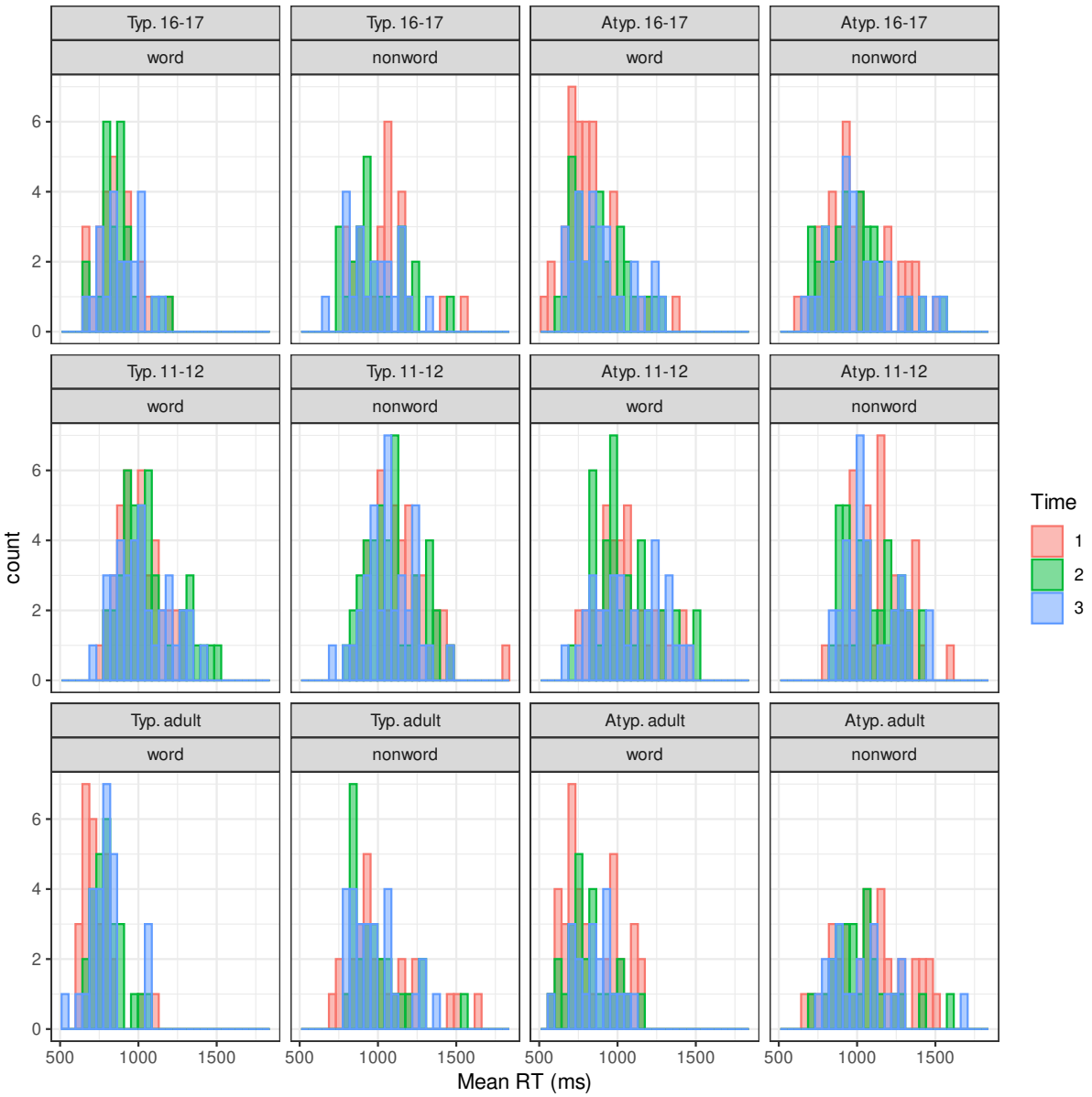


Figure 7.20

Estimates from the Posterior Distribution of the Preferred Model for ID and Psycholingustic Predictors on the Lexical Decision Reaction Time Data

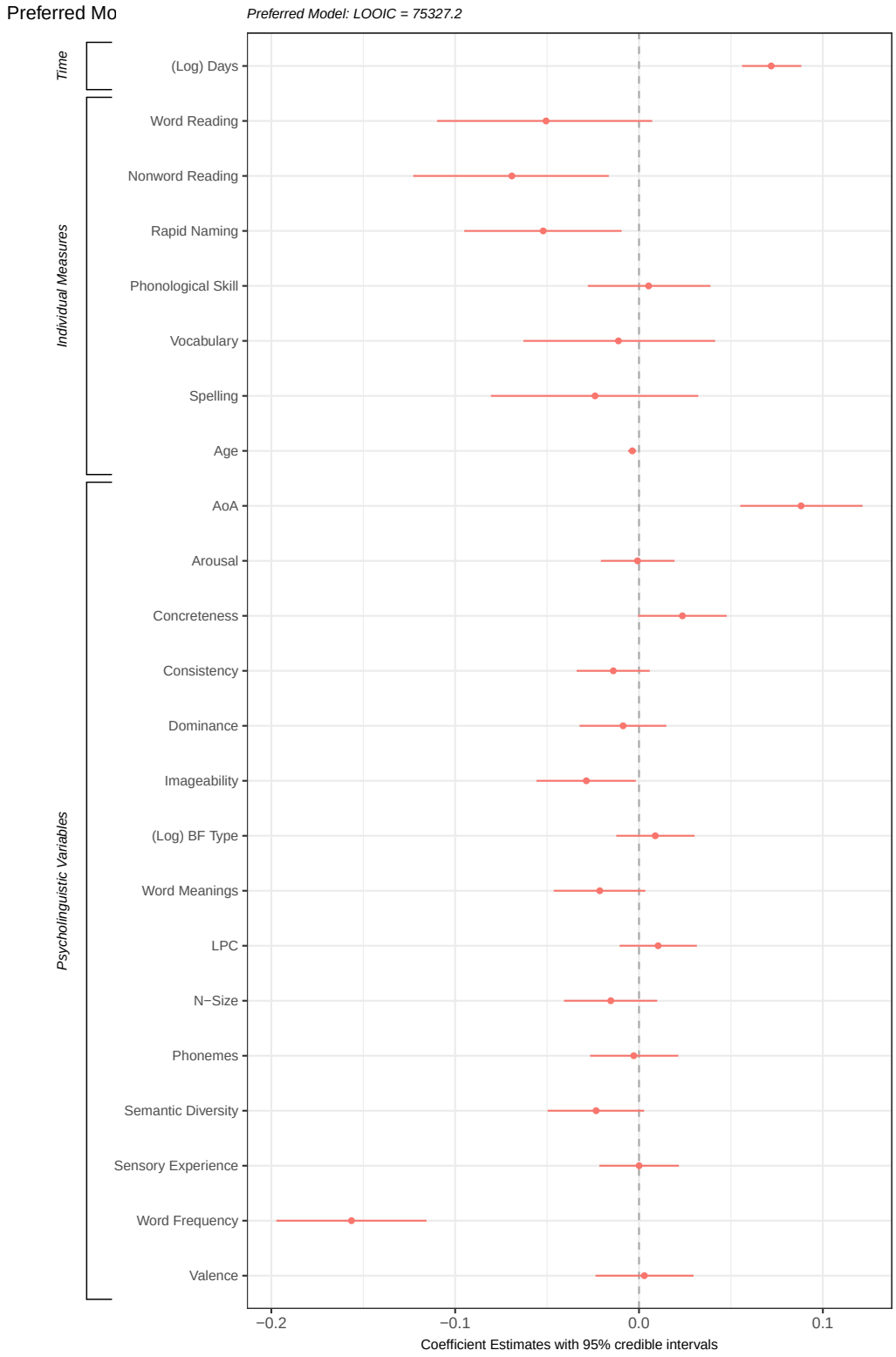


Table 7.8

Summary of Standardised Fixed Effects for Lexical Decision Reaction Time

Term	Estimate	SE	Lower CI	Upper CI
Intercept	-0.02	0.04	-0.10	0.06
Time				
(Log) Days	0.07	0.01	0.06	0.09
Individual Differences				
Word reading	-0.05	0.03	-0.11	0.01
Nonword reading	-0.07	0.03	-0.12	-0.02
Rapid naming	-0.05	0.02	-0.10	-0.01
Phonological skill	0.01	0.02	-0.03	0.04
Vocabulary	-0.01	0.03	-0.06	0.04
Spelling	-0.02	0.03	-0.08	0.03
Age	0.00	0.00	-0.01	0.00
Psycholinguistic Variables				
AoA	0.09	0.02	0.06	0.12
Arousal	0.00	0.01	-0.02	0.02
Concreteness	0.02	0.01	0.00	0.05
Consistency	-0.01	0.01	-0.03	0.01
Dominance	-0.01	0.01	-0.03	0.01
Imageability	-0.03	0.01	-0.06	0.00
(Log) BF Type	0.01	0.01	-0.01	0.03
Word meanings	-0.02	0.01	-0.05	0.00
LPC	0.01	0.01	-0.01	0.03
Neighbourhood size	-0.02	0.01	-0.04	0.01
Phonemes	0.00	0.01	-0.03	0.02
Semantic diversity	-0.02	0.01	-0.05	0.00
Sensory experience	0.00	0.01	-0.02	0.02
word-frequency	-0.16	0.02	-0.20	-0.12
Valence	0.00	0.01	-0.02	0.03

Note:
CI = Credible intervals. AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency.

approximately 23 ms faster at responding correctly. Higher scores for RON (β = -0.05 [-0.10, -0.01]) also decreases reaction time by approximately 17 ms. Just as with the accuracy outcome, the individual difference effects are incredibly small.

AoA and word-frequency are also identified as influential predictors for reaction time, as they were in the accuracy outcome model for lexical decision. AoA shows a positive, reliable association with reaction time (β = 0.09 [0.06, 0.12]) , describing how later learned words are slower (by approximately 30 ms) to be identified correctly as words than earlier learned words. We know that the AoA SD is approximately 2.5 years, so when words are separated in ratings by this amount, this model predicts that the later learned word will be slower by about 30 ms.

Word-frequency is reliably, negatively associated with reaction time (β = -0.16 [-0.20, -0.12]): words of higher frequency are associated with shorter reaction times. Each 1 SD increase in frequency (SD = 1.3) results in a decrease in reaction time of about 53 ms.

An additional influential predictor in the reaction time model is concreteness. It is showing a very small, positive relationship, with its lower credible interval resting on zero (β = 0.02 [0.00, 0.05]). It suggests that for every 1 SD increase in concreteness ratings (SD = 1.02), reaction times slow by approximately 7 ms. The concreteness measure within this study uses a Likert scale from 1 - 5, suggesting that as the ratings move up a single scale value, associated reaction times are 7 ms slower than the average rated word for concreteness.

Turning to predictors worthy of note given the exploratory nature of the study. Imageability (β = -0.03 [-0.06, 0.00]), number of word meanings (β = -0.02 [-0.05, 0.00]) and semantic diversity (β = -0.02 [-0.05, 0.00]) all appear to be potential candidates for predictors that show some influence. You will notice that their credible intervals contain positive and negative values, with their upper credible interval limits resting on zero. The model has estimated a posterior distribution that includes zero, however the limit of the posterior is so close to zero that the digits are not printed here. These estimates and credible intervals could change with a change in sample or study design and so are worthy of consideration here for replication purposes.

Each has a negative relationship with reaction time. In this model, words whose referents are easier to call to mind are decided upon quicker than words that do not easily call up a mental image (difference = -10 ms); words that have many synonyms are categorised more quickly than words that have few alternatives (difference = -7 ms) and words that can be used in numerous contexts are faster to be responded to than words that have a limited breadth of use (difference = -7 ms).

Complete Case and Outlier Analyses. At the time of writing, the full sample model with no outliers was still running.

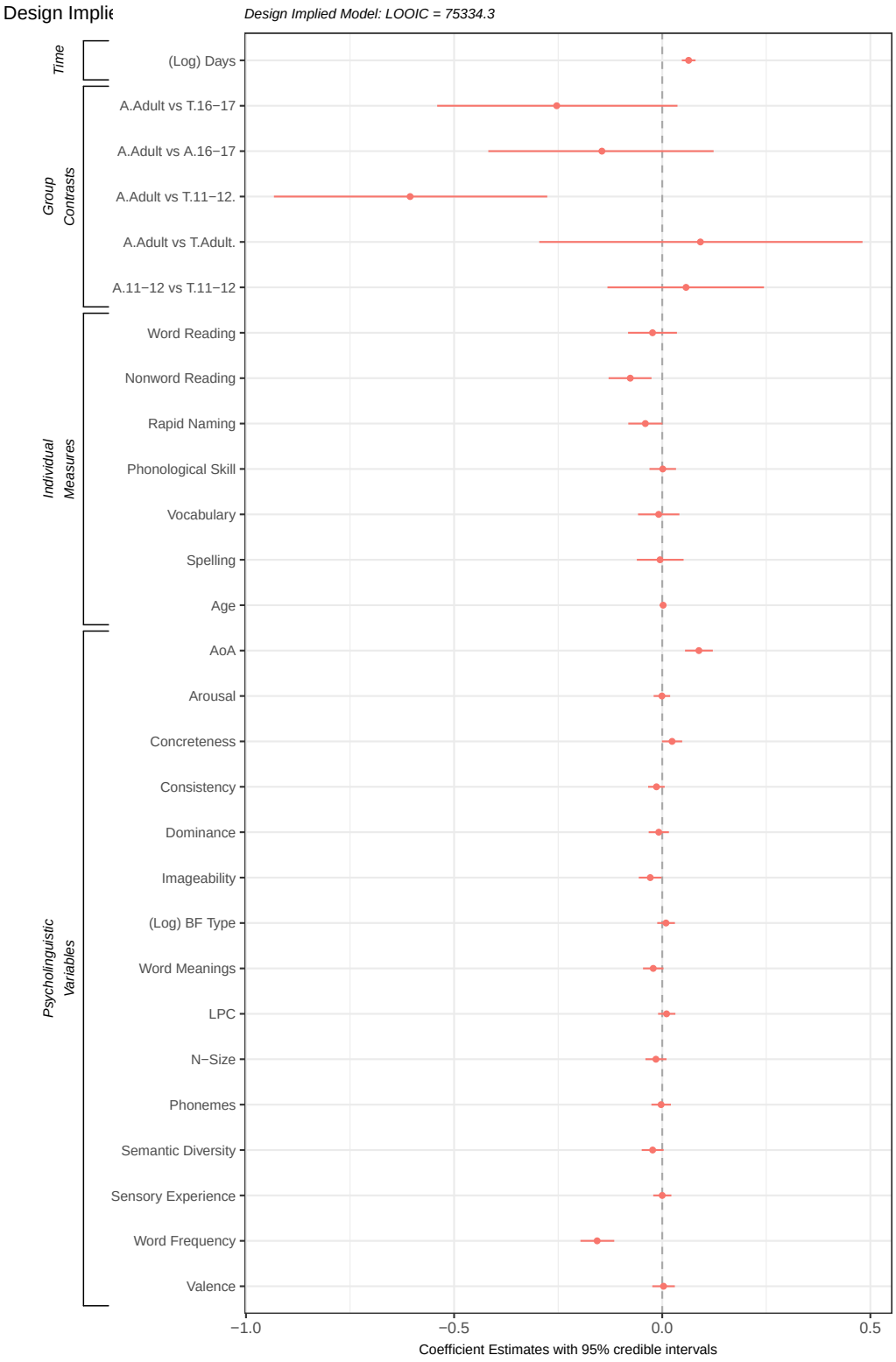
In the complete cases data set, log (days) remains reliable and positive. Nonword reading becomes unreliable but slightly larger while RON remains reliable but shrinks slightly towards zero. AoA and word-frequency remain reliable but are reduced slightly in size. For the purposes of an exploratory study, concreteness is just reliable but smaller. Imageability and semantic diversity remain the same but number of word meanings is now estimated as negative and reliable. Consistency joins the small band of predictors that should be considered under an exploratory label. It shows a negative effect that is very small. Bigram frequency is also resting on zero in this data set, is a very small effect and positive in sign.

The complete cases with no outliers model shows the same pattern of effects as the complete cases data set with the addition of N-size to the group of exploratory predictors. It shows a very small effect size and a negative relationship with reaction time.

Design Implied Model. The design implied model includes the group predictor (see Figure 7.21). All group contrast are unreliable except for the contrast between atypically-reading adults and typical 11-12-year-old readers, where the adults are estimated to be faster in reaction time than the younger readers. It explains the same amount of variance as the preferred model (R^2_{bayes} = 35.5% [34.9, 36.2]). Given the similarities in LOOIC values and variance explained, we performed a model averaging check to see how much weight each model would be apportioned if we considered the

Figure 7.21

Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on the Lexical Decision Reaction Time Data



two side by side. The model averaging result gives 97.1% of the weighting to the model with no group predictor.

It would be easy to be fooled into thinking that the coefficients in the study design implied model are estimated with much greater precision than the preferred model but this is due to the massive uncertainty by which the group contrast coefficients are estimated, such that the x-axis scale is compressed, relative to the x-axis for the preferred model.

The estimates are very similar, however. With the increase in days, reaction times lengthen. RON's upper credible interval now rests on zero but higher skill still indicates faster reaction time. The estimate for nonword reading is stable. AoA and concreteness remain positive and associated with slowed responses for higher values and word-frequency remains negatively associated with reaction time, as in the preferred model. Imageability, number of word meanings and semantic diversity also remain as candidate predictors showing the same direction of effects.

Model Predictions. Model implied predictions for lexical decision reaction time are displayed in Figure 7.22 and 7.23. Recall that because predictors and outcome variables are standardised, where the solid black line crosses the dashed blue line is where the average response is located. In this model, with this data, the average reaction time is based upon that of a typical reading 16-17-year-old. Where intervals and the solid black line are below the dashed line, the predictions are for faster reaction times. Where intervals and the solid black line are above the dashed line, the predictions are for slower reaction times.

The model predicts faster times for higher nonword reading and RON skill. Although not indicated in the model, model implied predictions do indicate a role for word reading (plot b), with higher skills predicting faster reaction times than average. There is also a very slight trend for higher skills in spelling to facilitate reaction time (plot g). The model implied predictions indicate that repeated sessions (plot a) will be associated with slower reaction times, as observed in the model.

Figure 7.22

Preferred Model Predictions for the Effects of Individual Differences on Lexical Decision Reaction Time Data

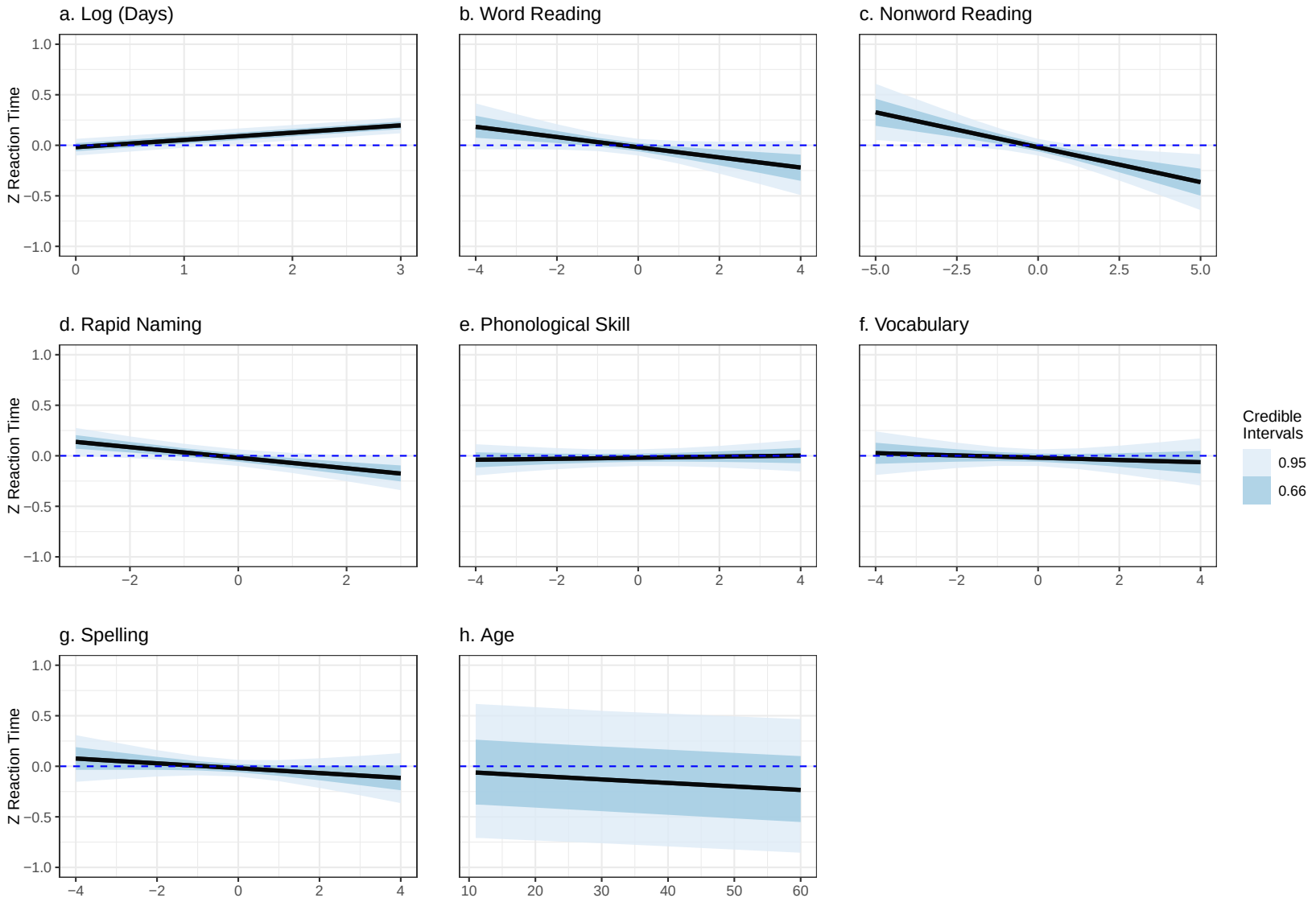
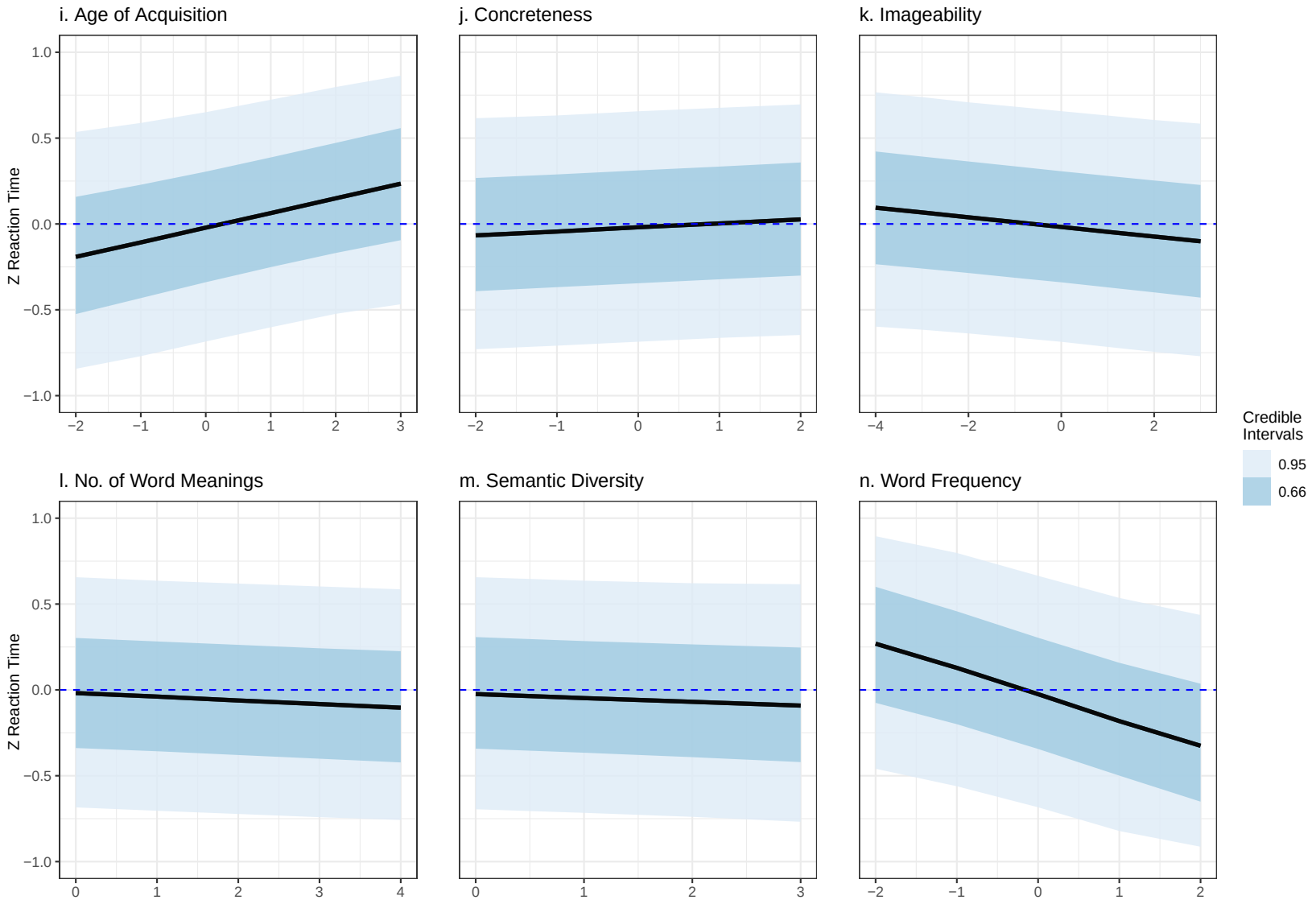


Figure 7.23

Preferred Model Predictions for the Effects of Age of Acquisition, Concreteness, Imageability, Number of Word Meanings, Semantic Diversity and Word-Frequency on Lexical Decision Reaction Time Data



Just as in the accuracy predictions, AoA and word-frequency predictions support the observed estimates of the model. We simulated the semantic predictors since their credible intervals touch on zero. Words that have lower concreteness ratings will have a slight advantage over words of higher concreteness ratings. Yet words that are highly imageable will also have a slight advantage over words that are more abstract. Number of word meanings and semantic diversity are clearly negative for higher values.

Summary and Discussion. Lexical decision reaction time, given this model and this data, shows some very small effects from a couple of ID measures and some small effects from a couple of psycholinguistic predictors. The model that included the group contrast was not the preferred model, suggesting that for this model and sample, the participants approach lexical decision for their speed of responses similarly.

Participant responses become slower with successive data collection points, and There is a very small difference on accuracy outcomes for participants who are slightly older than other participants. Given an updated range of data for age, older participants are faster than child participants by approximately a quarter of a standard deviation. Word-frequency and AoA effects are reliable. The predictors that touch on zero are all of semantic domain - concreteness, imageability, number of word meanings and semantic diversity.

The surprising finding is a positive effect for concreteness. Prior studies have observed negative relationships between estimates for concreteness and reaction time outcomes (Cohen-Shikora and Balota, 2016). Kousta et al. (2011) found a latency advantage for abstract words when imageability ratings were controlled. We have not controlled for imageability by design, but we have statistically adjusted for the independent influence of imageability by including it as a predictor in the model. We suggest that the presence of imageability as a predictor in the model has partialled out the effect of imageability and gives rise to this observed effect. It is an incredibly small effect and in need of confirmation through replication.

7.3 Word Naming

Participants were asked to name a presented word as quickly and accurately as possible. Details of the sample items and list construction are reported in section 5.2.4.2. Our research question was whether the speed and accuracy by which atypically-reading adults named single words was different to that of other groups of readers. Difference between groups could be characterised as rates of recognition but also if the groups differed in their use of psycholinguistic information. If the preferred model includes the group predictor, this may indicate that there are differences between the atypically-reading adult readers for accuracy rates or speed of single word naming. If the preferred model included interaction effects between group and any of the ID or psycholinguistic measures, this could indicate a difference in either strategy or knowledge for completing trials. No interactions between the group variable and ID or psycholinguistic measures would suggest that the groups are approaching the task similarly.

7.3.1 Item Properties

At each time point, a participant would see 70 words (50 words and 20 words for the isolation condition of the sentence reading task). Mean frequency scores for high and low ratings and mean number of letters across lists are in Table 7.9.

A two-way ANOVA for effects of frequency category (high vs low) and list (1-3) on frequency ratings confirmed a significant main effect of frequency category ($F(1, 204) = 856.36$, $p < .001$) and a non-significant main effect of list ($F(2, 204) = 0.05$, $p = .952$). A one-way ANOVA for the effect of list on length confirmed a non-significant finding ($F(1, 207) = 0.09$, $p = .91$). Thus, there is a difference between levels of high and low frequency within lists but lists are equivalent with each other for frequency and length.

The properties for the other psycholinguistic variables and findings of inferential test for differences within variables across list are in Table 7.10 and

Table 7.9

Descriptive Statistics for Frequency and Length for Three Item Lists in the Word Naming Task

	Mean Frequency (SD)		
List	High	Low	Length
1	5.2 (0.7)	2.9 (0.4)	4.7 (1.2)
2	5.3 (0.7)	3 (0.4)	4.6 (1.2)
3	5.2 (0.7)	2.9 (0.4)	4.7 (1.3)

displayed in Figure 7.24. There is a difference for consistency values ($p = .038$), between the second and the third list. None of the other variables show statistical differences across lists (all ps > 0.22).

7.3.2 Analyses

In word naming, a correct response reflects that the participant pronounced an item correctly. An incorrect response represents that either part of the word was pronounced incorrectly or an different word from the target item was pronounced.

7.3.2.1 *Number of Observations*

Full Sample. We collected 40,950 word naming observations. We excluded 420 observations from six participants data as duplicates of items from previous waves of data collection. A further 210 observations were excluded due to microphone errors from three participants. There were two trials that were measured as < 200 ms and five observations that were recorded as above 4,000 ms, all were excluded as mis-trials from malfunctions of equipment. This left 40,313 observations for accuracy analyses. After removing incorrect trials ($n = 2{,}322$) there were 37,991 observations available for reaction time analyses.

Table 7.10

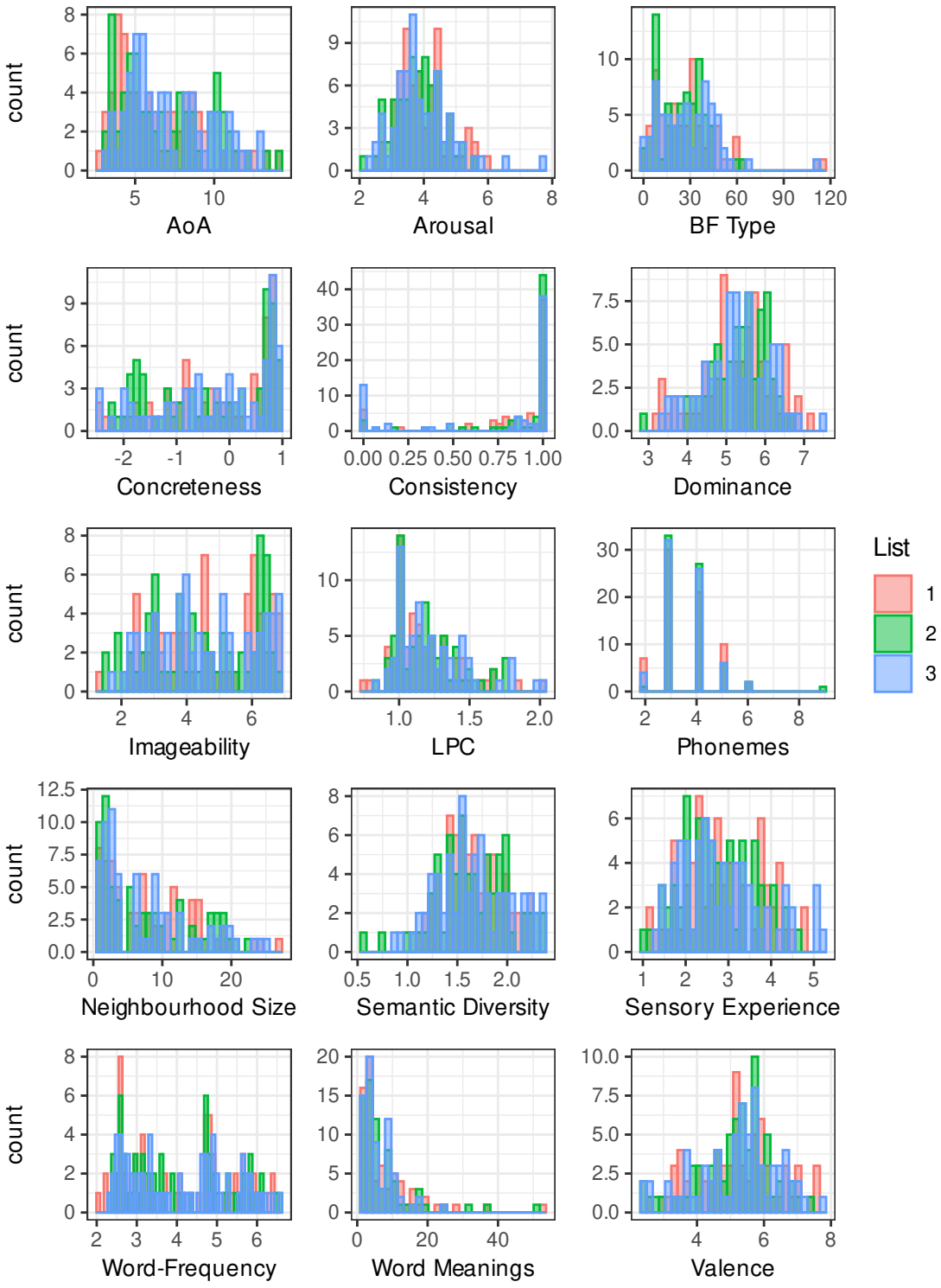
Summary of Psycholinguistic Variable Measures for Word Naming Items with F-Ratio and P Values to Signify Differences Between Item Lists

					ANOVA	
Psycholinguistic Variables	Mean	SD	Min	Max	F(2, 57)	p
AoA	6.8	2.7	2.9	14.2	1.29	0.276
Arousal	3.9	0.8	2.1	7.7	0.7	0.498
BF Type	27.2	18.7	1.0	116.2	1.5	0.226
Concreteness	3.8	1.0	1.7	5.0	0.24	0.788
Consistency	0.8	0.3	0.0	1.0	3.34	**0.038**
Dominance	5.3	0.9	2.8	7.4	0.2	0.819
Imageability	4.5	1.5	1.4	6.9	0.27	0.763
LPC	1.2	0.2	0.8	2.0	0.59	0.553
Phonemes	3.6	0.9	2.0	9.0	0.54	0.584
Neighbourhood size	7.9	6.5	1.0	27.0	0.12	0.887
Semantic diversity	1.7	0.3	0.6	2.4	0.1	0.909
Sensory experience	2.8	0.9	1.0	5.2	0.88	0.417
Word frequency	4.1	1.3	2.0	6.6	0.01	0.99
Word meanings	7.2	7.6	1.0	52.0	1.5	0.226
Valence	5.2	1.2	2.3	7.7	0.12	0.886

Note:
AoA = Age of acquisition. BF = Bigram frequency. LPC = Levenshtein Phonological Consistency.

Figure 7.24

Histograms Showing the Distribution of Psycholinguistic Properties for Items on the Word Naming Task Across Three Lists



Complete Case Analysis. There were 165 participants with complete data for word naming trials across three data collection sessions. The number of observations available for a complete case analysis is 34,393. The preferred model was re-run on this dataset for accuracy and for correct trials in reaction time (n = 32,778).

Outlier Analysis. In the full sample dataset, we removed 320 (0.8%) trials that were at the time-out value and performed the calculations for identifying outliers for each participant, removing them (n = 5,033, 12.5%) leaving 34,960 observations. The preferred model for accuracy was was re-run on this dataset and for reaction time on all correct trials (n = 33,708).

In the complete case dataset with no outliers, there were 30,059 observations for the accuracy model. For analysis of reaction time outcomes on correct trials only, there were 29,061 observations.

7.3.3 Accuracy Results

7.3.3.1 *Descriptive Statistics*

We calculated mean accuracy performance per participant and display the distributions within groups across time in Figure 7.25; averages across accuracy and reaction time by time and group are displayed in Figure 7.25. As may be apparent from the numbers of observations above, and from the plots, accuracy rates were very high in this dataset. There is very little variance observed for the typically-reading 16-17-year-olds and adult readers. Atypical reading groups show a greater spread of mean accuracy rates, however their distributions are still display quite extreme negative skew. Estimation of effects under these conditions can be tenuous and results may be unstable.

Figure 7.25

Histograms Showing the Distribution of Mean Accuracy Rates per Participant By Group Across Time Points in the Word Naming Task Across Time

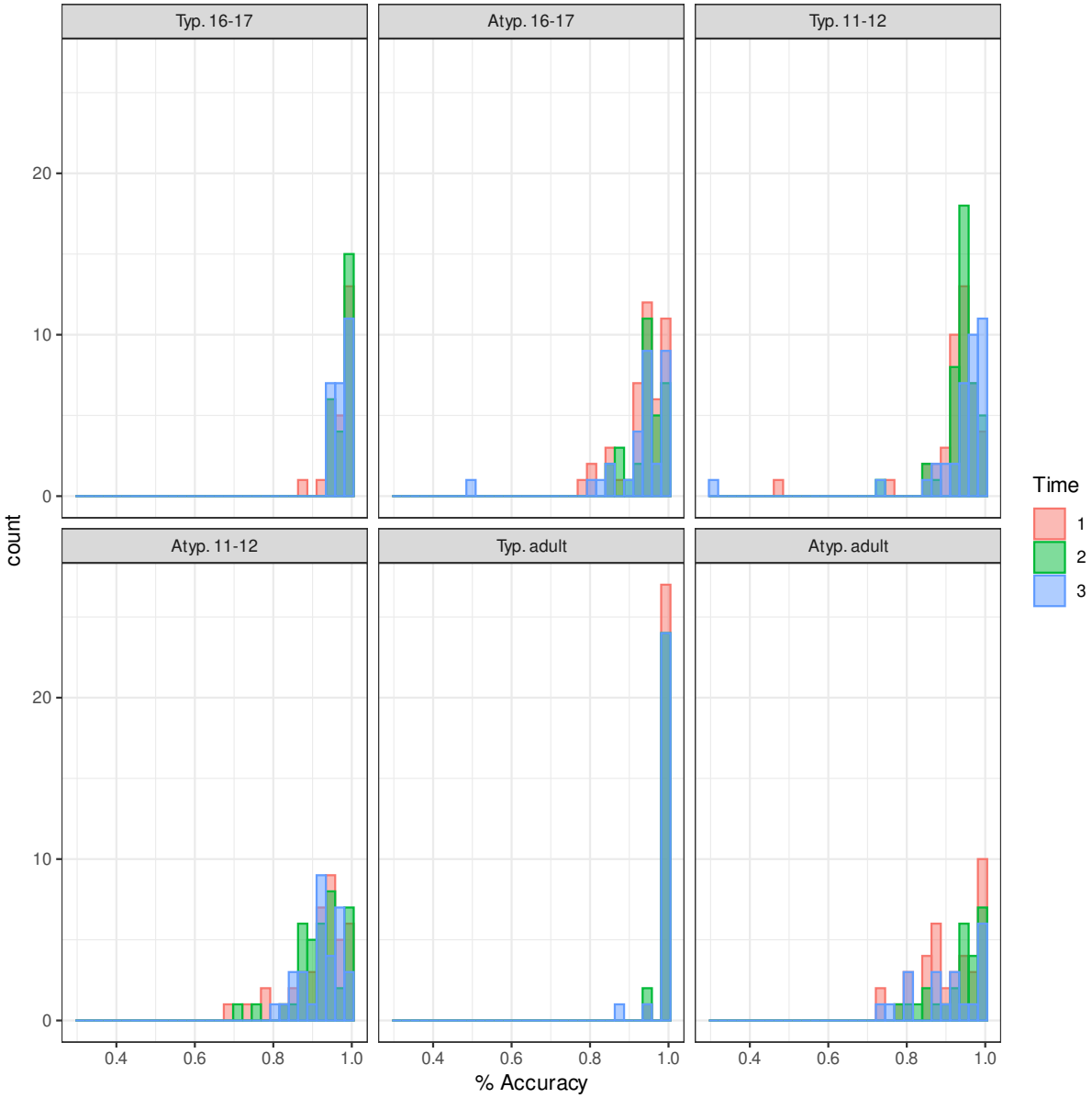


Figure 7.26

Mean Accuracy and Raw Mean RT for Correct Trials By Group and Time for Word Naming

	Time 1			Time 2			Time 3		
Group	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typ. 16-17	97.3	691.5	239.8	97.8	723.3	244.6	97.4	726.6	254.5
Atyp. 16-17	93.9	719.3	267.8	94.7	755.2	293.3	92.6	776.1	318.8
Typ. 11-12	92.5	779.1	331.9	94.2	771.6	306.0	93.4	792.2	309.8
Atyp. 11-12	91.5	801.2	320.4	91.9	836.7	339.1	92.5	861.5	356.7
Typ. adult	99.4	673.5	185.2	99.1	685.4	195.9	98.7	670.3	211.5
Atyp. adult	91.0	803.5	326.7	93.2	788.0	334.2	90.6	788.2	300.0

Note:
Acc % = Percentage Accuracy; RT = Reaction time; Atyp. Adult = Atypically-reading adult; Typ. 16-17 = Typically-reading 16-17-year-old; Atyp. 16-17 = Atypically-reading 16-17-year-old; Typ. 11-12 = Typically-reading 11-12-year-old; Typ. Adult = Typically-reading adult; Atyp. 11-12 = Atypically-reading 11-12-year-old.

7.3.3.2 *Preferred Model*

The preferred model for word naming accuracy data was the Additive-RIS model that included the predictors for time, group contrasts, ID and psycholinguistic variables, with random intercepts and slopes for ID and psycholinguistic predictors on participants and items. This model explained R^2_{bayes} = 37.5% [36.1, 39.0] of the variance in the accuracy outcome. This model is also that implied by the study design.

Table 7.11

Summary of Standardised Fixed Effects for Word Naming Accuracy

Term	Estimate	SE	Lower CI	Upper CI
Intercept	5.42	0.22	5.00	5.86
Time				
(Log) Days	-0.08	0.04	-0.16	-0.01
Group Contrasts				

Table 7.11

Summary of Standardised Fixed Effects for Word Naming Accuracy (continued)

Term	Estimate	SE	Lower CI	Upper CI
A. Adult vs T. 16-17	-1.46	0.41	-2.28	-0.65
A. Adult vs A. 16-17	-1.04	0.38	-1.78	-0.28
A. Adult vs T. 11-12	-1.24	0.47	-2.14	-0.29
A. Adult vs T. Adult	-1.86	0.65	-3.17	-0.63
A. 11-12 vs T. 11-12	-0.10	0.22	-0.53	0.32
Phonemic Onsets				
Voice	0.07	0.14	-0.20	0.33
Nasal	0.08	0.12	-0.16	0.32
Fricative	0.16	0.16	-0.16	0.47
Liquid_SV	0.13	0.13	-0.13	0.39
Bilabials	-0.27	0.18	-0.62	0.08
Alveolars	0.02	0.17	-0.32	0.36
Palatals	0.04	0.13	-0.21	0.30
Velars	-0.08	0.17	-0.42	0.25
Glottals	0.11	0.13	-0.15	0.37
Individual Differences				
Word reading	0.00	0.12	-0.23	0.24
Nonword reading	0.28	0.10	0.08	0.47
Rapid naming	0.03	0.07	-0.10	0.16
Phonological skill	-0.01	0.05	-0.11	0.08
Vocabulary	0.44	0.08	0.27	0.60
Spelling	0.22	0.08	0.06	0.39
Age	0.01	0.01	-0.01	0.02
Psycholinguistic Variables				
AoA	-0.29	0.18	-0.64	0.05

Table 7.11

Summary of Standardised Fixed Effects for Word Naming Accuracy (continued)

Term	Estimate	SE	Lower CI	Upper CI
Arousal	-0.11	0.11	-0.33	0.11
Concreteness	0.09	0.13	-0.17	0.35
Consistency	0.37	0.11	0.16	0.58
Dominance	0.04	0.13	-0.22	0.30
Imageability	-0.04	0.15	-0.34	0.26
(Log) BF Type	-0.35	0.13	-0.61	-0.09
Word meanings	0.19	0.14	-0.08	0.48
LPC	-0.09	0.12	-0.32	0.14
N-Size	0.19	0.14	-0.09	0.47
Phonemes	0.04	0.13	-0.23	0.30
Semantic diversity	-0.08	0.15	-0.37	0.21
Sensory experience	0.04	0.12	-0.20	0.27
Word-frequency	0.95	0.21	0.54	1.36
Valence	0.00	0.14	-0.28	0.29

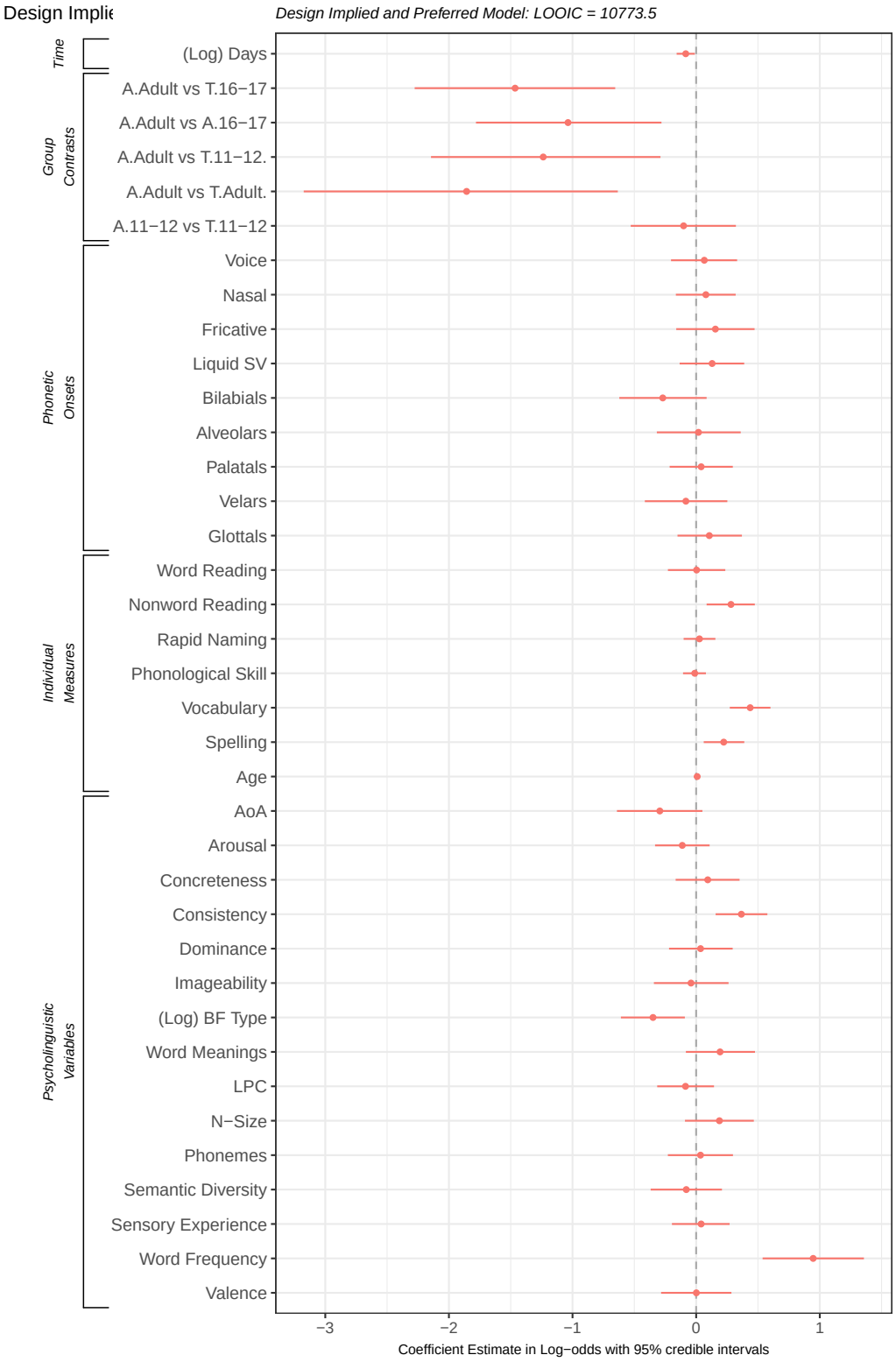
Note:

CI = Credible intervals. A. Adult = Atypically-reading adult; T. 16-17 = Typically-reading 16-17-year-old; A. 16-17 = ATypically-reading 16-17-year-old; T. 11-12 = Typically-reading 11-12-year-old; T. Adult = Typically-reading adult; A. 11-12 = Atypically-reading 11-12-year-old; AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency. N-Size = Neighbourhood size.

Model Inference. As suggested by Figure 7.25, the accuracy rate for the word naming task was very high. Within the full sample observed data, accurate answers

Figure 7.27

Estimates from the Posterior Distribution of the Preferred Model for Group, Phonemic Onsets, ID and Psycholingustic Predictors on the Word Naming Accuracy Data



were at 97.3%. The intercept indicates that when all predictors are at their mean, accuracy is a log-odds of 5.4 - which transforms to a probability rate of 99.5%.

With accuracy above 95%, we have to interpret the coefficients with caution. Although the model returns some coefficients whose credible intervals lie far from zero, and are measured with some certainty, the biggest effect only increases probability of being correct by 0.3%.

There is a small negative effect of log (days) (log-odds = -0.08 [-0.16, -0.01]) indicating that, over time there are lower odds that a response will be accurate. The coefficient credible intervals do not include zero, so we can be confident that the model implied effect is negative. However the magnitude of the effect is so small so as to have no discernible effect on the intercept term (-0.03% difference on the intercept).

Model implied estimates for the group contrast predictor appear to be much stronger, with the coefficients for differences between atypically-reading adults and the other groups being far away from zero and negative. Atypically-reading adults are less likely to be accurate than the typically- and atypically-reading 16-17-year-olds, as well as the typically-reading 11-12-year-olds and adults. The model is inconclusive about the sign of any difference between the two 11-12-year-old reading groups. Atypically-reading adults and typically-reading 16-17-year-olds differ by approximately 3%, this rises to a difference of approximately 4.5% between atypically-reading adults and typically-reading adults. The difference between the atypically-reading adults and the typically-reading 11-12-year-olds is a lower accuracy rate of approximately 2% and the difference between atypically-reading adults and atypically-reading 16-17-year-olds is a lower accuracy rate of approximately 1.6%.

In this model, none of the phonetic onset terms predict a reliable difference in accuracy rates.

Nonword reading skill, vocabulary knowledge and spelling show positive relationships with word naming accuracy. Higher scores in each of these measures is associated with higher odds of producing a correct pronunciation. Nonword reading (log-odds = 0.28 [0.08, 0.47]) and spelling (log-odds = 0.22 [0.06, 0.40]) are very small effects, each predict an increase in the probability of being accurate of 0.1% for a 1 SD

increase in the respective skill scores. Vocabulary (log-odds = 0.44 [0.27, 0.60]), as a small effect size, increases probability of a correct pronunciation by 0.2% for a 1 SD increase in knowledge scores. The remaining individual difference estimates are unreliable.

Consistency, word-frequency and log bigram frequency (type) are indicated as reliable psycholinguistic predictors. Consistency shows a very small, positive relationship (log-odds = 0.37 [0.16, 0.58]), with a 1 SD increase in consistency value increasing the probability of being correct by 0.2%. As word-frequency increases by 1 SD, accuracy increases by 0.3% (log-odds = 0.95 [0.54, 1.36]). Log bigram frequency for the type of word shows a very small negative relationship (log-odds = -0.34 [-0.61, -0.09]). A 1 SD increase in log bigram frequency values decreases accuracy by 0.2%.

Complete Case and Outlier Analyses. The complete case model shows a smaller, now unreliable effect for log (days). The model is still certain that atypically-reading adults have lower odds than typically-reading adults, 11-12- and 16-17-year olds for naming a word correctly. The coefficient for the group contrast between atypically-reading adults and atypically-reading 16-17-year-olds is not reliable in the complete cases model. The accuracy rates between the two 11-12-year-old groups remains equivalent. Nonword reading shows a stronger effect, while vocabulary and spelling are slightly reduced in size. All remain reliable. Consistency, bigram frequency and word-frequency also remain reliable and are stronger effects in this model.

While the group contrasts remain stable when outliers are removed in the full sample dataset, the contrast effects between atypical reading adults and atypical reading 16-17-year-olds and typical reading 11-12-year-olds are attenuated in the complete case dataset with no outliers. The model is no longer confident that the atypically-reading adults are likely to be less accurate than these other groups. All other predictor effects remain stable, reflecting those of the full sample model.

Model Predictions. The model implied predictions for word naming accuracy are shown in Figure 7.28. As before, these plots are drawn on the probability scale with the y-axis ranging from 0 - 1. The granularity of the axis does not match the small effects well but we felt that to change the axis for these plots would distort the image of these effects relative to the other outcomes for other tasks.

The model implied predictions reflect the ceiling level of the observed accuracy rates with the solid black line to the uppermost part of plots a - k. In plot c, f and g, the model predicts that even very extreme low scores for nonword reading, vocabulary and spelling decrease the certainty around the probability of a correct response to a very small extent.

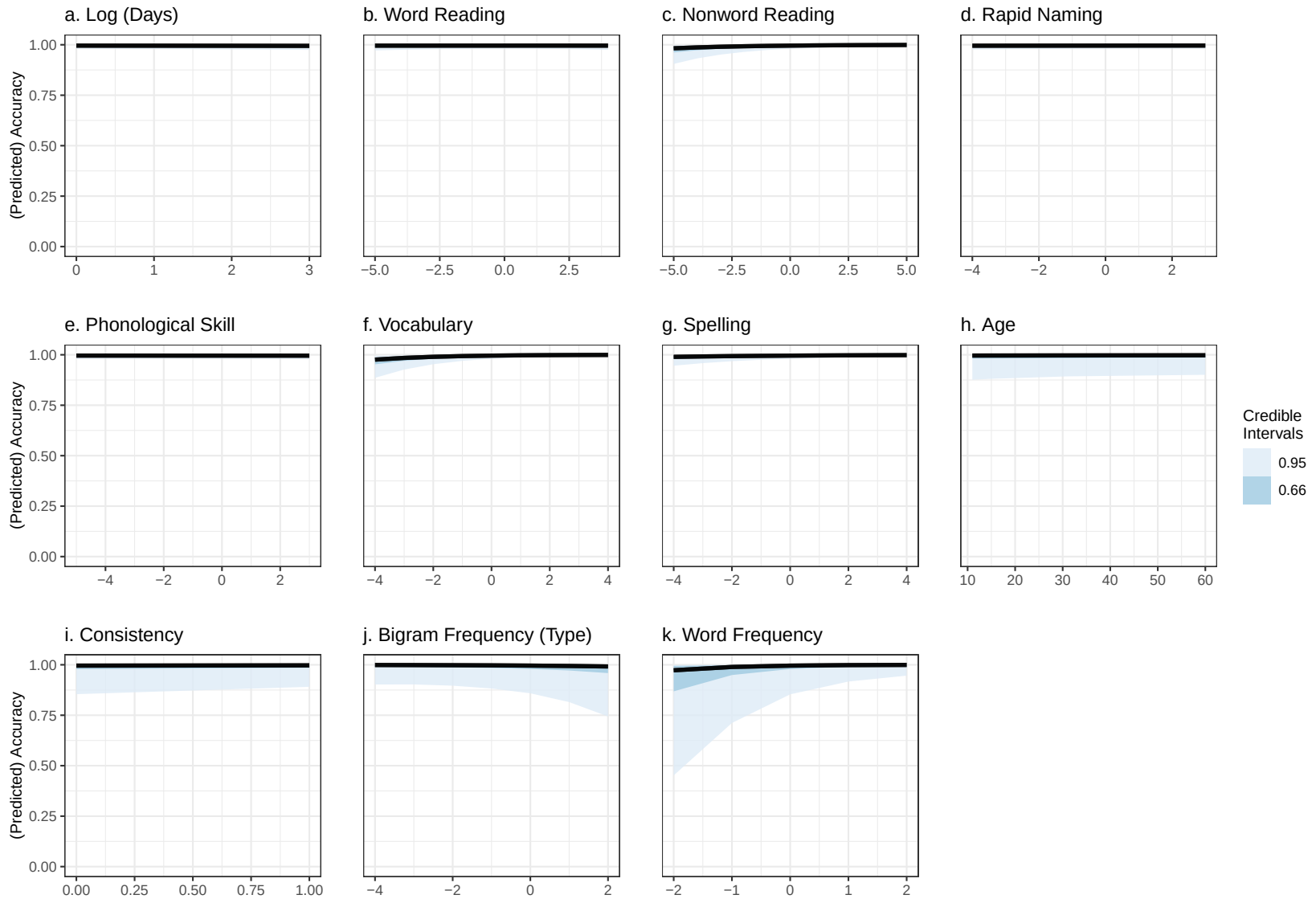
Predictions for consistency and word-frequency show high certainty at the upper levels of their ranges, with probability of a correct response falling potentially below 75% and 50% respectively, at the lower extremes. In contrast, target words that have a low number of bigrams that are shared with other words show a narrower range of uncertainty while words that show a high number of shared bigrams have greater uncertainty, with the model predicting accuracy rates of 75% as plausible.

Summary and Discussion. The preferred model for word naming accuracy includes the group contrast predictor. It is not the interaction model. Consequently, it is quantitative rates of accuracy that differ rather than a qualitative difference in approach. This is the first set of results that describe lower performance for atypically-reading adults across all of the contrasted groups. It is an unstable effect since it disappears in the sensitivity analysis. An inspection of the group means for word reading skill shows that the atypically-reading adults resemble the atypically-reading 11-12-year-olds most closely.

Nonword reading skill, vocabulary knowledge and spelling skill are positively associated with word naming accuracy. This feels like a nexus of phonological, semantic and orthographic skills that reflect the three domains of the lexical quality hypothesis. We are surprised that the estimates for word reading skill are inconclusive in this model.

Figure 7.28

Preferred Model Predictions for the Effects of Individual Differences, Consistency, Bigram Frequency and Word-Frequency on Word Naming Accuracy Performance



Interpreted with respect to the lexical quality hypothesis, we suggest that the three components are working together, but are not integrated. Interpreted through a connectionist model of reading, this implicates a role for the phonological and semantic route as suggested by the division of labour hypothesis. Lack of integration from a lexical quality hypothesis perspective could also mean the interdependence of the two components is weak.

The role of spelling might also be cast as phonological in a feed-backward consistency mechanism (Balota et al., 2004; Ziegler et al., 1997). Whichever is chosen, spelling represents a fragmented source of information for the orthographic function. The absence of a definitive word reading skill effect may suggest that word units as a whole are not the most efficient route to word identification for this sample, i.e. the orthographic route of the connectionist model is not playing a strong role.

Consistency, log bigram frequency (type) and word-frequency are implicated as psycholinguistic predictors of word naming accuracy. We need to interpret the consistency effect with caution because there were significant differences between items on list 2 and 3. The consistency measure in this sample of items is constructed as a ratio of friends and enemies, with the rime of a word always sharing letters but not necessarily the same pronunciation. The bigram frequency for type of word is a count of letter pairs across the entire body of a word, irrespective of the sound of the pair. This could be understood as a phonological effect of a larger section of a word facilitating greater accuracy while the sublexical attributes of letter-pair frequency, that can occur at any point within the letter string and leaves the larger part of a word still to be activated, acts as competition. Where the pairs of letters have high frequency, many candidate words are activated. A larger selection pool introduces a greater probability of making an error.

7.3.4 Reaction Time Results

Reaction time data was log transformed to reduce skew before standardising using the typically-reading 16-17-year-olds as our reference level (mean RT = 691.5 ms, SD =

239.8). Consequently, positive coefficients indicate slower reaction times and negative coefficients indicate faster reaction times than that of a typically-reading 16-17-year-old.

7.3.4.1 *Descriptive Statistics*

Distributions for the raw, mean reaction time in milliseconds on correct trials per participant are displayed at the group level in Figure 7.29 and Figure 7.25 for correct responses to words.

7.3.4.2 *Preferred Model*

The preferred model for reaction time data in the word naming task is the Additive-RIS model, containing fixed effects for time, phonetic onsets, ID measures and psycholinguistic variables, random intercepts and slopes for participants and items. The explained variance in the reaction time outcome was $R^2_{\text{bayes}} = 39.3\%$ [38.8, 39.9]. The model coefficients are plotted in Figure 7.30 with the model summary of fixed effects in Table 7.12.

Model Inference. Neither number of days between data collection sessions nor any of the ID measures are reliable in this model. Each predictor's credible interval crosses zero. Estimates are very small for this model and this sample.

For exploratory purposes, we note the results whose credible intervals just cross zero. Although inconclusive, the general trend across the ID measures is for negative estimates that suggest individuals with higher scores are quicker to begin a correct pronunciation. RON (β = -0.04 [-0.09, 0.01]) and vocabulary (β = 0.03 [-0.02, 0.08]) need exploring further. The vocabulary measure suggests a positive relationship and that readers of higher vocabulary knowledge are slower to begin a correct pronunciation than those with lower vocabulary knowledge. Age shows absolutely no influence on reaction time in this model for this sample (β = 0.00 [0.00, 0.00]).

Figure 7.29

Histograms Showing the Distribution of Raw, Mean Reaction Time (ms) By Participant, Group and Time Point for Correct Pronunciations in the Word Naming Task

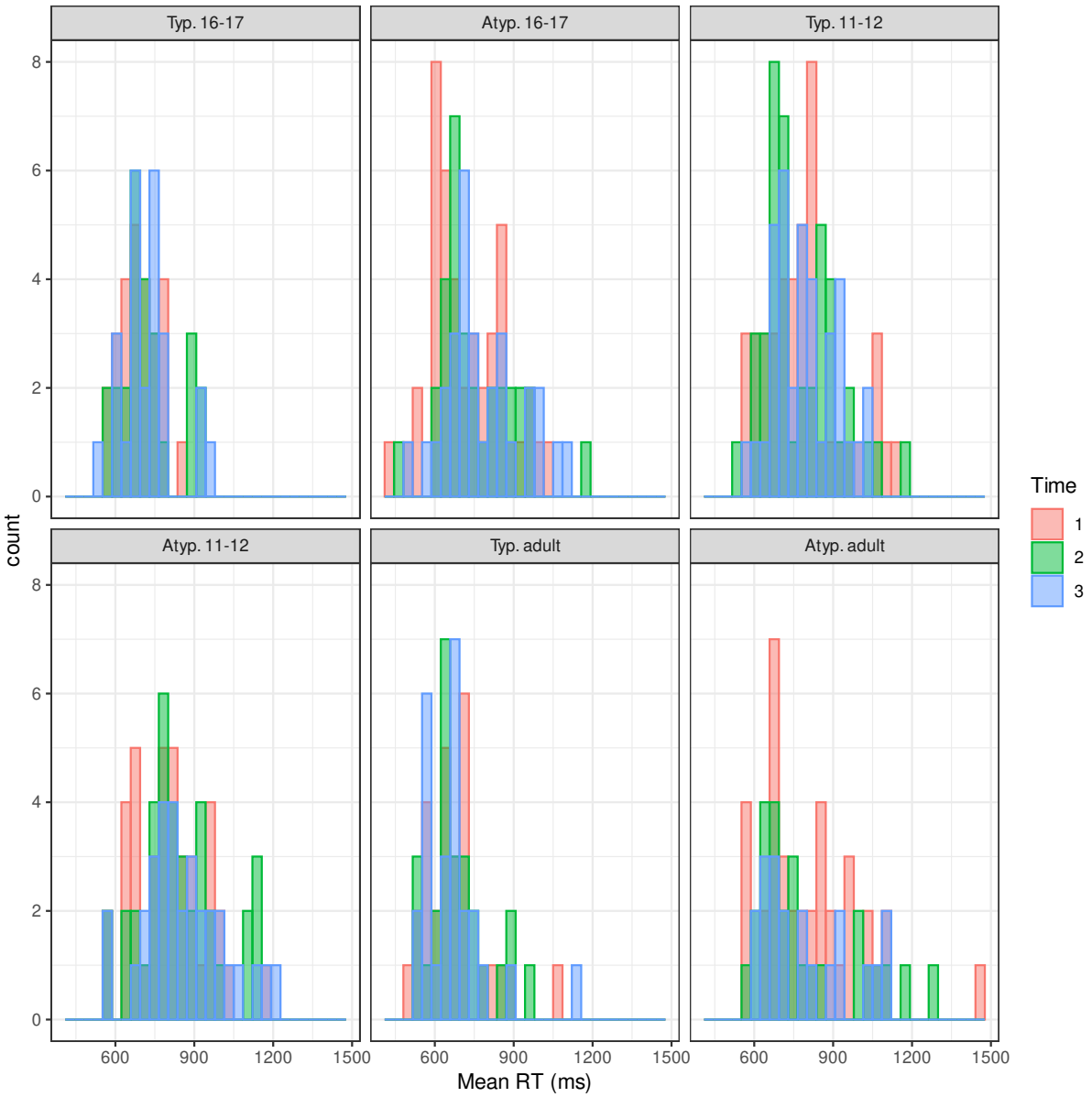


Figure 7.30

Estimates from the Posterior Distribution of the Preferred Model for ID and Psycholingustic Predictors on the Word Naming Reaction Time Data

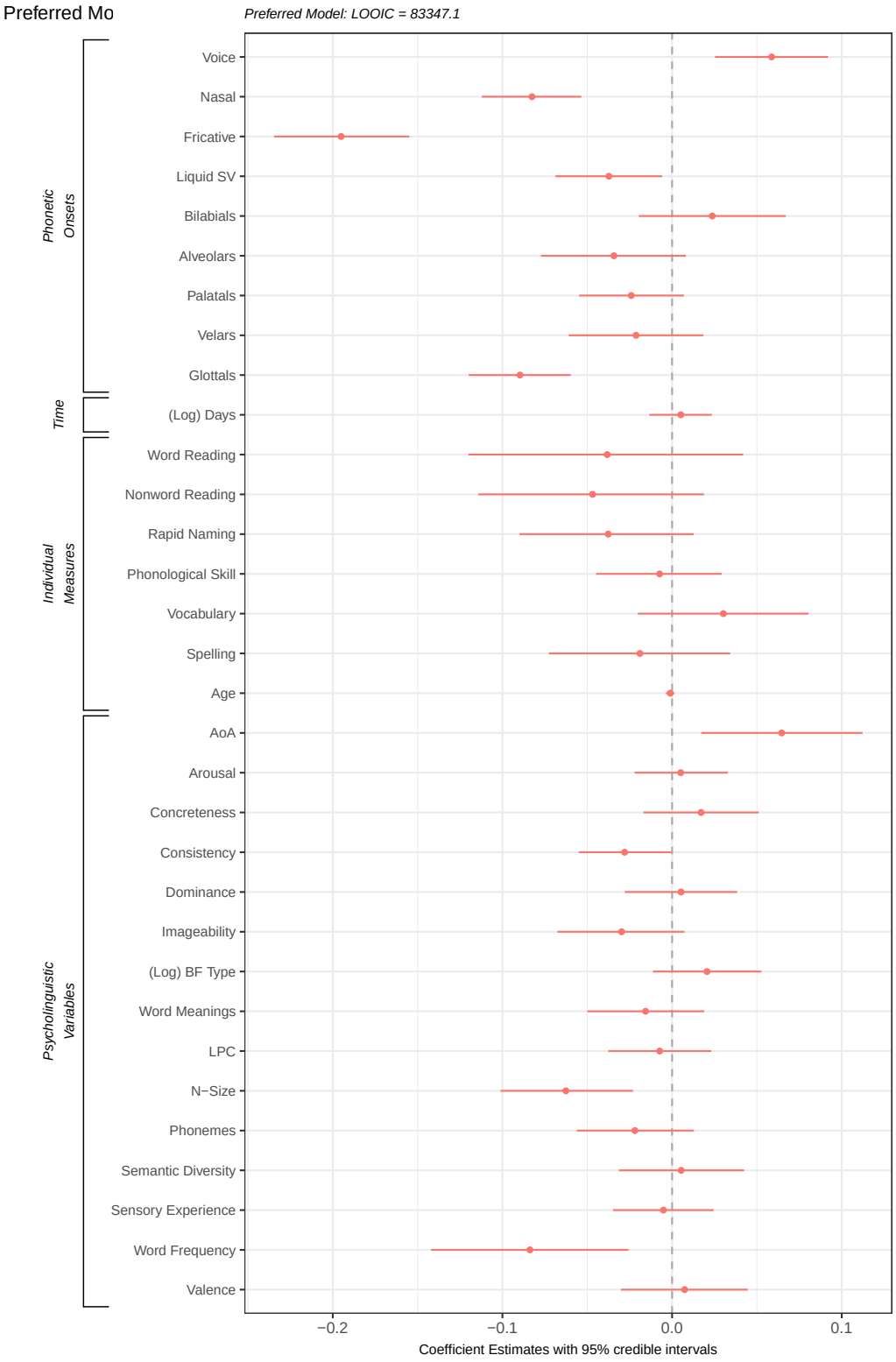


Table 7.12

Summary of Standardised Fixed Effects for Word Naming Reaction Time

Term	Estimate	SE	Lower CI	Upper CI
Intercept	0.23	0.05	0.13	0.34
Phonemic Onsets				
Voice	0.06	0.02	0.03	0.09
Nasal	-0.08	0.02	-0.11	-0.05
Fricative	-0.20	0.02	-0.23	-0.15
Liquid_SV	-0.04	0.02	-0.07	-0.01
Bilabials	0.02	0.02	-0.02	0.07
Alveolars	-0.03	0.02	-0.08	0.01
Palatals	-0.02	0.02	-0.05	0.01
Velars	-0.02	0.02	-0.06	0.02
Glottals	-0.09	0.02	-0.12	-0.06
Time				
(Log) Days	0.01	0.01	-0.01	0.02
Individual Differences				
Word reading	-0.04	0.04	-0.12	0.04
Nonword reading	-0.05	0.03	-0.11	0.02
Rapid naming	-0.04	0.03	-0.09	0.01
Phonological skill	-0.01	0.02	-0.04	0.03
Vocabulary	0.03	0.03	-0.02	0.08
Spelling	-0.02	0.03	-0.07	0.03
Age	0.00	0.00	0.00	0.00
Psycholinguistic Variables				
AoA	0.06	0.02	0.02	0.11
Arousal	0.01	0.01	-0.02	0.03
Concreteness	0.02	0.02	-0.02	0.05
Consistency	-0.03	0.01	-0.05	0.00
Dominance	0.01	0.02	-0.03	0.04
Imageability	-0.03	0.02	-0.07	0.01
(Log) BF Type	0.02	0.02	-0.01	0.05
Word meanings	-0.02	0.02	-0.05	0.02
LPC	-0.01	0.02	-0.04	0.02
N-size	-0.06	0.02	-0.10	-0.02
Phonemes	-0.02	0.02	-0.06	0.01
Semantic diversity	0.01	0.02	-0.03	0.04
Sensory experience	-0.01	0.02	-0.03	0.02
Word-frequency	-0.08	0.03	-0.14	-0.03
Valence	0.01	0.02	-0.03	0.04

Note:
CI = Credible intervals; AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency. N-Size = Neighbourhood size.

The control variables for phonetic onsets show a slowing of reaction time for voiced onsets and faster responses for onsets in the nasal, fricative, liquid SV and glottal position.

AoA, N-size and word-frequency are identified as reliable psycholinguistic predictors for reaction time. AoA shows a positive association with reaction time (β = 0.06 [0.02, 0.11]), describing how later learned words are slower (by approximately 14 ms) to be pronounced correctly than earlier learned words for a 1 SD increase in AoA values.

N-size is estimated as having a reliable negative relationship with word naming reaction time for this model and sample (β = -0.06 [-0.10, -0.02]). A word from a neighbourhood that is 1 SD larger is faster to be named by approximately 14 ms.

Word-frequency is negatively associated with reaction time (β = -0.08 [-0.14, -0.03]): words of higher frequency are associated with shorter reaction times. Each 1 SD increase in frequency (SD = 1.3) results in a decrease in reaction time of about 19 ms.

Consistency's upper credible interval touches on zero (β = -0.03 [-0.05, 0.00]). The negative sign of the coefficient suggests that for a 1 SD increase in consistency value, a word will be correctly pronounced approximately 7 ms faster.

Complete Case and Outlier Analyses. At the time of writing, the full sample with no outliers model was still to converge.

The complete case analysis shows the same pattern of effects as the full sample model for time passing and ID measures - all remain unreliable.
Word-frequency and AoA remain reliable and relatively stable in size. N-size becomes unreliable. Consistency remains just reliable and imageability is also just touching on zero so needs including as an exploratory study.

The complete case with no outliers is essentially the same as the complete cases with outliers model. The phonemes predictor is suggested as reliable as an addition. It is negative in size and very small. This describes faster reaction times for

words with a greater number of phonemes.

Design Implied Model. The model coefficients for the design implied model are plotted in Figure 7.31. The model adds the group contrast predictor. Only two of the contrasts are reliable, although the trend of the estimates is for the atypically-reading adults to be slower than the group with which they are contrasted. A reliable estimate is present for the contrast between atypically-reading adults and atypically-reading 16-17-year-olds ($\beta = 0.31$ [0.06, 0.63]) which equates to a difference of approximately 74 ms, and between atypically- and typically-reading 11-12-year-olds ($\beta = 0.23$ [0.02, 0.46]) which is a difference of approximately 55 ms. All other predictors remain the same.

Model Predictions. Model implied predictions for word naming reaction time are displayed in Figure 7.32. The model implied predictions follow the observed data well for the effect of log (days) with plot (a) drawn as a flat line across repeated sessions. The predictions for all ID measures are negative but for vocabulary. The predictions for age show a slight negative influence, suggesting that older individuals may be slightly faster than younger individuals. The credible intervals also show a much wider range of values than for the other ID measures, showing how uncertain the model is for this variable.

We simulated AoA, consistency, N-size and word-frequency. No matter the ratings of the variables, none drop below the observed data mean value. Overall, the model predicts slower reaction times than observed in the data.

Figure 7.31

Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on Word Naming Reaction Time Data

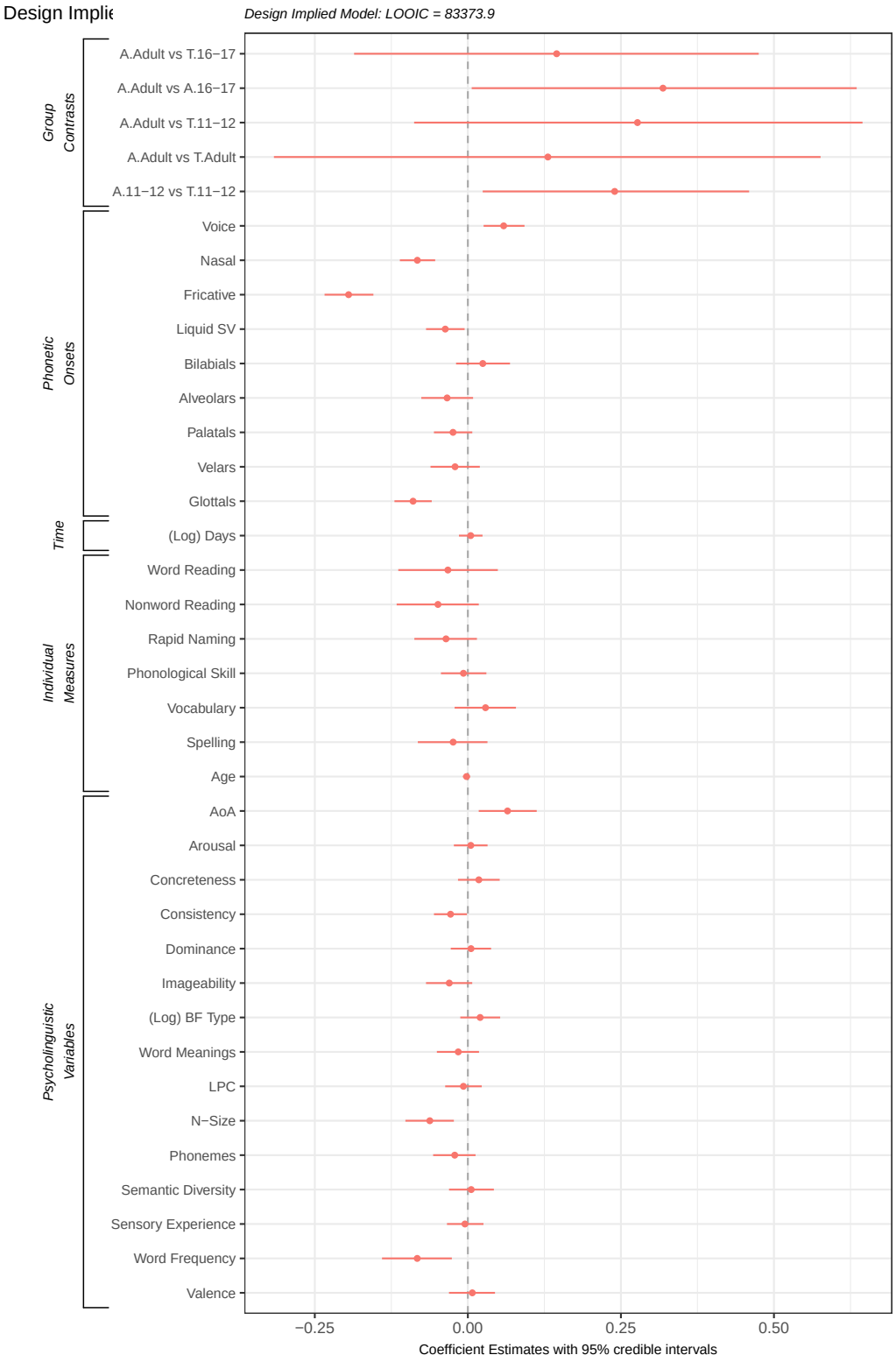
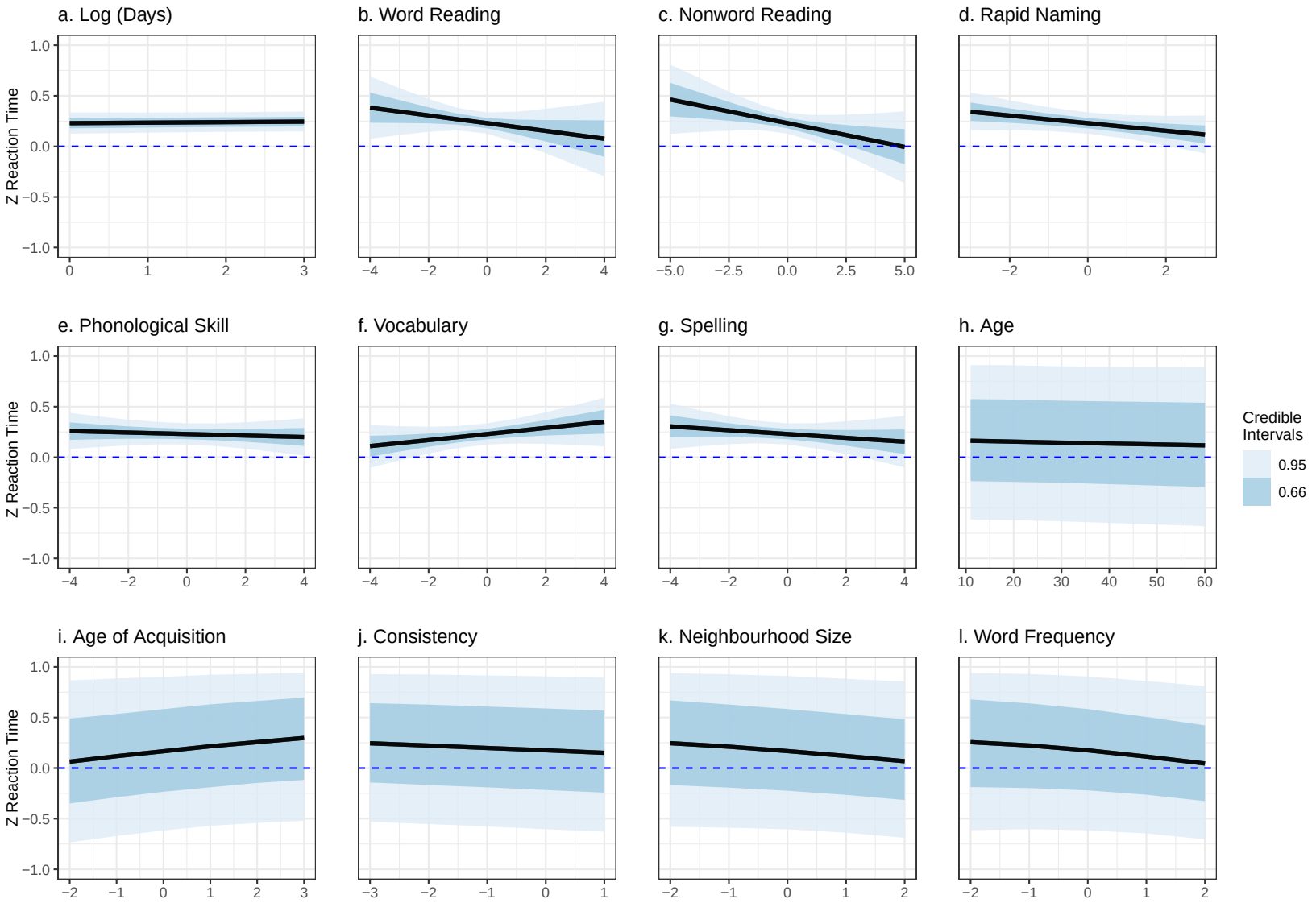


Figure 7.32

Preferred Model Predictions for the Effects of Individual Differences, Age of Acquisition, Consistency, Neighbourhood Size and Word-Frequency on Word Naming Reaction Time Data



Summary and Discussion. The word naming models are estimated on a dataset that is characterised by a high level of accuracy, suggesting that, possibly the sample items were too easy. Well known words are likely to be responded to much more quickly, which reduces the variability on the outcome measure by which to detect effects.

For this model and data, the ID measures are important for accuracy. The atypically-reading adults are observed as having lower odds of naming a word correctly than the group with which they are contrasted. They are predicted to have lower odds than the typically-reading 11-12-year-olds. This, being the task that in prior literature, is supposed to be a relative strength for atypically-reading adults, is a weakness.

Two of the ID measures - vocabulary and spelling - that are implicated to assist with word naming accuracy, are the relatively strong ID measures for the atypically-reading adults. The nonword reading skill is relatively weak in atypically-reading adults. Consequently, Plaut et al. (1996)'s division of labour hypothesis would have semantics supporting the phonological skill of nonword reading to accomplish the recognition of the orthographic form, which appears to be best represented by the spelling predictor.

This is not true for the reaction time data, however. The preferred model does not include group and shows no influence of ID measures. All things being equal across person-level measures, the psycholinguistic predictors that assist speeded word recognition are a canonical set of AoA, consistency, N-size and word-frequency. Early learned words, with straightforward decoding patterns of high frequency are named faster. However, the accuracy rate is very high and replication is needed to confirm these results, especially since one) this task implicates group differences for the atypically-reading adults and two) in the sensitivity analyses the psycholinguistic predictors changed across models. While AoA, word-frequency and consistency were relatively stable, under the different conditions of the data sets, N-size, imageability and phonemes were observed to change.

7.4 Sentence Reading

Participants were asked to silently read a stem sentence that was missing the final word. The final word then appeared in isolation on the next screen and the task was to pronounce the single word as quickly and accurately as possible. Details of the sample items and list construction are reported in section 5.2.4.3.

There was a context predictor in the task, of three levels. The first was no context - reading a word in isolation. These words were presented as additional items in the word naming task. The second and third levels involved a sentential context, neutral and meaningful. The neutral reading context is set as the reference level for this variable.

There were two research questions for this task. The first was whether the speed and accuracy by which atypically-reading adults named single words across each sentence reading context was different to that of other groups of readers. The second research question was how the addition of context affected accuracy and speed of word naming.

We expected the meaningful context to facilitate both accuracy and reading rate for all groups, compared to the neutral condition because it contributes semantic priming information. We were agnostic about where the isolated reading condition would place, relative to the other two conditions. As an extension, we expected vocabulary knowledge to be positively associated with accuracy and negatively associated with reaction time on this task since meaningful context suggests that semantic properties of a word may be activated while reading the sentence stem and preparing candidate words for pronunciation. This would elevate both the probability of the correct word being pronounced and also the speed with which the pronunciation begins.

If the preferred model includes the group predictor, this may indicate that there are differences between the atypically-reading adult readers for accuracy rates or speed of word identification in the context of sentence reading. If the preferred model included interaction effects between group and any of the ID or psycholinguistic

Table 7.13

Descriptive Statistics for Sentence Reading Items Across Three Lists

	Mean Frequency (SD)		
List	High	Low	Length
1	4.6 (0.4)	3.4 (0.2)	4.6 (1.2)
2	4.7 (0.3)	3.3 (0.2)	4.6 (1.2)
3	4.6 (0.4)	3.3 (0.1)	4.5 (1.1)

measures, this could indicate a difference in either strategy or knowledge for completing trials. No interactions between the group variable and ID or psycholinguistic measures, or a preferred model that did not include the group variable would suggest that the groups are approaching the task similarly.

7.4.1 Item Properties

At each time point, a participant would see 20 words in the isolation task embedded within the word naming task, and a further 40 (20 x 2) words for the neutral and meaningful conditions within a separate sentence reading task. Mean frequency scores for high and low ratings across lists are in Table 7.13. A two-way ANOVA for effects of frequency category (high and low) and list (1-3) on frequency ratings confirmed a significant main effect of frequency category ($F(1, 174) = 986.32$, $p < .001$) and a non-significant main effect of list ($F(2, 174) = 0.4$, $p = .674$). Thus, there is a difference between high and low frequency of words within lists and no evidence to suggest differences in levels of high and low frequency between lists.

We display the mean and SD values for the measured ratings across words for other psycholinguistic variables in Table 7.14. The final column of Table 7.14 list the F-ratio and p values of a series of ANOVA testing for differences between lists within psycholinguistic properties.

There is a number of psycholinguistic predictors that are showing differences for values between lists. This raised number of variables that are different across lists

Table 7.14

Summary of Psycholinguistic Variable Measures for Sentence Reading Items with F-Ratio and P Values to Signify Differences Between Item Lists

					ANOVA	
Psycholinguistic Variables	Mean	SD	Min	Max	F(2, 177)	p
AoA	5.8	1.9	3.0	11.3	2.07	0.13
Arousal	3.7	0.8	2.7	7.7	3.83	**0.023**
BF Type	27.8	15.9	1.5	63.8	3.78	**0.025**
Concreteness	4.9	0.1	4.7	5.0	3.07	**0.049**
Consistency	0.8	0.3	0.0	1.0	0.81	0.445
Dominance	5.4	0.7	3.6	6.6	2.58	0.079
Imageability	6.1	0.6	4.1	6.8	4.96	**0.008**
LPC	1.2	0.2	0.9	1.8	2.86	0.06
Phonemes	3.6	0.8	2.0	6.0	0.08	0.922
Neighbourhood size	8.6	6.3	1.0	24.0	0.28	0.753
Semantic diversity	1.5	0.3	0.7	2.1	4.66	**0.011**
Sensory experience	3.3	0.7	1.9	5.2	0.62	0.54
Word frequency	4.0	0.7	3.1	4.9	0.18	0.833
Word meanings	5.8	3.6	1.0	15.0	6.19	**0.003**
Valence	5.4	0.9	3.2	7.1	3.5	**0.032**

Note:
AoA = Age of acquisition. BF = Bigram frequency. LPC = Levenshtein Phonological Consistency.

could introduce results that confound our interpretation of model effects. Consequently, results must be interpreted with caution. The distribution of different psycholinguistic properties of items across three lists are displayed in Figure 7.33.

7.4.2 Analyses

7.4.2.1 *Number of Observations*

Full Sample. We collected 34,920 observations in the sentence reading task. We excluded 360 observations from six participants due to them being duplicated items from previous waves of data collection. We excluded a further 40 observations from one participant due to technical issues. We also excluded 53 observations that were <200ms and two observations that were recorded as above 4,000 ms, all were excluded as mis-trials due to technical malfunctions. This left 34,465 observations correct and incorrect word trials were available for accuracy analyses. We further removed incorrect trials (n = 1,079) to leave 33,386 observations of correct trials for reaction time analyses.

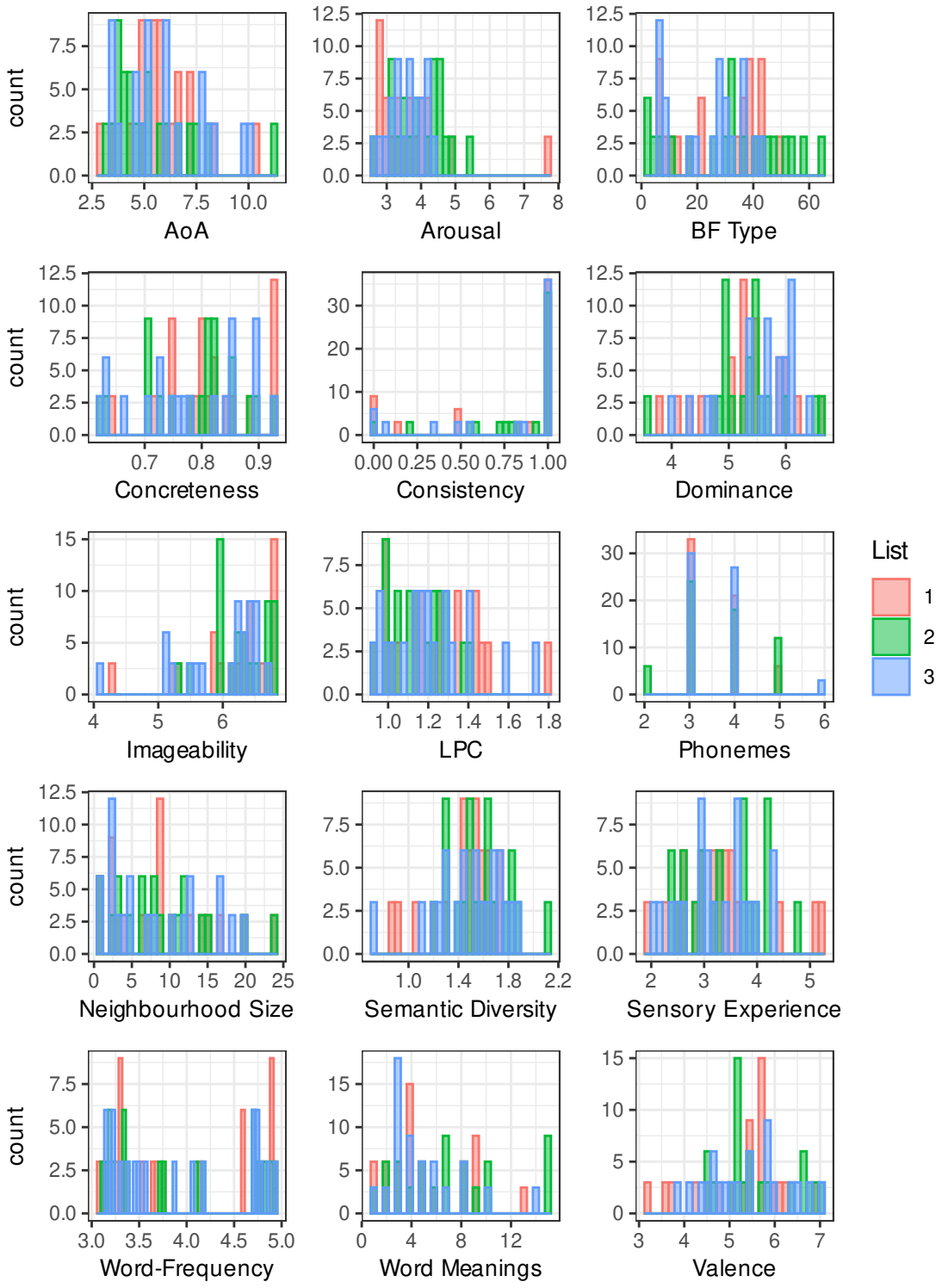
Complete Case Analysis. There were 169 participants who completed the sentence reading task at three data collection sessions. This gave 29,866 observations for a complete case analysis. The preferred model was run on this dataset for accuracy. Accuracy was at 97.7% for this dataset. For correct trials in reaction time data there were 29,201 observations.

Outlier Analysis. We removed 233 trials (0.7%) that were at the time-out value before calculating the outliers for each participant, removing them (n = 4,039, 11.7%) leaving 30,193 observations by which to re-run the preferred model for accuracy data. Accuracy was at 98.3% for this dataset. We re-ran the preferred models on correct trials for reaction time (n = 29,675).

In the complete dataset with no outliers, there were 26,464 observations for

Figure 7.33

Histograms Showing the Distribution of Psycholinguistic Properties of Items in the Sentence Reading Task Across Three Lists



the accuracy model. Accuracy was at 98.4% for this dataset. For analysis of reaction time outcomes on correct trials, there were 26,037 observations.

7.4.3 Accuracy Results

7.4.3.1 *Descriptive Statistics*

We calculated mean accuracy performance per condition per participant and display the distributions within groups across time in Figure 7.34; averages across accuracy and reaction time by time, condition and group are displayed in Figure 7.35. Total mean accuracy performance was above 95% (isolation = 98.5%; meaningful = 99.4%; neutral = 99.0%). This ceiling level means that any effects are likely to be extremely small and that estimation of those effects may be highly unstable. We must exercise caution in the interpretation of the preferred model.

7.4.3.2 *Preferred Model*

The preferred model for the sentence reading accuracy data was the Additive-RIS, the model with predictors for time, ID and psycholinguistic variables, with random intercepts and slopes on participants and items. This model explained 28.6% of the variance in the accuracy outcome (R^2_{bayes} = 28.6% [26.2, 31.0]). Figure 7.36 and Table 7.15 display the fixed effect coefficients for the model.

Model Inference. The intercept coefficient of the preferred model shows a log-odds value of 6.07, which transforms to a probability accuracy rate of 99.8% when all predictors are at their mean. With accuracy above 95%, we have to interpret the coefficients with caution. Although the model returns some coefficients whose credible intervals lie far from zero, the largest coefficient translates to an increase in the probability of being accurate of only 0.001% because accuracy in the model at the ceiling level.

Figure 7.34

Histograms Showing the Distribution of Mean Accuracy Rates per Participant By Group, Condition and Time in the Sentence Reading Task

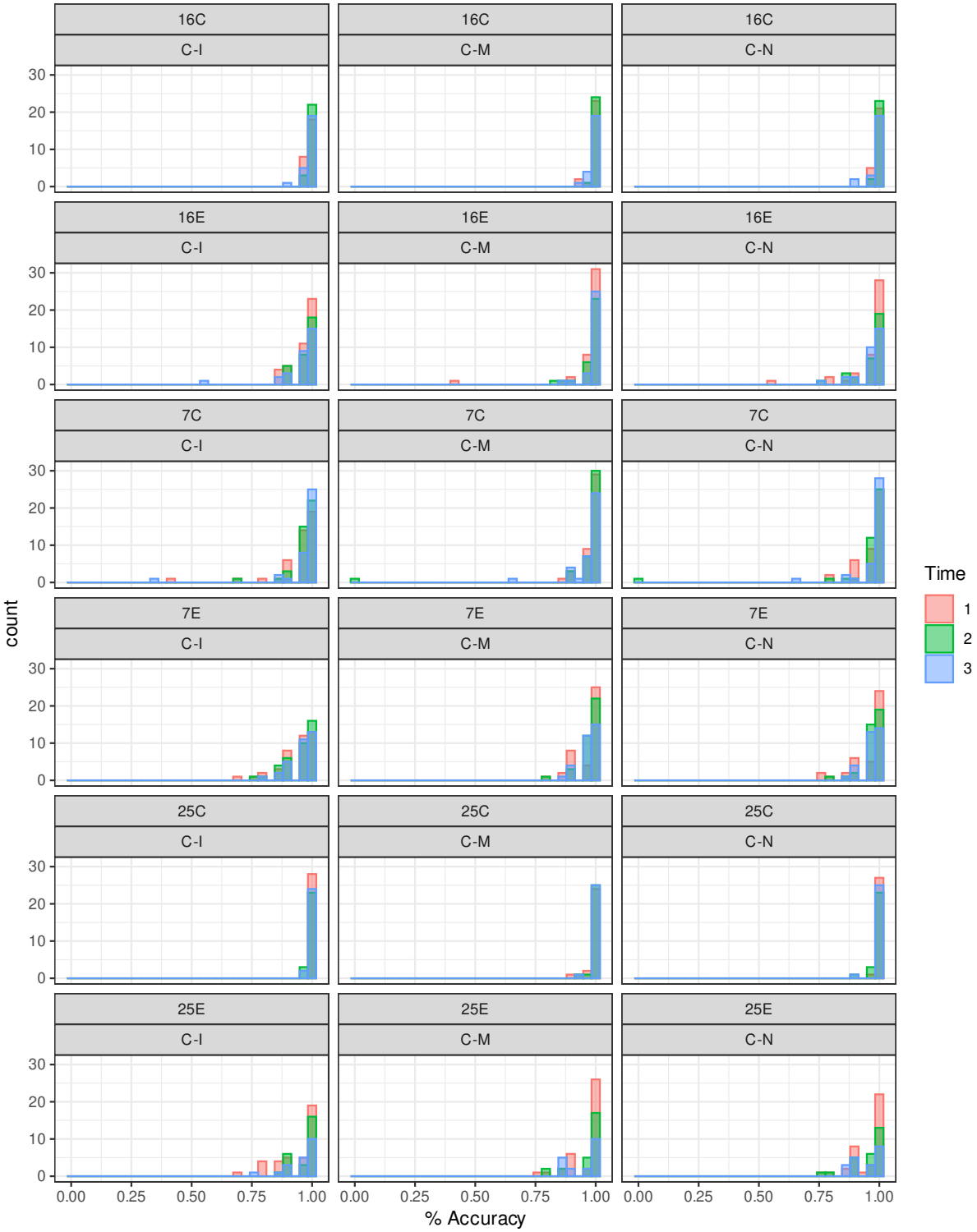


Figure 7.35

Accuracy and Raw Mean RT for Words By Group, Time and Condition for Sentence Reading

	Isolation			Meaningful			Neutral		
Time	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD	Acc %	RT (ms)	SD
Typically-reading 16-17									
1	98.5	680.9	205.7	99.4	621.4	181.7	99.0	676.6	219.2
2	99.4	706.0	209.1	99.8	703.4	205.2	99.6	765.0	216.7
3	98.6	714.7	218.7	98.9	706.6	247.8	98.5	792.6	256.2
Atypically-reading 16-17									
1	96.2	714.8	260.8	96.9	647.8	226.4	96.0	691.3	219.3
2	97.1	740.1	254.4	97.8	723.9	328.1	96.1	756.9	293.1
3	95.0	769.9	301.4	98.7	741.8	280.3	95.8	816.0	329.4
Typically-reading 11-12									
1	94.1	763.6	305.5	97.8	713.9	226.4	96.5	777.9	286.2
2	96.4	754.2	264.7	96.0	765.2	290.8	95.0	832.0	292.6
3	96.1	779.7	288.7	96.9	749.4	283.8	97.3	842.2	325.1
Atypically-reading 11-12									
1	93.0	792.3	282.2	96.2	786.3	257.9	95.4	817.6	274.8
2	94.3	820.3	354.6	97.1	808.6	290.4	96.6	858.5	337.6
3	95.2	850.2	350.4	96.4	822.1	296.8	96.2	878.5	329.3
Typically-reading Adult									
1	100.0	671.7	180.7	99.1	586.9	143.7	99.8	644.7	151.8
2	99.4	691.2	196.7	99.6	606.8	163.5	99.1	697.7	196.2
3	99.6	677.5	209.4	99.8	613.3	155.6	99.6	699.0	186.4
Atypically-reading Adult									
1	93.6	789.4	298.0	96.3	705.0	296.8	95.5	771.1	316.8
2	96.5	793.6	348.4	96.3	742.0	303.0	95.2	796.9	270.2
3	95.2	789.1	301.2	94.5	740.9	257.9	94.2	812.9	244.0

Note:
Acc % = Percentage Accuracy; RT = Reaction time in milliseconds

Figure 7.36

Estimates from the Posterior Distribution of the Preferred Model for Time, Sentence Context, Phonemic Onsets, ID and Psycholingustic Predictors on the Sentence Reading Accuracy Data

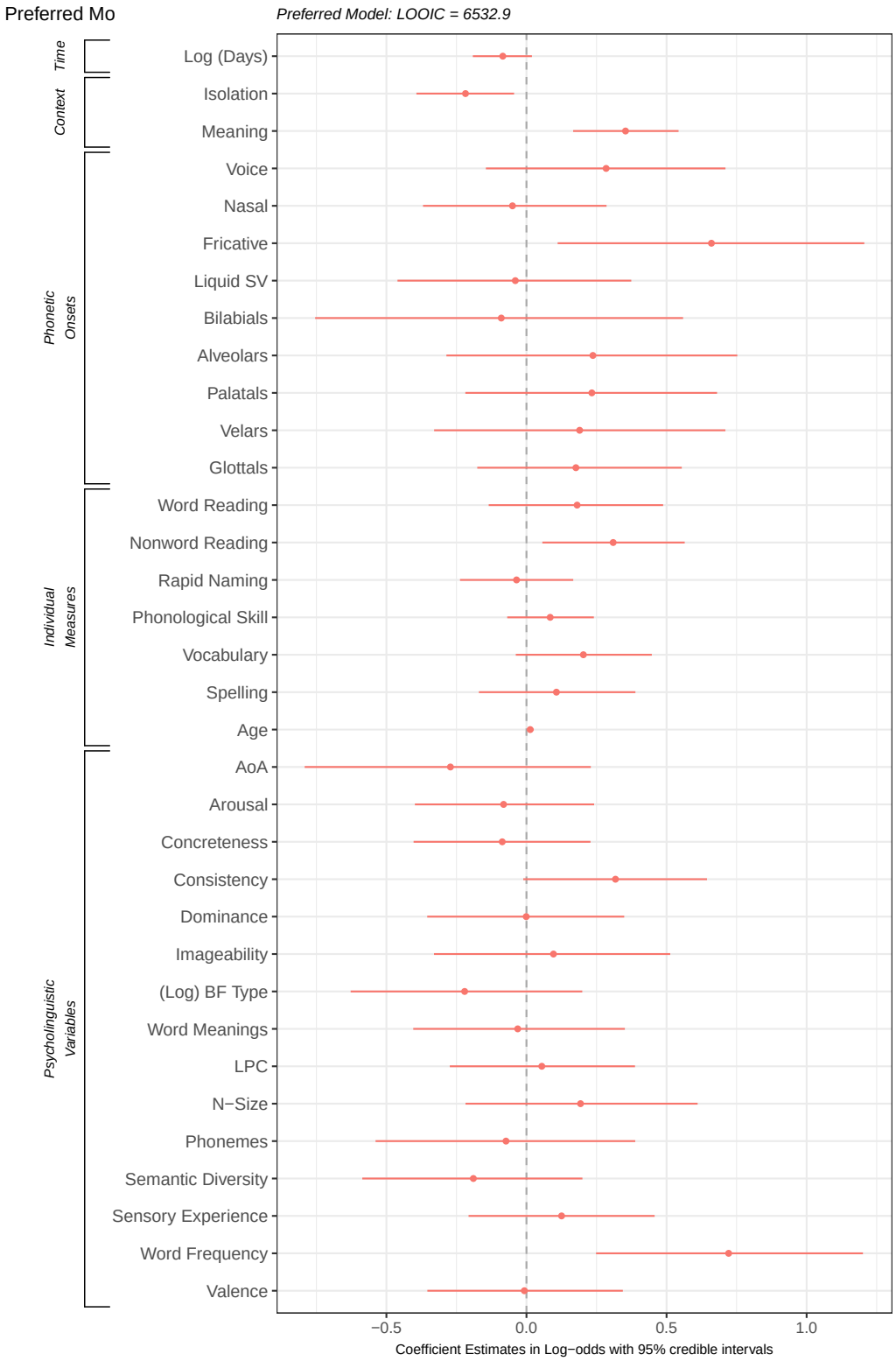


Table 7.15

Summary of Standardised Fixed Effects for Sentence Reading Accuracy

Term	Estimate	SE	Lower CI	Upper CI
Intercept	6.07	0.27	5.57	6.61
Time				
(Log) Days	-0.08	0.05	-0.19	0.02
Reading Context				
Isolated context	-0.22	0.09	-0.39	-0.04
Meaningful context	0.35	0.10	0.17	0.54
Phonemic Onsets				
Voice	0.28	0.22	-0.14	0.71
Nasal	-0.05	0.17	-0.37	0.29
Fricative	0.66	0.28	0.11	1.21
Liquid_SV	-0.04	0.21	-0.46	0.37
Bilabials	-0.09	0.33	-0.75	0.56
Alveolars	0.24	0.26	-0.29	0.75
Palatals	0.23	0.23	-0.22	0.68
Velars	0.19	0.26	-0.33	0.71
Glottals	0.18	0.19	-0.18	0.55
Individual Differences				
Word reading	0.18	0.16	-0.13	0.49
Nonword reading	0.31	0.13	0.06	0.56
Rapid naming	-0.04	0.10	-0.24	0.17
Phonological skill	0.08	0.08	-0.07	0.24
Vocabulary	0.20	0.12	-0.04	0.45
Spelling	0.11	0.14	-0.17	0.39
Age	0.01	0.01	0.00	0.03
Psycholinguistic Variables				
AoA	-0.27	0.26	-0.79	0.23
Arousal	-0.08	0.16	-0.40	0.24
Concreteness	-0.09	0.16	-0.40	0.23
Consistency	0.32	0.17	-0.01	0.64
Dominance	0.00	0.18	-0.35	0.35
Imageability	0.10	0.21	-0.33	0.51
(Log) BF Type	-0.22	0.21	-0.63	0.20
Word meanings	-0.03	0.19	-0.40	0.35
LPC	0.05	0.17	-0.27	0.39
N-Size	0.19	0.21	-0.22	0.61
Phonemes	-0.07	0.24	-0.54	0.39
Semantic diversity	-0.19	0.20	-0.59	0.20
Sensory experience	0.13	0.17	-0.21	0.46
Word-frequency	0.72	0.24	0.25	1.20
Valence	-0.01	0.18	-0.35	0.34

Note:
CI = Credible intervals. AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency. N-Size = Neighbourhood size

The coefficients for the context predictor levels were reliable. Reading a word in a meaningful context gives higher odds of an accurate response (log-odds = 0.35 [0.17, 0.54]) relative to a neutral context, increasing the probability of an accurate response by 0.0007%. As you can see, in real terms, the effects sizes are infinitesimal. Reading a word in isolation gives lower odds of an accurate pronunciation compared to the neutral context of reading (log-odds = -0.22 [-0.39, -0.04]), decreasing the probability of accuracy by approximately 0.0006%.

We predicted that vocabulary would be a reliable predictor in the sentence task. It was not (log-odds = 0.20 [-0.04, 0.45])). Only nonword reading and age variables have credible intervals that do not include zero. The nonword reading coefficient is positively related to accuracy (log-odds = 0.31 [0.06, 0.56]), suggesting that people of higher nonword reading skill have higher odds of giving an accurate response. Incredibly small but present is a positive coefficient for age (log-odds = 0.01 [0.00, 0.03]). Older participants are more likely to be accurate than younger participants, to an extremely small degree. None of these effects move the accuracy rate from 99.8 to 99.9 they are that small.

Word-frequency has credible intervals that do not include zero (log-odds = 0.72 [0.25, 1.20]) and is a big enough effect to register a change in accuracy rate from the intercept term. For a 1 SD increase in frequency values, the odds of being accurate increase by 0.1%. In terms of exploratory effects, consistency has a log-odds of 0.32 [-0.01, 0.64] with a lower credible interval that has just crossed zero. It is worthy of consideration in this instance, given that the predictor is reliable for word naming for this sample, but it is just as tiny as the ID measures coefficients.

Complete Case and Outlier Analyses. The full sample with no outliers model shows a strengthening of the effects for the context condition. The isolation condition moves from log-odds = -0.22 [-0.39, -0.04] to log-odds = -0.61 [-0.87, -0.35], almost three times the size of the full sample data set model. Vocabulary remains an unreliable predictor. Nonword reading and age remain reliable. Age remains the same size but nonword reading becomes a stronger effect log-odds = 0.40 [0.10, 0.70].

Word-frequency doubles in size in this data set log-odds = 1.24 [0.50, 1.94]. The model is very certain that consistency is not a reliable effect in the full sample with no outliers data set.

The negative effect attributed to accuracy over time is attenuated in the complete case and outlier analyses. In each of these models, the model implied coefficient for log (days) lies on zero. The effects for word reading in isolation or meaningful sentences remain the same.

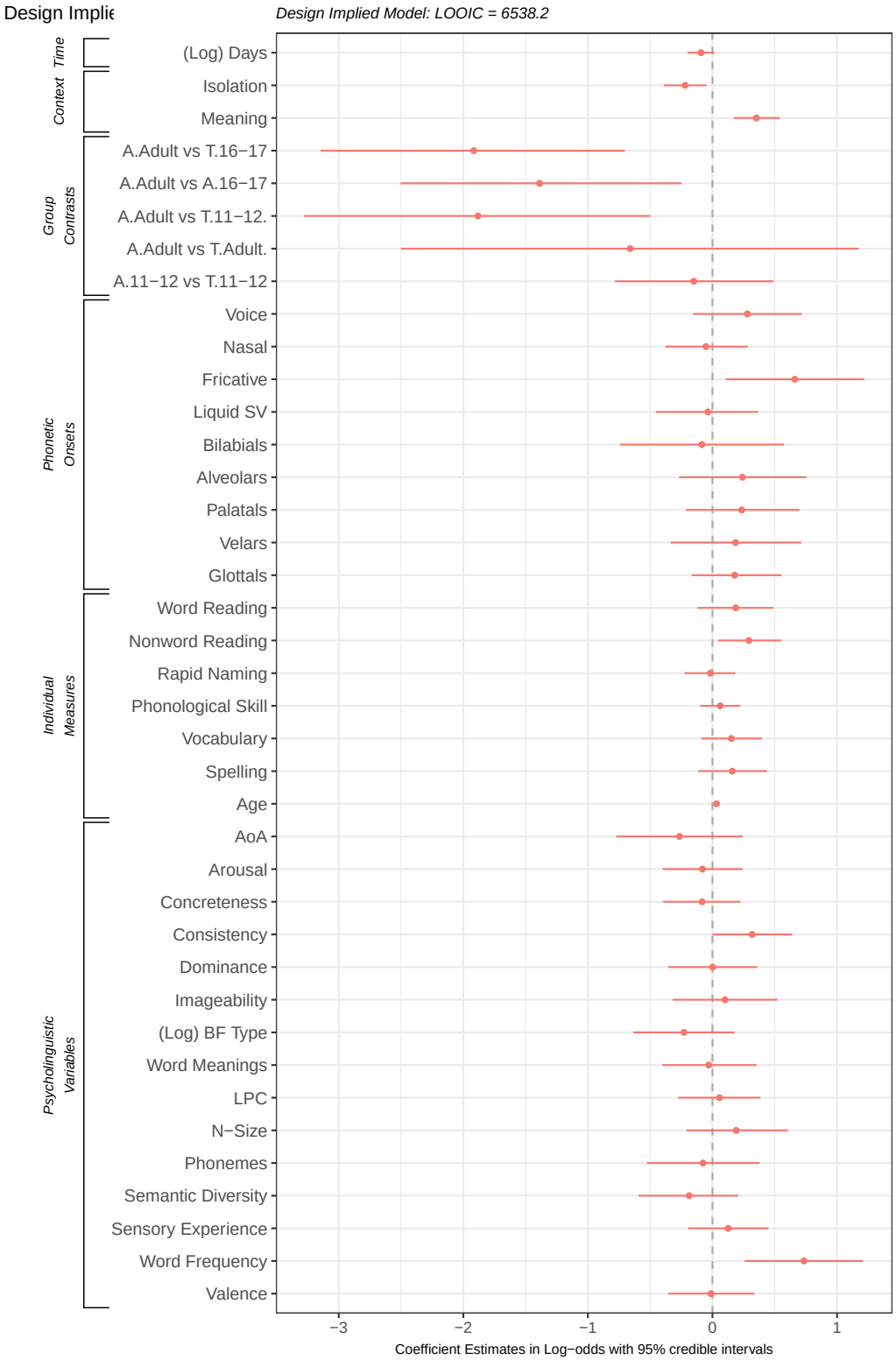
Nonword reading remains the same across the models and vocabulary becomes influential also, showing an extremely small, positive coefficient with a credible interval whose lower bound lies on zero. The effects across the psycholinguistic predictors remain the same as in the full sample model.

Design Implied Model. The design implied model is in Figure 7.37. As with the earlier models, the difference here is only the addition of the group contrast predictor. The group predictor does show coefficients that predict differences in accuracy rates between the atypically-reading adults and other groups of readers. The model implied trend is that atypically-reading adults are less likely to accurately pronounce a word than either the typically-reading 16-17-year-olds (log-odds = -1.92 [-3.15, -0.70]), atypically-reading 16-17-year-olds (log-odds = -1.39 [-2.50, -0.25]), typically-reading 11-12-year-olds (log-odds = -1.88 [-3.28, -0.50]) and the typically-reading adults (log-odds = -0.66 [-2.50, 1.17]). The credible interval for the contrast between atypically- and typically reading adults includes zero which implies that the opposite effect is also plausible, with atypically-reading adults reading more accurately than their typical reading peers. Each of these effects is of a size worthy of notice in that they are greater than -0.5. The estimates for nonword reading, age and word-frequency are the same.

Model Predictions. Figure 7.38 shows the model predictions for the sentence reading accuracy outcome. We draw the plots here with a truncated range on the y-axis because the effects are so small. In contrast to earlier predicted effects plots

Figure 7.37

Estimates from the Posterior Distribution of the Design Implied Model for ID and Psycholingustic Predictors on Sentence Reading Accuracy Data



where the y-axis begins at 0, here the y-axis begins at 0.7 to allow some resolution of effects to be displayed.

The ceiling levels of accuracy means that predictions show little variation across the range of low to high skills for RON, phonological skill, vocabulary and spelling. There is some decrease in accuracy for predictions of low word reading and nonword reading skill. The plot visualises that the lower bound of the 95% interval suggests that accuracy could fall as low as 97% and 95% for individuals at the lowest values of the simulated range. The lower bound of the 95% interval on word-frequency estimates an accuracy value of between 80% - 85% for the lowest frequency words.

7.4.3.3 Summary and Discussion.

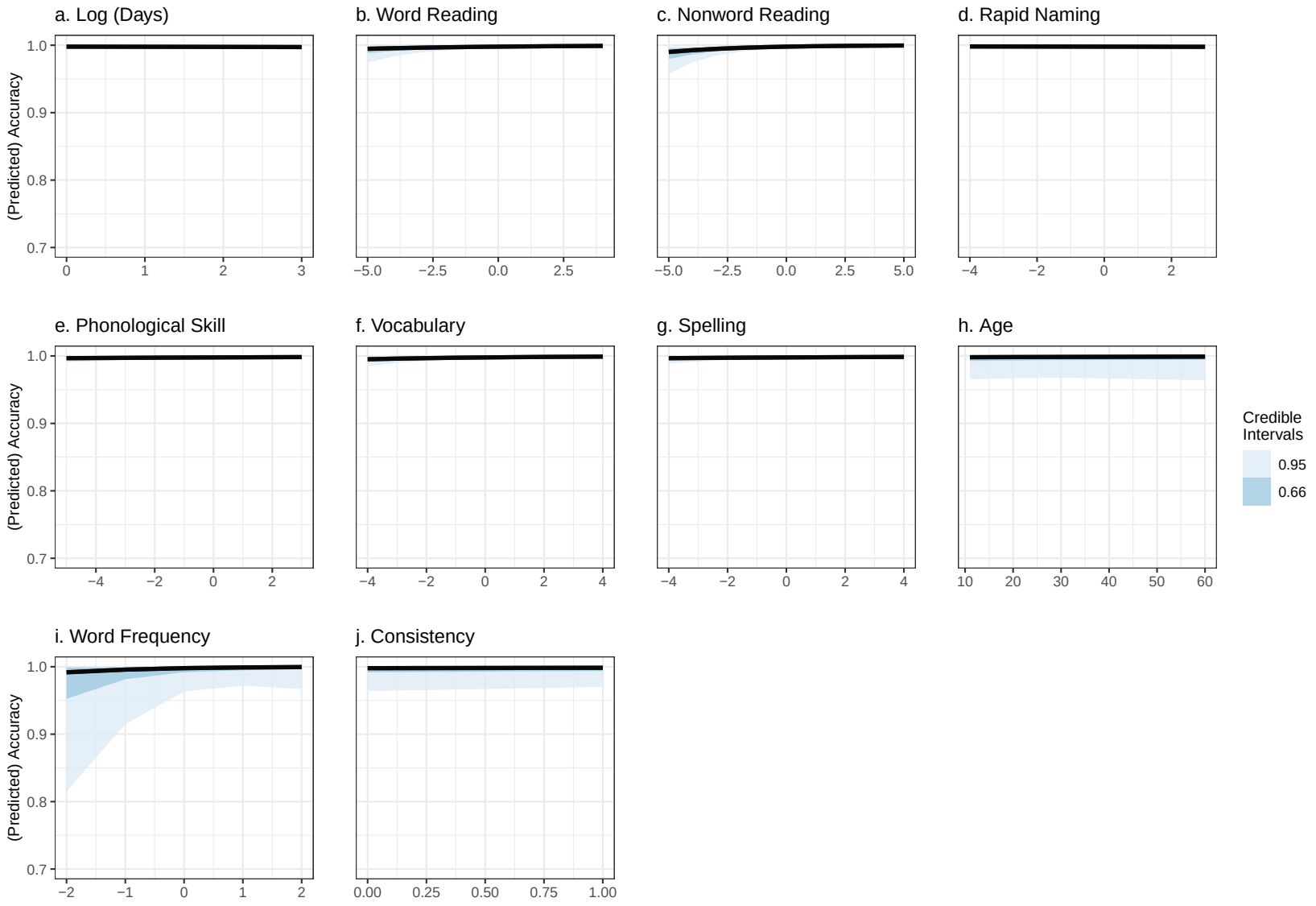
We asked participants to name single words under three levels of sentence reading context that provided different levels of information for word recognition. The preferred model did not include the group contrast predictor nor interactions, suggesting that for this model and sample, groups approached the task similarly, using the same information and strategy.

The three conditions of reading context differed from each other. A meaningful context reliably facilitates accuracy relative to a neutral sentence context or no sentence context at all, confirming our prediction. This finding aligns with Bruck (1990), Ben-Dror et al. (1991), and Ricketts et al. (2016).

Only nonword reading and frequency were estimated with any certainty. People of higher nonword reading skill are likely to be more accurate than people of lower nonword reading skill. High frequency words are likely to be pronounced at a higher rate of accuracy than lower frequency words. As in earlier models, consistency's lower bound of its credible level rested just on zero. Vocabulary comes online as a reliable predictor in the sensitivity analyses. Future design of this task should take this into consideration.

Figure 7.38

Preferred Model Predictions for the Effects of Individual Differences and Word-Frequency on Sentence Reading Accuracy Performance



Nonword reading has a very small effect in this model. It is the only predictor amongst the range of ID measures to not include zero inside its credible intervals. For words that are well known and for skilled readers, the lexical quality hypothesis and computational models of reading both suggest that the lexical / orthographic form and route for words takes prominence over phonological and semantic forms / routes as the word becomes better known and reader skill increases. If orthographical information were sufficient, then we might expect word reading to show greater influence. The very high accuracy rates within task suggest that with the added information of sentence context, the words are known. Consequently, a nonword reading skill effect and an absence of a word reading skill effect is surprising.

Interpretation of this model is limited because of ceiling levels of accuracy in the data. While the relative redundancy of most of the predictors mirrors that of lexical decision and word naming models, greater variation in the outcome variable needs to be present to have greater trust in these results. This suggests either a higher level of challenge in the items or a higher level of challenge in the reading context is mandated. It is also possible that the small number of items (20 items repeated three times), compared to other tasks in the study (150 individual items) reduced the variation to further reduce information available for modeling.

7.4.4 Reaction Time Results

7.4.4.1 *Descriptive Statistics*

Distributions of raw mean reaction time in milliseconds across correct trials per participant are displayed at the group level in Figure 7.39 and Figure 7.35. Reaction time data was log transformed to reduce skew before standardising within condition using the typically-reading 16-17-year-olds as our reference level (Isolation mean = 680.9 ms, SD = 205.7; Meaningful mean = 621.4 ms, SD = 181.6; Neutral = 676.6 ms, SD = 219.2). Consequently, positive coefficients indicate slower reaction times and negative coefficients indicate faster reaction times than the mean of an average

typically-reading 16-17-year-old.

7.4.4.2 *Preferred Model*

The preferred model for reaction time data in the sentence reading task is the Additive-RIS model with predictors for the effects of time, phonetic onsets, ID and psycholinguistic variables, random intercepts and slopes for participants and items. The explained variance in the reaction time outcome was $R^2_{\text{bayes}} = 38.8\%$ [38.2, 39.4]. The plot for fixed effects of the model is presented in Figure 7.40. A summary of the coefficients for the fixed effects can be seen in Table 7.16.

Model Inference. Log (days) is positively associated with reaction time, suggesting that responses slowed, on average, across data collection sessions ($\beta = 0.08$ [0.06, 0.10]). Every 95 days, reaction time increased on average by ~ 16 ms.

Both isolation β = -0.21 [-0.23, -0.19] and meaningful contexts β = -0.10 [-0.12, -0.08] are reliably faster than the neutral context for reading. The isolation context is on average 42 ms faster than the neutral context, while the meaningful context's advantage is estimated as approximately 20 ms. Taken together, this gives and estimated difference of approximately 22 ms between the isolation and the meaningful context for reaction time.

The control variables for phonetic onsets with credible intervals that exclude zero show a slowing of reaction time for voiced onsets and faster responses for onsets in the nasal, fricative, palatal and glottal position.

The model is confident for the sign and direction for word reading skill (β = -0.06 [-0.13, 0.00]), RON (β = -0.05 [-0.11, 0.00]) and spelling (β = -0.05 [-0.11, 0.00]). For a 1 SD increase in word reading skill, a person would be approximately 12 ms faster. For RON and spelling skill, the gain is 10 ms. These are very small effects which indicate that people of higher skill in these tasks are faster in pronouncing the target word across all conditions of reading contexts. In absolute terms, the data are inconclusive about the direction of effect for the remaining ID measures for this model

Figure 7.39

Histograms Showing the Distribution of Raw, Mean Reaction Time (ms) By Participant, Group and Condition Across Time in the Sentence Reading Task

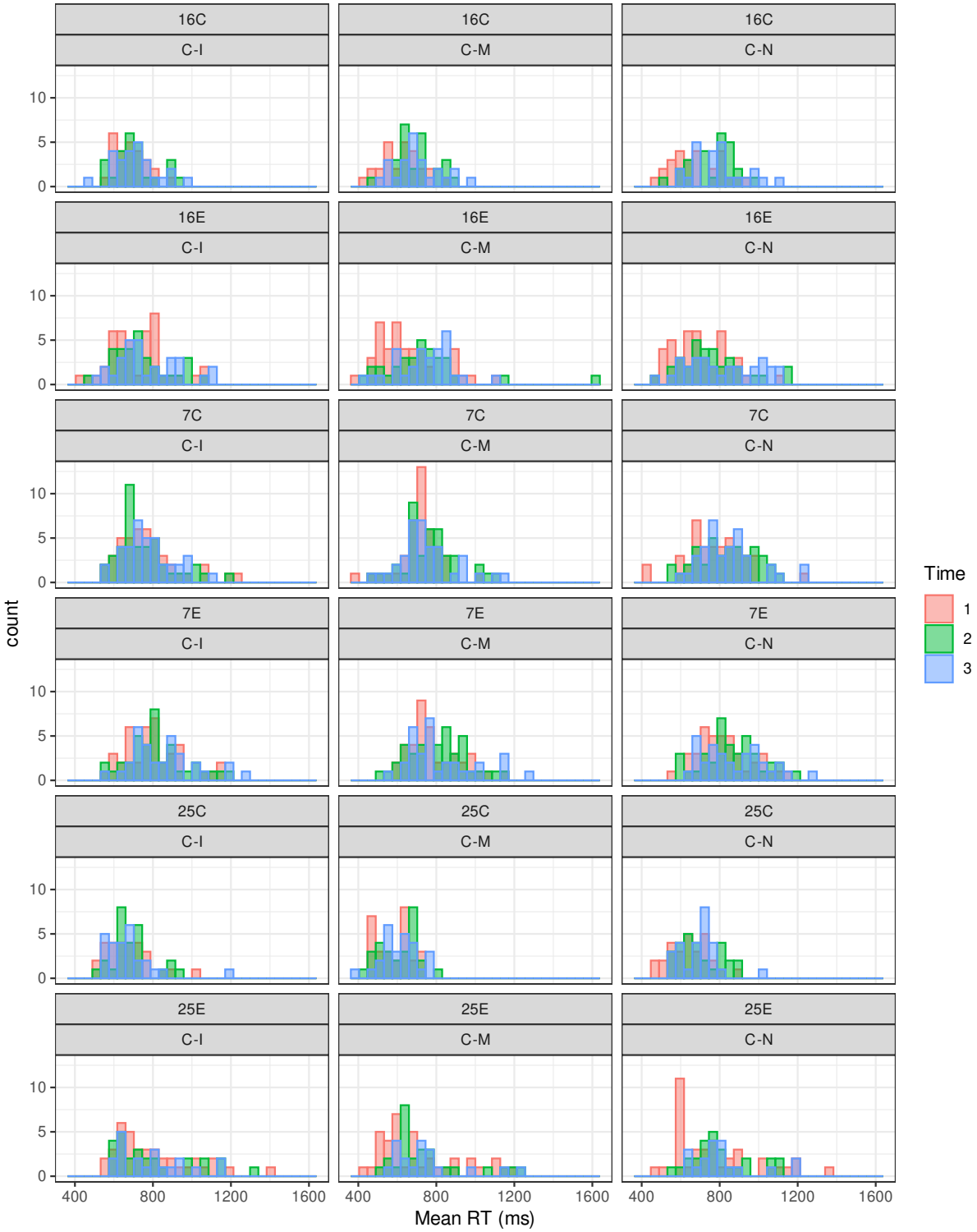


Figure 7.40

Estimates from the Posterior Distribution of the Preferred Model for Time, Sentence Context, Phonemic Onsets, ID and Psycholingustic Predictors on the Sentence Reading Reaction Time Data

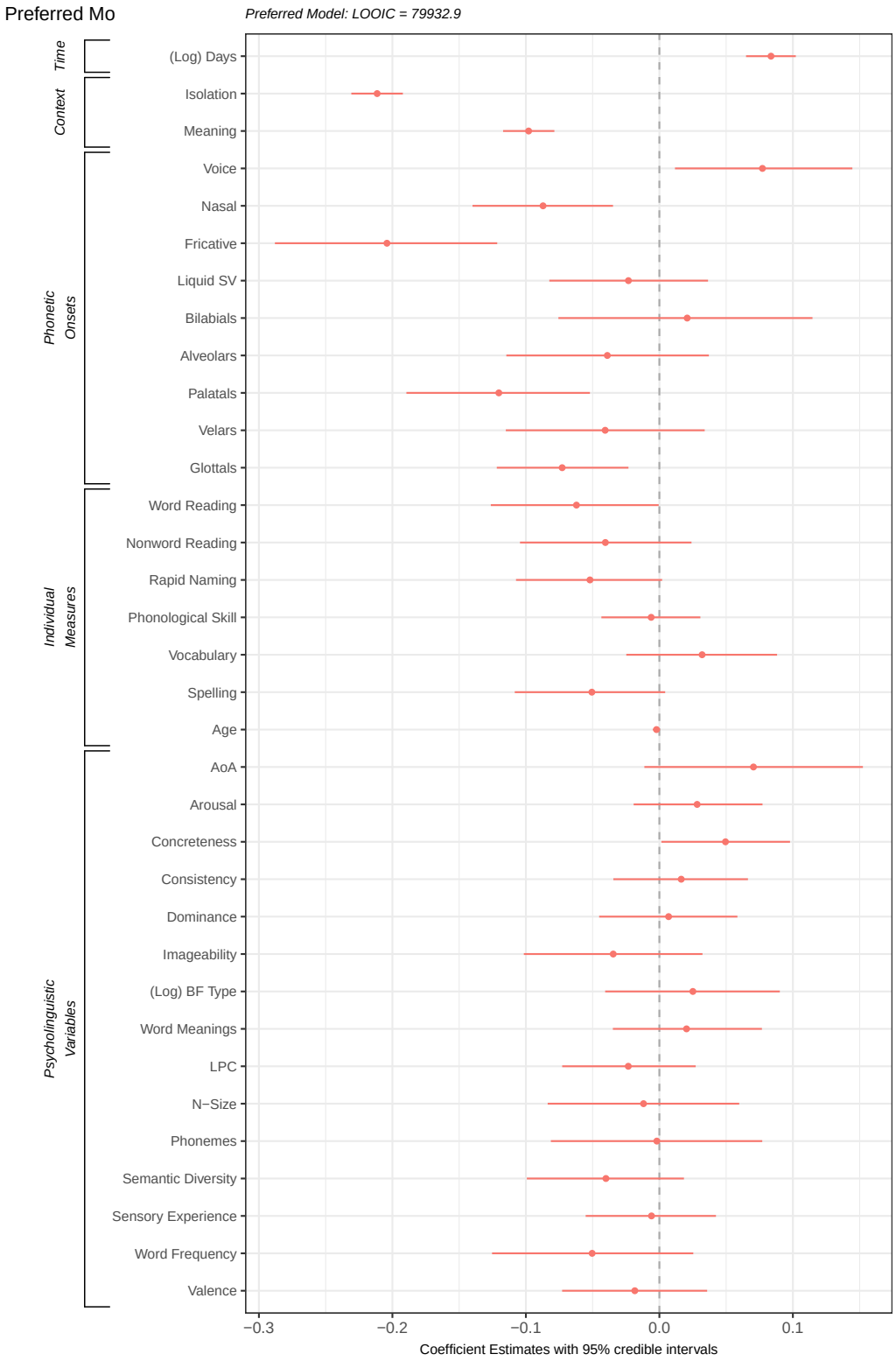


Table 7.16

Summary of Standardised Fixed Effects for Sentence Reading Reaction Time

Term	Estimate	SE	Lower CI	Upper CI
Intercept	0.34	0.05	0.23	0.45
Phonemic Onsets				
Voice	0.08	0.03	0.01	0.14
Nasal	-0.09	0.03	-0.14	-0.03
Fricative	-0.20	0.04	-0.29	-0.12
Liquid_SV	-0.02	0.03	-0.08	0.04
Bilabials	0.02	0.05	-0.08	0.11
Alveolars	-0.04	0.04	-0.11	0.04
Palatals	-0.12	0.04	-0.19	-0.05
Velars	-0.04	0.04	-0.12	0.03
Glottals	-0.07	0.02	-0.12	-0.02
Reading Context				
Isolated context	-0.21	0.01	-0.23	-0.19
Meaningful context	-0.10	0.01	-0.12	-0.08
Time				
(Log) Days	0.08	0.01	0.06	0.10
Individual Differences				
Word reading	-0.06	0.03	-0.13	0.00
Nonword reading	-0.04	0.03	-0.10	0.02
Rapid naming	-0.05	0.03	-0.11	0.00
Phonological skill	-0.01	0.02	-0.04	0.03
Vocabulary	0.03	0.03	-0.02	0.09
Spelling	-0.05	0.03	-0.11	0.00
Age	0.00	0.00	0.00	0.00
Psycholinguistic Variables				
AoA	0.07	0.04	-0.01	0.15
Arousal	0.03	0.02	-0.02	0.08
Concreteness	0.05	0.02	0.00	0.10
Consistency	0.02	0.03	-0.03	0.07
Dominance	0.01	0.03	-0.05	0.06
Imageability	-0.03	0.03	-0.10	0.03
(Log) BF Type	0.03	0.03	-0.04	0.09
Word meanings	0.02	0.03	-0.03	0.08
LPC	-0.02	0.03	-0.07	0.03
N-size	-0.01	0.04	-0.08	0.06
Phonemes	0.00	0.04	-0.08	0.08
Semantic diversity	-0.04	0.03	-0.10	0.02
Sensory experience	-0.01	0.02	-0.06	0.04
Word-frequency	-0.05	0.04	-0.13	0.03
Valence	-0.02	0.03	-0.07	0.04

Note:
CI = Credible intervals. AoA = Age of acquisition. (Log) BF Type = Log Bigram Frequency Type. LPC = Levenshtein Phonological Consistency. N-Size = Neighbourhood size

and this sample.

The sole psycholinguistic predictor to show a marginally conclusive effect is concreteness ($\beta = 0.05$ [0.00, 0.10]). Words of a higher concreteness rating predict slower pronunciation times, equating to an inhibitory effect of approximately 10 ms with each move up the rating scale.

Complete Case and Outlier Analyses. When outliers are removed from the full sample, word reading and spelling remain as stable, negative estimates. The estimates for context and concreteness are also stable. However, there are some changes to the model estimates. AoA also becomes a reliable, positive estimate, suggesting that later learned words are read more slowly than early learned words.

The complete case data set continues to show a slowing of responses over time. The context condition shows the same pattern and strength as the full sample model. Word reading, RON and spelling all become unreliable but concreteness is a stable effect.

When outliers are removed from the complete cases data set, the context condition estimates retain their direction of effect but change size. While the isolation estimate remains stable, the meaningful condition is reduced in size. Word reading, RON and spelling remain unreliable but concreteness is a stable effect. Just as in the full sample with no outliers data, AoA is indicated as a positive, reliable predictor.

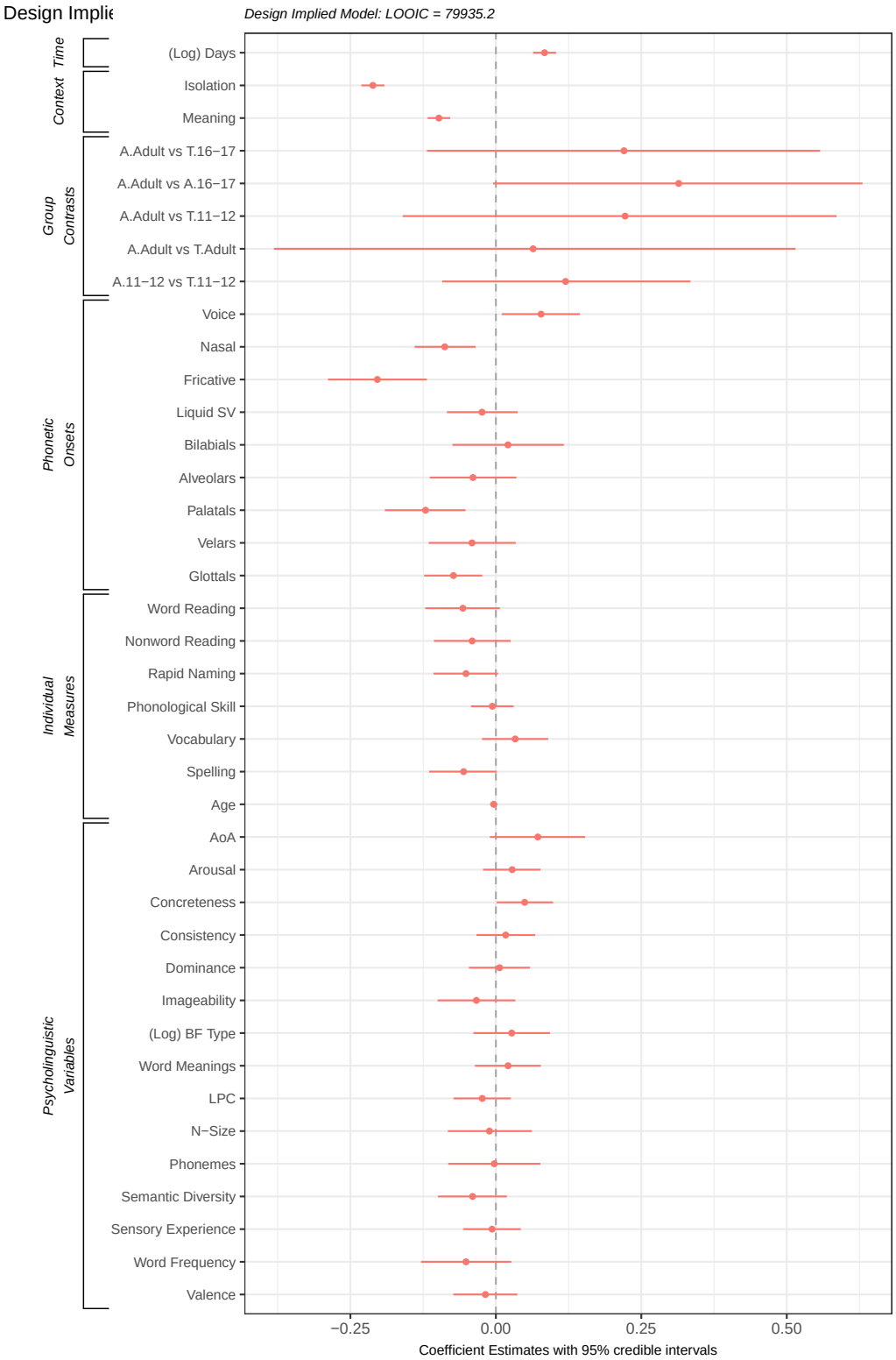
Design Implied Model. The design implied model includes the group contrast predictor. Coefficient estimates for this model are displayed in Figure 7.41.

The estimates are similar in direction and size in the design implied model compared to the preferred model. The upper bound for the credible interval for word reading skill now crosses zero, and so is similar to the other ID measures. Concreteness remains reliably positive.

The group contrasts all show very wide credible intervals. The coefficient for the contrast between atypically-reading adults and atypically-reading 16-17-year-olds has a lower boundary that is resting on zero. The trend in the model is for

Figure 7.41

Estimates from the Posterior Distribution of the Design Implied Model for Sentence Reading Reaction Time Data



atypically-reading adults to be slower than other groups (16-17-year-olds, typically-reading 11-12-year-old and typically-reading adults), however credible intervals also suggest that the opposite effect is plausible, with each estimate interval including zero.

The model shows the same uncertainty around the contrast between atypically- and typically-reading 11-12-year-olds, however the trend is the same with atypically-reading 11-12-year-olds tending to be slower in responding, to a very small degree, than their peers.

Model Predictions. We show the model implied predictions Figure 7.42. The predictions follow the model estimates. Reading and nonword reading skill, RON and spelling all show negative relationships while higher vocabulary knowledge is implicated as an inhibitory effect, as are words of high concreteness ratings.

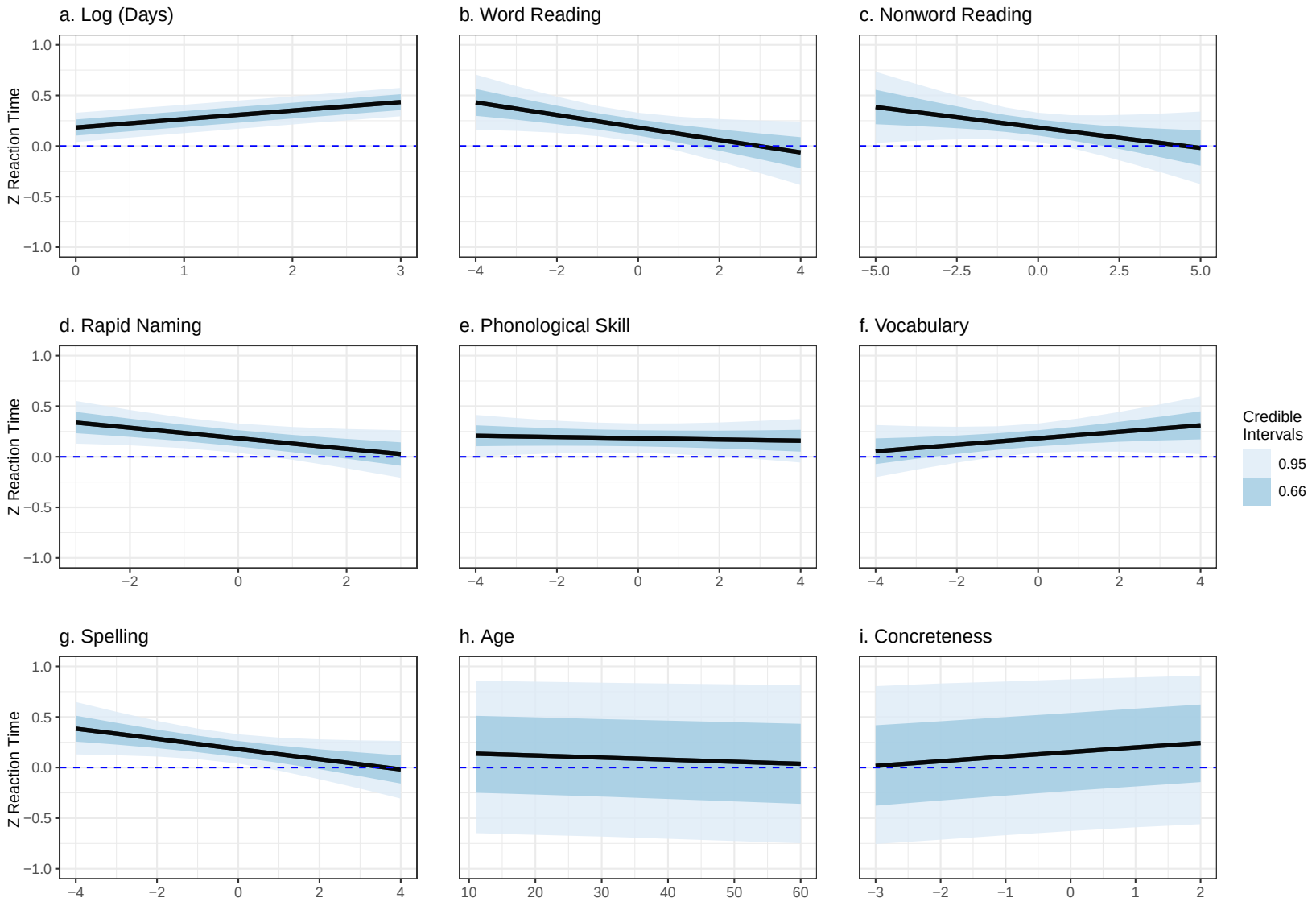
Summary & Discussion. We expected the meaningful condition to facilitate both accuracy and reading speed relative to the neutral condition, which it did, reliably. It was the most accurate condition and the second fastest for reading. We were unsure of how the isolation condition would place - the models estimate a lack of context to be less accurate but the fastest condition amongst the three.

We also expected vocabulary to show a positive relationship with sentence reading accuracy data and facilitate reaction times. In the full sample model for accuracy, it was unreliable but in the sensitivity analyses, vocabulary became reliable. We believe this may be because accuracy rates were at ceiling and so there was no need for assistance from other sources outside of the items themselves. Constructing a more challenging item sample may test this assumption.

Nor did vocabulary facilitate reaction time responses. The model estimated its effect as unreliable and model predictions imply a positive effect, slowing reaction time for people with stronger vocabulary knowledge. We interpret this as a competitive effect for those people with stronger vocabulary knowledge having more candidate words from which to choose the correct response (Ramscar et al., 2013).

Figure 7.42

Preferred Model Predictions for the Effects of Individual Differences and Concreteness on Sentence Reading Reaction Time Performance



Both preferred models did not include the group contrast predictor, suggesting that on average, this model and this sample describe participants who are approaching the task in a similar way and use similar types of information and skill. Of interest is the repeated occurrence of an estimate for concreteness that is positive, when previous studies tend to find that it shares a negative relationship with reaction time (Kousta et al., 2011).

8 Discussion

The present study explored the single word recognition performance of a group of atypically-reading adults. We compared this group with typically-reading adults, and younger, typical- and atypical-readers. We measured reading-related skills through individual difference (ID) measures to connect to previous studies. We also conducted four experimental tasks – letter search, lexical decision, word naming and sentence reading – to estimate single word reading processes from letter level to sentence level and modelled the impact of psycholinguistic variables on reaction time and accuracy outcome measures. Only one study that we know of has explored the influence of psycholinguistic variables on single word reading processes for a similar group of adults (McKoon and Ratcliff, 2016, – word-frequency), with the preponderance of research focusing on individual difference measures.

Our exploratory research questions were whether atypically-reading adults were similar or dissimilar to the other groups of readers and what the form of those differences may take. We constructed several statistical models to see if independent or interaction effects were better fits for the observed data. Crucially, we used mixed-effects models to account for the within-participant-between-item and within-item-between-participant dependencies in the data due to the crossed effects and longitudinal design of the study. We estimated our effects using frequentist and Bayesian Inference (BI) approaches, knowing prior to beginning data analysis that model convergence could be problematic and that BI approaches offer some mitigation of these problems. To construct strongly informative priors for the BI models, we conducted a meta-analysis of a core set of eight predictors. The meta-analysis is a living project, openly accessible for researchers to use for their own research purposes.

Findings from the ID measures suggest that the atypically-reading adults' reading-related skills are broadly in line with atypically-reading 16-17-year-olds and

typically-reading 11-12-year olds. We found no evidence of significant change in ID measures across the three data collection sessions of the study. Differences that were present at T1 tended to remain stable or disappear by T3 due to increases in the other groups' scores. A lack of significant findings for separate clusters in a cluster analysis supports a finding of similarity between groups.

Atypically-reading adults present with relative strengths, in decreasing order, in spelling, vocabulary, word reading and nonword reading. Previous studies suggest that adult-learners may be stronger in their semantic skills (Braze et al., 2007; Greenberg et al., 1997; Mellard et al., 2012b) but at first look, atypically-reading adults in the present sample would appear to be relatively stronger in their orthographic skills, as represented by spelling scores. As an orthographic measure of precision (Andrews et al., 2020), however, spelling skill is a fragmented source of information for word recognition compared to word reading skill. Word and nonword reading skill are at the lowest ranks. Furthermore, they appear to be of equivalent strength. This may be problematic if word recognition needs word reading skill to be stronger in order for orthographic learning to develop (Perfetti and Hart, 2002; Perfetti, 2011; Share, 2004).

All of the preferred models for the experimental tasks were for independent effects. The lack of support for any models that include interaction terms suggests that participants across the sample are approaching the experimental tasks in a similar way and using similar sources of information or skill to complete trials.

Across the tasks, the range of variance explained for accuracy was 18.7% - 41% and for reaction time models was 32.8% - 39.9%%. Every preferred model included random intercepts and slopes for participants and items, rather than only random intercepts. Two conclusions arise from this. The first is that individual differences within and between participants may exist which are not captured by the predictors in the study and second, accounting for the dependencies within outcome measures that exist intra-individuals and items is worthwhile for effects estimation.

8.1 Design Effects

The design of the study mandated that certain predictors needed to be entered into every model. These were the passage of time, age, word-frequency measure and the group contrast predictor. We briefly discuss these below before turning to the model effects across tasks.

The longitudinal design of the study captured a very general effect of slowed responses and reduced levels of accuracy across all tasks with each passing data collection session. Adelman et al. (2014) and Yap et al. (2012) show attenuation in reaction time measures over two sessions amongst skilled readers. In the context of no detectable changes across ID measures, we have interpreted this as a habituation effect. It is worth noting as correlational designs that capture data at one time point only may consider their findings in the context of this repeated finding.

The lack of a general effect of age across time and tasks is surprising. The age range in the study was 11 – 79 years. Age as a group contrast variable is prevalent in the literature, as attested to by the number of included studies that focus on age differences in the meta-analysis. One explanation is that for older typical readers in this sample, reading development may be at an asymptote level of skill such that differences are too small to detect. Additionally, small changes in younger groups, not yet at an asymptote level of reading skill, may be better captured by the ID measures. Future studies could consider that age effects are a proxy measure for ID effects that could be measured more directly.

The recovery of a word-frequency effect for the majority of task outcomes is not surprising. The item sample was manipulated for high and low frequency values. That being said, an interesting finding is the lack of a word-frequency effect for sentence reading reaction time measures. In connected text, the effect of frequency for a speeded response appears to be washed out for all participants[1]. We interpret this

[1]We also found this effect in a pilot study of items with two groups of 11-12-year-olds. Items were sampled from a reading book from their school. Frequency values were collected rather than manipulated. No frequency effect was evident in any of the subsequent models and accuracy was very high. This suggested a manipulation on items for the main study to ensure that items would present

as a demonstration of the priming capacity of the sentential context. The identification and recognition of the target word enjoys an extended time course as the contextual information builds prediction for the target word. The advantage that high frequency words show over low frequency words when read in an isolated context is nullified (Zevin and Balota, 2000). In the absence of sentence level support, the word-frequency effect assisted participants' reaction times for items that were unfamiliar in the usual way.

Another surprising finding is that the observed word-frequency effect for word naming accuracy is larger than that observed for lexical decision. This differs from previous studies (Adelman et al., 2014; Balota et al., 2004; Spieler and Balota, 2000; Yap et al., 2012). Vocabulary effects also show this same reversal of the typical trend, with a larger effect in word naming than lexical decision. This may be symptomatic of both the requirements of each task and the wide range of experience in the sample. Lexical decision does not rely on specific word knowledge to be able to make a correct response. The mixing of words and nonwords promotes a focus upon sublexical structures where orthographical knowledge is the stronger source of information and is shared between all items. Information boundaries between items are less clear. Consequently, the influence of word-frequency and vocabulary as lexical level variables is diminished.

In contrast, word naming requires a specific item for a correct response. People of higher vocabulary skill are more likely to know a greater range of words and words of higher frequency are more likely to be known by a greater range of people. Both assist each participant in selecting the specific item for a correct response. Consequently, word-frequency and vocabulary show larger effects.

Perfetti (2007) suggests that for persons of lower vocabulary knowledge, words of objective high frequency are experienced as medium frequency, with low frequency words becoming very low frequency. Kuperman and Van Dyke (2013) support this observation. They critically reviewed the objectivity of word frequency values and their interaction with varying levels of reading experience and found that

---

a challenge.

objective measures of frequency do not explain word recognition behavior so well as subjective word familiarity values. In the present sample, the word-frequency effect for the same item could therefore work differently across two people who differ in vocabulary. While high frequency words likely behave similarly across participants, the differences will begin to be observed lower down the distribution, with objectively "low" frequency valued items showing the largest amount of disparity amongst participants. The word-frequency effect is likely amplified in these conditions because of the amount of "very low" frequency words for some participants, extending the lower range of frequency.

In word naming, although accuracy rates were very high, the presence of the group contrast predictor suggests that specific items were unfamiliar to the participants and that this was true to a greater extent for the atypically-reading adults than other groups. This may be the differential impact of word frequency in operation for atypically-reading adults. In lexical decision, half of the items are nonwords and are unfamiliar to all participants. This disperses the spread of items that are unfamiliar across the groups making them more alike in their knowledge for the task, attenuating effect sizes.

The group contrast predictor was not supported in any of the reaction time models. We conclude that the range of reaction times observed for the atypically-reading adult group, across tasks, was similar to other groups. Any differences in accuracy performance between groups can then be viewed as occurring within a similar time frame. The group contrast predictor was included in the preferred models for letter search and word naming accuracy. Atypically-reading adults show unreliable higher odds of being correct on a letter search task and reliable lower odds of being correct in a word naming task.

This cross-task difference in performance is intriguing. Intuitively, the similarity in word and nonword reading scores as observed in the ID measures would seem to suggest that atypically-reading adults' performance in both tasks would echo that of the younger readers. However, atypically-reading adults are weaker still than all groups in the word naming accuracy performance. Contrast this with the trend for

higher accuracy in the letter recognition task.

We have already suggested above that observed high frequency values may behave as lower frequency values in the context of low vocabulary knowledge. The atypically-reading adults do have low vocabulary compared to the younger readers, as shown in Figure 6.3. This may explain the reliable group differences predicted by the word naming accuracy model. Yet this does not explain the superior performance in the letter search task.

We suggest that the difference in performance across groups in the two tasks is due to the atypically-reading adults demonstrating a preferred reading strategy that applies sublexical processing. Sublexical processing and a letter-level of analysis for the letter search task are congruent with each other for grain size. It confers an advantage to the atypically-reading adults observed in the trend for higher accuracy scores. Sublexical processing and the word level of analysis for the word naming task are incongruent with each other. Using this processing strategy, many more parts of a word will need to be identified, remembered and integrated to produce a specific word item, inevitably increasing the chance of error.

The interpretation of a sublexical processing strategy fits for equivalent accuracy performance on the lexical decision task between the atypically-reading adults and other groups. The mixture of word and nonword items in lexical decision supports a sublexical approach for any reader however the *benefit* for the atypically-reading adults may be greater than for other typical-readers. The task requirement is not for identification of a specific word. A response can be made by way of recognition for an unlikely letter sequence within the letter string. Consequently, the sequencing and integration of all identified letters in a letter string is not a necessary condition for a successful trial, and the probability of making an error is reduced. This elevates the performance of the atypically-reading adults so that their accuracy is equivalent with the other groups.

Meanwhile, there is a question as to whether the processing of information for typically-reading 16-17-year-old and adults means they experience a cost in their performance for both the letter search task and to a lesser extent, the lexical decision

task. The lower odds of accuracy in the letter search task, relative to the atypically-reading groups, may be due to being unable to switch off phonological or lexical processing in the letter search task. The lexical quality hypothesis posits that for highly-skilled readers, the orthographic and phonological information is essentially one factor (Perfetti and Hart, 2002), so this assumption is plausible. If this processing style was applied in the lexical decision task, the cost maybe lower as now there are words as items. The cost is not fully extinguished because specific word recognition may not be the optimal strategy under these sample conditions. The reduction in the cost experienced for the letter search task, plus the suggested elevation of the atypically-reading adults' performance closes the performance gap between themselves and the atypically-reading adults. The model reflects this by recommending that the model with no group is the best fitting for the data.

The sentence reading task is harder to interpret. An intuitive explanation is that the surrounding, supportive presence of words semantically primes the word for correct pronunciation (Perfetti and Stafura, 2014). However, there is also priming through repetition of items across conditions in this task. While it is perhaps easy to assume that the priming properties of sentence context either equalises performance or negates the need for sublexical processing strategies, task design may be inducing practice effects and masking other possible effects in this instance.

8.2 Cross Task Comparisons

In this section, we look across tasks for similar and different predictors and discuss them in the context of task specific demands and the participant sample, with a focus on the atypically-reading adult group. We begin by looking at individual difference measures first and then examine the psycholinguistic predictors for their level of support across the preferred models. We discuss findings with respect to the lexical quality hypothesis (Perfetti and Hart, 2002) and the division of labour hypothesis (Plaut et al., 1996).

8.2.1 Individual Difference Measures

Rapid naming skill (RON) is suggested as an influential predictor across all experimental tasks on reaction time measures (credible intervals on word naming just cross zero). This is intuitive if we accept that speeded tasks depend upon a general domain skill such as processing speed. We find this quite surprising however, as the weight of the literature for RON with respect to word recognition suggests that RON may be more relevant to studies with younger children. Meyer et al. (1998b) and Hulslander et al. (2010) found that measures of RON predicted word reading into later years of school for samples that included atypically-reading individuals.

Meyer et al. (1998b) documented that it was the object/colour versions of the task that continued to show the prolonged relationship with reading development over the years. By choosing the object form of the task we may therefore be seeing a specific result that looks like a difference but the reviewed literature may have used letter and digit naming versions of the task.

An alternative explanation is that an element of working memory is implicated. Katz et al. (2012) suggested that effects of RON (as observed in student readers with a range of reading difficulties) could be acting as a proxy measure for working memory capacity. However, only letter search and sentence reading explicitly involve holding items in memory to be able to complete a trial.

The RON effect on lexical decision may arise as a lagged effect of the letter search task as it preceded the lexical decision trials in each data collection session. As a design element, this could be confirmed by separating the tasks in future studies.

Alternatively, the lexical decision task could have a working memory demand if, as suggested above, the participants were heavily reliant upon a sublexical processing strategy. Parts of unfamiliar items would need to be held in working memory while a decision was made. Approaching the lexical decision task at a sublexical level of processing when the task requires a lexical level unit to be identified increases the cognitive load on all trials, not just unfamiliar trials, possibly invoking working memory and reflected in an effect for RON.

Future studies could test for this by collecting a measure of working memory and asking for immediate recall of lexical decision items after a decision is made. The level of recall could be left free to vary across trials which may allow for some inference about preferred levels of processing.

Which version of memory to operate would be the important question. Talwar et al. (2018) finds that verbal working memory is a better predictor than short term working memory in a sample of adult-learners, however Swanson (1994) found that both types of memory measures were related to reading ability, short term memory for readers with disabilities and WM for readers without disabilities. Mellard et al. (2016) tested auditory working memory and found it did not differ across two subgroups across the sample but remained important for predicting reading progress. Just this small sample of studies shows that this question may have some relevance to further study, however the best way to operationalise the construct is not entirely clear.

Modeling RON and a memory measure simultaneously may show that RON is no longer a relevant predictor if it is indeed acting as a proxy for memory. The persistence of RON across reaction time models suggests that, an explanation notwithstanding, statistically adjusting for processing speed is important to account for differences when comparing a wide range of skills in a participant sample.

The remaining three experimental tasks showed a wider range of ID measures as reliable effects. Nonword reading skill and vocabulary were indicated across accuracy outcomes in the preferred models, showing higher odds for accurate responses in the context of stronger skill. On reaction time outcomes, nonword reading was present for lexical decision. Word reading skill was indicated for sentence reading reaction time outcomes; neither was present for letter search or word naming. Spelling was indicated on word naming accuracy measures and sentence reading reaction time measures as very small effects. This converges with the meta-analysis by Swanson et al. (2003) that real word reading was best predicted by measures of nonword reading and spelling.

Nonword reading, spelling, and vocabulary skill represent the triad of

components for the lexical quality hypothesis. Spelling would be cast in the role of the orthographic component in the absence of word reading skill (Andrews et al., 2020; Treiman, 2018), with a critical difference between the two measures of the level at which they operate: word reading at the lexical level and spelling at the sublexical level. Does the presence of spelling over the presence of a reliable word reading measure indicate that sublexical variables are in the dominant roles for reading forever more? How do you move to a more lexical style of reading if a lexical level variable is not present to form a bridge. The three factor solution for less-skilled readers (Perfetti and Hart, 2002) also had spelling on one factor, with nonword reading and word reading loading together onto another factor. Crucially, it was nonword reading that linked the two factors together.

Nonword skill is estimated in the presence of vocabulary for lexical decision and word naming but not sentence reading. We have interpreted the lack of a vocabulary effect in the sentence reading accuracy as a difference in the trial level information between tasks. The target items in the sentence reading task are embedded in an external source of semantic information to which everyone has equal access. Individual differences in vocabulary are consequently less relevant.

In vocabulary and nonword reading, we have semantic and phonological information that can work together to support word recognition, as suggested by the division of labour hypothesis (Plaut et al., 1996). Over-reliance on such a path however may reduce orthographic learning and knowledge development over time. The mapping induced at each learning episode that is supported by semantic information will not be as useful for the next word learning episode as the mapping for an episode supported by phonological information. Orthographic-semantic relationships are less systematic than orthographic-phonological relationships. Knowing one orthographic-semantic relationship neither helps in knowing the next orthographic-semantic relationship nor the pronunciation of the next encountered word.

Furthermore, the utility of the semantic-phonological division of labour may vary as a function of vocabulary knowledge (Dilkina et al., 2008; Plaut et al., 1996).

Figure 6.3 displays the low vocabulary profile of atypically-reading adults as a function of their standard score. In the context of low vocabulary, sufficient semantic information may not be present to help, leaving the brunt of the work to be completed by phonological skills.

Spelling was ranked as one of the stronger skills for the atypically-reading adults in their reading-related skills, however we know that the orthographic-phonological relationships for atypically-reading adults are unstable and vary across performances. Spelling errors are more likely to be a poor match for the sound form of target items and are inconsistent for an item across time. The stronger ability to identify single letters as demonstrated in the letter search task but variable application across types of spelling errors could point to a under-developed knowledge of the relationships between adjacent letters. This reduces the ability to predict which letter is more likely to follow another from the sound of a word. It further constrains the development of orthographic knowledge for prediction and also orthographic redundancy that comes from knowing that frequently occurring letter grouping (Ziegler and Goswami, 2005). This increases cognitive demands to remember single letters in a sequence.

Previous studies have found that continuing use of nonword reading skill for word recognition is indicated in readers at risk of or with reading difficulties (Katz et al., 2012; Steacy et al., 2017a). Bruck (1990) found that their adult readers with dyslexia were slower than the typically-reading 11-12-year-olds. This is not the case here. For this sample, data and analyses, group differences are indicated for accuracy measures and not reaction time. Additionally, a hallmark symptom of phonological dyslexia is observed in a weaker nonword reading ability relative to word reading (Castles and Coltheart, 1993). We do not observe that here. We observe equivalent levels of skill as measured by word and nonword scores. These two behaviour markers would suggest that on the whole, the atypically-reading adults are not readers with undiagnosed phonological dyslexia.

This begs the question of whether the markers of developmental surface dyslexia are present. However this is out of scope for the present study and remains

an open question. A further study, designed to answer such a specific and important question, would need to be established.

An alternative explanation for the presence of nonword reading over word reading skill may be that, given that our sample is young and less-experienced or atypically reading (but for the typically-reading 16-17-year-olds and adults), many more words may be unfamiliar to this sample. Ricketts et al. (2011) consistently found that nonword reading washed out the effects of other predictors for orthographic learning of novel words in a non-selected sample of primary school aged children. We do not see that here, nonword reading is supported by multiple ID and psycholinguistic predictors. Nation and Castles (2017) state that phonological skills remain salient for all readers for unfamiliar words. This is not a satisfactory explanation either, since word naming accuracy rates were at 97.3% in the sample, suggesting that for the majority of words, familiarity was good enough for successful recognition.

Lexical level word recognition, as measured by word reading skill, may not be sufficiently strong at the person-level in this sample. Hence, the prevalence of nonword reading skill as a reliable predictor across task models. Nonword reading skill can approximate lexical level recognition and also accommodate the additional sublexical processing requirements of unknown nonwords in lexical decision and letter search tasks. In this way, the properties of nonword reading skill make it relevant to a wide range of tasks. A potential downside of this is that approaching every task with one strategy is likely inefficient. It is well known that stronger readers change strategy according to task demands (Brown and Deavers, 1999; Tamura et al., 2017; Treiman et al., 1990).

The similar strength of nonword reading skill effects across the three tasks suggests that nonword reading may be being used in the same way across tasks. The items across the lexical decision and word naming tasks are identical, yet the tasks demand different types of processing and different types of output (Andrews, 2012; Balota and Chumbley, 1984; Chumbley and Balota, 1984). The overarching presence of nonword reading in the face of different task demands suggests that for the

majority of the sample, switching strategy across tasks is not occurring.

In contrast, vocabulary does show different effect sizes (lexical decision accuracy log-odds = 0.14; word naming log-odds = 0.44) for the same items. In the context of tasks where sublexical processing can complete a trial, vocabulary as a lexical variable may not be called upon as frequently as in word naming. In word naming, where all items are words, when it is known to the individual, vocabulary knowledge is more helpful to identify the specific word, with boosts to the signal of words that are known creating a stronger vocabulary effect, as suggested by the division of labour hypothesis (Plaut et al., 1996).

We suggest that the strength of the vocabulary effect in word naming expresses an advantage for those words that are familiar to all participants but also reflect use by stronger participants in the sample and who can more easily switch between sublexical and lexical levels of processing as the task demands. The vocabulary effect size may be attenuated in lexical decision if, as mentioned earlier, the stronger participants switch from a lexical level processing strategy to sublexical processing to accommodate the mix of familiar (word) and unfamiliar (nonword) items, thereby reducing the weight of vocabulary's influence.

If we assume that the vocabulary effects are driven by the stronger participants this may mean that the mechanism of the division of labour hypothesis is not so useful to the atypically-reading adults and possibly atypically-reading 16-17-year-old readers. Their low vocabulary levels are not strong enough to render useful support for word recognition.

Alternatively, the vocabulary and phonological skills may be present in sufficient quantity and it is the interdependence part of the relationship that has yet to develop (youngest readers) or has not developed (atypically-reading 16-17-year-old and adults). In this case, an underlying problem as suggested by both the lexical quality and the division of labour hypothesis stems from an identical basis, a lack of strong interdependence between the critical components for word recognition.

This would suggest that representations are of low lexical quality. Effects for word reading skill were present, but for the exception of sentence reading reaction

time, they were always slightly smaller in magnitude than nonword reading skill effects and unreliable. The lexical quality hypothesis states that for words of high lexical quality the presentation of orthographic information is sufficient to bring about successful word recognition, making phonological and semantic information redundant in the process (Perfetti and Hart, 2002). No such redundancy is present for words of low lexical quality, which needs all three of the phonological, orthographic and semantic components for recognition. We observed reliable nonword reading skill effects supported by vocabulary and spelling, a triad of measures that reflect the information sources upon which the lexical quality hypothesis rests.

Taken together, we suggest that atypically-reading adults may show an integration difficulty between orthographical and phonological information. The evidence for this is suggested by the relatively superior performance on the accuracy outcome for the letter search task, equivalent performance on lexical decision and weak accuracy performance on the word naming task, all performed under equivalent reaction time performance.

Each of the tasks can be achieved with sublexical processing however the work that sublexical processes must do increases from letter search to lexical decision to word naming. The overt pronunciation for a discrete word in word naming further demands that whichever type of processing is used, sublexical or lexical, orthographical and phonological information must be integrated to produce a pronunciation.

Perfetti and Hart (2002) assert that the lack of integration of mapping within a learning episode slows learning of orthographic knowledge and learning in the long term. Words of low lexical quality are all affected since their overlapping attributes do not help each other over successive exposures to the same extent. The transition from low to high lexical quality is slowed.

Relatively good performance in the letter-search task may be symptomatic of single letter level processing. The letter search task could also be interpreted as being a pure test of visual encoding. Successful trials may be achieved merely by looking and recognising an object (a single letter) from a string of objects. If the

atypically-reading adults performed the letter search task by seeing, this may explain their stronger performance, relative to lexical decision and word naming. As Greenberg et al. (1997) hypothesised, adult-learners may 'read by seeing', leading to a weak performance in word naming where a specific item is required for accuracy and rules of pronunciation change according to the spelling of the item. Yet we deliberately limited the conditions by which a visual matching strategy could easily be used by mixing the case of the target letter presentation episode and the target letter identification episode. Letter names link the two visual forms of the letters which involves an element of integrated orthographic-phonological information. As a result we believe that simple orthographic-phonological relationships at this level look secure.

Orthographic-phonological information in a spelling task represents more complex levels of knowledge, where relationships are conditioned upon letter position. More information is brought to bear in this task and the stimulus of the target word has no visual cue. Not only are simple phonological-orthographic mappings tested in this task, but different grain sizes of information, distributional characteristics representing frequency of the most likely spelling and also selection of a correct orthographic form for a semantic context. This knowledge of how spellings and sounds may change in relation to one another appears truncated (Bruck, 1990; Masterson et al., 2007).

Equivalent performance in the lexical decision task for atypically-reading adults may arise from the application of sublexical processing for entire letter strings and the absence of needing to integrate orthographical and phonological information for a word pronunciation. The requirement of a pronunciation in the word naming task further lowers accuracy rates for atypically-reading adults. Even with good sublexical processing, if the knowledge of the orthographic-phonological information is not present, a correct answer cannot be derived. The atypically-reading adults' reliably lower odds for accuracy on word naming may suggest that although sublexical processing can achieve word recognition, a difficulty with integration for, or incorrect or under-developed orthographic-phonological knowledge will lead to incorrect pronunciations.

There is a connectionist implementation of this kind of reading behaviour. Harm and Seidenberg (1999) damaged a version of a PDP model by reducing the number of connections between orthographic and phonological domain layers. As a consequence, the model could not learn the relationships between letters. It did not learn the different grain sizes available to readers of the English language and developed instead a preferred reading style of letter-by-letter decoding[2].

In summary, atypically-reading adults in this sample show a reading-related skills profile that is similar to that of older secondary school children. The equivalent scores between word and nonword reading skills may suggest that orthographic learning is insufficiently strong to work as the dominant reading strategy, and nonword reading skill with support from vocabulary and spelling form a complement of skills to facilitate word recognition. These multiple components of information need integrating for each item, which is costly and error prone in the long-term. The array of skills supports an interpretation that familiar words are of low lexical quality. The coupling of nonword reading skill with vocabulary suggests that phonological and semantic sources are working together, as the division of labour hypothesis states. However, each of the component skills are relatively weak.

When the output of a trial is a letter or making general decisions, this produces relatively good performance for atypically-reading adults. When the target word is surrounded by other words that prime word recognition, the sentence reading preferred model suggests that accuracy performance is equivalent. Yet once a specific word is required in isolation, over the same time course, accuracy performance falls to below that of the youngest readers. The specificity of this decrement to performance as located in word naming suggests to us that the integration of orthographic-phonological information as a potential site of difficulty and may be truncating or slowing development of a broad orthographic knowledge, with consequent impingement on orthographic learning over the longer term.

[2]Apropos of the (out of scope) question regarding atypically-reading adults and surface dyslexia: this implementation precedes the next simulation that involves much more severe lesioning across several locations in the architecture and that demonstrated developmental surface dyslexia type behaviour.

8.2.2 Psycholinguistic Variables

We come to the level of psycholinguistic variables. This is the first study we know of that includes multiple psycholinguistic variables and explores their influence for atypically-reading adult word recognition processes.

None of the preferred models were interaction models, which suggests that atypically-reading adults are similar in their use of psycholinguistic predictor information, at least for the group of readers in this sample. Group differences were indicated across two accuracy models, although the letter search preferred model did not include psycholinguistic predictors. All of the remaining preferred models included psycholinguistic predictors.

Predictors for arousal, dominance, LPC, and valence were not reliable in any of the models. We believe this to be due to sampling reasons and the niche properties of these variables. Yarkoni et al. (2008) designed the measures underpinning the LPC predictor to accommodate longer words where N-size could not. In the company of N-size and a sample of monosyllabic items (maximum no. of letters = 8), the N-size effect has probably appropriated all the relevant variance for the construct.

Arousal, dominance and valence, all predictors that capture an affective type of semantic information, often work together with concreteness, imageability and in interaction with their context (Snefjella and Kuperman, 2016). We did not focus on interaction effects between psycholinguistic predictors in this first look at the atypically-reading adult population. However, the effects we observed for concreteness (see below) may suggest that they could be relevant in future studies.

The preferred lexical decision model for reaction time enjoys a high level of semantic support from psycholinguistic predictors. By this token, the cumulative amount of semantic effects do appear to be greater for lexical decision than word naming, however the sources of semantic support are disparate. In contrast, apart from AoA, the type of psycholinguistic predictors for word naming are working at a phonological and sublexical level.

AoA predicts reaction times for both lexical decision and word naming. It is

also a reliable predictor for lexical decision accuracy. The estimate for AoA is larger, in absolute terms, for lexical decision compared to word naming reaction times (~30 ms vs ~14 ms) and accuracy (log-odds = 0.61 vs 0.29, though unreliable for word naming accuracy). AoA is not indicated as a reliable predictor in the sentence reading task. It would seem that for this sample and these models, the AoA effect attenuates when the support of semantic information lessens (Ellis and Lambon Ralph, 2000; Monaghan and Ellis, 2010; Morrison et al., 2002). We interpret this as AoA acting as a semantic predictor in this sample, rather than an adjunct to word-frequency, contributing to the accumulation of semantic information by which the correct decisions and pronunciations are facilitated.

The AoA effect may be spurious however, in that some of the AoA ratings are above the age of the youngest participants. When errors were trimmed from the data set for reaction time analyses, this could mean that the remaining trials showed a bias for earlier learned words. We looked at the sample of words for which errors were made in both lexical decision and word naming (16 items), splitting the sample into two groups of ratings: 11 – 11:6 and 11:6 – 14:3 years.

In absolute terms, younger readers made significantly more errors across both groups of AoA words than 16-17-year-old and adult groups in lexical decision. The difference in error rates between the two AoA ratings groups was significant only for the typically-reading 11-12-year-old group. Word naming was a little more divided: The atypically-reading groups made similar levels of errors to the typically-reading 11-12-year-old group while the typically-reading older groups were consistently lower on errors. In summary, the AoA effects in the models do not appear to be because those items rated as being learned at an older age than our youngest participants were unfamiliar to only the younger readers. They were just as likely to be unfamiliar to some of the older, atypically-reading participants.

In both tasks, the AoA effect predicts that earlier learned words are faster and more accurately recognised (as words). In lexical decision, number of word meanings, semantic diversity and imageability also facilitate faster response times for words of higher values. Conversely, concreteness inhibits a response (reliable also in

sentence reading – see below). Each of these remain reliable after the differences in vocabulary between participants are accounted for.

Words of multiple meanings are believed to excite faster activation because the independent meanings contribute multiple increments of semantic activation to patterns across orthographic units for a word (Balota et al., 2004; Jastrzembski, 1981; Jastrzembski and Stanners, 1975). Words that are used in multiple contexts contribute nuanced meanings that add to cumulative semantic contribution (semantic diversity and number of word meanings show a correlation of $r = .48$).

The plurality of the ways in which a word can be experienced, gives them an advantage over words that have a more niche application. Convergent with semantic diversity, words of greater abstractness (lower concreteness) or multiple meanings, are likely to have more possible meanings than concrete words and are more able to be used across many contexts. Often labels for specific objects carry the fixed meaning across a narrower selection of contexts (Adelman et al., 2006; Kousta et al., 2011).

All of these effects are very small, between 7 – 11 ms. It is the type of predictors that is interesting. Steyvers and Tenenbaumb (2005) suggested that early learned words as represented by AoA form the centre of hubs of semantic networks, with spokes of later learned words forming connections as a function of overlapping semantic features. The selection of predictors gives support to the argument that breadth as much as depth of reading type is key to efficient word recognition (Hsiao and Nation, 2018; Keuleers and Balota, 2015). In this sense, AoA is a depth metric while semantic diversity and number of word meanings provide a breadth metric (Hoffman et al., 2018).

Imageability and concreteness are type measures that share a positive relationship with each other. We observe a very small effect for imageability in lexical decision reaction time (β = -0.03). The direction of effect is similar to previous studies in that imageability shares a negative relationship with reaction time (Balota et al., 2004; Davies et al., 2017). Crucially, this effect is present over and above an independent effect of AoA (Baddeley et al., 1988; Klose et al., 1983; Woollams, 2005). This is important because imageability is often used interchangeably with

concreteness (Kousta et al., 2011) and as early learned words are often of high concreteness values, in the absence of an AoA effect, imageability is often interpreted as a pseudo-AoA effect.

At the same time as a facilitatory effect for imageability, we observed a reliable, inhibitory concreteness effect on lexical decision. The opposite direction of this effect is often observed, with a facilitatory effect for words of high concreteness ratings (Cohen-Shikora and Balota, 2016; Strain and Herdman, 1999). However, in the present study, this inhibitory concreteness effect is also observed on sentence reading reaction time models, and the letter search design implied model, without the complementary presence of a reliable imageability effect.

There was evidence of a significant difference in concreteness ratings between items for list 2 and 3 in the sentence reading task ($p = .049$). These items are also included in lexical decision and word naming trials. We considered if the concreteness ratings were driven by this difference. The difference was not present in lexical decision lists ($p = .824$), nor in the word naming lists ($p = .788$).

Sentence reading items were not in the letter search sample items. Letter search always preceded lexical decision trials, negating the possibility of a lagged effect as an explanation. Additionally, the sentence reading items were in the word naming trials but concreteness is not a reliable predictor for word naming reaction time. Consequently, it is unlikely that the difference between concreteness ratings in list 2 and 3 of the sentence reading task are the only source of the concreteness effect.

Kousta et al. (2011) and Barber et al. (2013) found that when paired concrete and abstract items were identical for imageability ratings, then abstract words were identified faster than concrete words. Although in the absence of one another, facilitatory effects on either concreteness or imageability are observed, in the presence of each other imageability partials out the indirect effects of concreteness on the items with the residual abstractness effect eliciting faster reaction times.

Two theories proffer mechanisms by which words with higher concreteness values have an advantage over words of lower values: Dual Coding Theory (Paivio, 1991) and Contextual Availability (Schwanenflugel et al., 1988). Although different in

their details, both essentially claim that a multiplicity of sources of information for concrete items relative to abstract items, contributes to greater activation for the item pattern over the same time course, resulting in faster responses for words of higher concreteness ratings.

Kousta et al. (2011) uses the basis of these accounts to explain the inhibitory effects of concreteness. The richer or greater number of sources of information on concrete items need greater integration, with words of lower concreteness ratings exciting fewer sources of information and so needing less integration. Integration processes incur a time cost and result in the observed inhibitory effect of concreteness when imageability has been partialled out. Within this sample, across the wide range of vocabulary scores and word reading skills, semantic properties assist at the item level for recognition of more challenging words, with richer and denser definitions of items taking longer to integrate and activate for each participant.

Quite apart from the item-level explanation, the implication of integration processes aligns with our interpretation of findings that integration of multiple sources of information may be a location of difficulty for atypically-reading adults.

But for AoA (discussed above) and vocabulary, none of the predictors on word naming are for semantic properties. They represent phonological and by extension, sublexical levels of word processing (Hofmann et al., 2007). Consistency is indicated on both outcomes for word naming with N-size on reaction time and bigram frequency on accuracy. The lack of an effect for consistency and bigram frequency in the lexical decision accuracy model, while using the same items, suggests that the effects for word naming are related to the phonological and articulatory processes of the word output (Andrews, 1992; Balota et al., 2004; Coltheart et al., 1977; Thompkins and Binder, 2003).

The bigram frequency effect predicts lower odds of an accurate response for words that contain high frequency bigrams. Bigram frequency effects in previous studies are inconclusive. Some studies have found effects (Broadbent & Gregory, 1968, Orsowitz 1963 cited in Gernsbacher, 1984). Other studies have not: Chetail (2017) observed no effect for bigram frequency on accuracy outcomes for a typical-reader

sample. Davies et al. (2017) found a small, positive effect for bigram frequency in an interaction with reading skill on reaction time in a sample that included young children and elderly participants. The difference in findings may be due to the inclusion of a typically-reading child sample in Davies et al. (2017). Appropriately lower skill in the typically-reading children compared to the typically-reading adults may mean a greater reliance on sublexical processing that is then expressed as a small cost for older skilled readers.

In Chetail (2017), this variation is not present and so no bigram frequency effect is observed. In the present study, we have younger and older readers of both typical and atypical reading skills and sublexical parcels of information appear to be relevant. The influence of bigram frequency in the word naming task data suggests that this sample of participants is not processing words via a lexical, whole word strategy.

The positive bigram frequency effect occurs in the presence of a reliable negative consistency effect. Bigrams are two letters in length. The consistency predictor is constructed at the level of the rime of words. Consequently, it is at least two letters in length but more often larger. The concept of grain sizes (Ziegler and Goswami, 2005) and their use for word recognition becomes relevant in the presence of these two effects occurring side by side. The consistency effect is located toward the end of a word, while bigrams are constructed along the length of a word. It is not clear whether the location of the effects are overlapping or occurring at different parts of a word. Andrews (1992) controlled their items for initial phonemes and the bigram effect of an earlier experiment disappeared. We have performed a statistical adjustment that should approximate such a control. Future studies could control initial phonemes of items to explore this finding in greater depth for a similar sample of participants.

In terms of the lexical quality hypothesis, a bigram frequency effect is further evidence that many words within participants are of low lexical quality. Low lexical quality for an item implies a lack of orthographic redundancy, with letters at all positions of the word remaining salient (Perfetti and Hart, 2002). A connectionist

interpretation describes an interactive competitive process where words that overlap in features are activated in parallel. Pairs of letters with high frequency activate all those words with the bigram in the same position (Plaut et al., 1996), increasing the number of items for recognition thus decreasing the probability of the correct item being selected for recognition. Each of the explanations suggest that sublexical processing is the dominant mechanism for word recognition.

On reaction time, consistency and N-size are reliable. N-size shows a very small, negative effect (β = -0.06). The items contributing to the consistency measure are a subset of a neighbourhood since our consistency measure is constructed by words that are similar and dissimilar at the level of the rime. Consequently, the N-size effect may be attenuated because the majority of its influence is explained in the consistency effect.

In the context of low spelling skills and low vocabulary, a facilitatory effect for N-size for efficient naming on accurate responses suggests partial decoding processes (Andrews, 1989; Andrews and Hersch, 2010). In lexical quality hypotheses terms, word representations are not yet unitary; in connectionist terms, since the pattern of activation for the discrete word has not yet stabilised, there is sufficient time for the co-activation of words containing similar letters to activate and assist with pushing the activation signal strength to threshold. In each explanation, this means that orthographic similarity can help, with words of higher consistency having an advantage.

In the context of low orthographic learning, the difference between words may be as minimal as one letter, as suggested by the N-size effect, making the selection of the correct word highly error prone because words are very similar in appearance. Additionally, for accuracy, the location of a bigram may be critically important, due to the complexity of the English spelling sound system. The position of some bigrams in relation to adjacent letters, may alter the pronunciation of letters in a word, introducing multiple possible pronunciations for parts of, or whole, words. Over and above the number of orthographic forms, the number of possible pronunciations may be larger. For participants of low orthographic knowledge, this further amplifies the

chance for error.

Together with spelling knowledge on word naming accuracy, this suggests that where the orthography and phonology of a word are systematic, accuracy and speed of pronunciation are promoted (Andrews and Lo, 2013; Dilkina et al., 2008). Andrews posited that orthography drives recognition when spelling is implicated as a reliable predictor. In the word naming accuracy model, spelling is a reliable predictor, however the size of its influence is very small, compared to the other effects. On balance, phonology and semantics are working together and the orthographical component appears to play a weaker role.

The sentence reading task has not featured much in the discussion. As a bridging task between a word level and sentence level of reading, and with so few predictors being indicated as influential, it feels distanced from the nexus of lexical decision and word naming. Consequential to the current argument, however, is that nonword reading and – tentatively – consistency facilitate accurate responses over and above the larger sentential context in which the target word is placed. For reaction time, word reading skill takes the lead over nonword reading, in the presence of a RON effect. Participants of higher word reading skill and processing speed are faster, over and above the assistance of context. Spelling also facilitates faster responses on correct trials.

There may be a confound inherent in the sentence reading task that enables the word reading skill to come through, however. This is a possible practice effect from the repeated items across the meaningful and neutral condition within a data collection session. We chose to maintain the pairs of words within a data collection session to provide some fidelity of the data should participant attrition occur. In support of this argument, as part of the neutral condition, half of the sentence contexts were also repeated. Since word reading skill has been notably absent in all other models within the study, the high content of repetition within this task suggests that this the word reading skill effect could be an artefact of the task design.

In summary, the single word recognition skills of the atypically-learning adults in this sample are equivalent to a sample of late secondary school readers. Across all

ID measures, they perform significantly lower than their typically-reading adult peers. They are supported by an array of skills, of which spelling and vocabulary are relative strengths. Application of these skills to experimental tasks show that they are better at identifying letters in a letter string than recognising discrete single words. They are equivalent in their ability to recognise a word from a nonword and use sentence context to identify words to the same extent as the larger sample of participants.

Word and nonword reading skill appear to be as strong as each other, while being weak compared to other reading-related skill measures. The downstream effect of this appears to be a weak orthographic learning capacity and a suggestion that reading is performed by sublexical processing. The lexical quality hypothesis (Perfetti and Hart, 2002) assumes that orthographic information from word reading skill is dominant for words of high lexical quality. Orthographic knowledge as measured by spelling was the stronger of the reading-related skills for the atypically-reading adults. Critically, however, types of spelling errors showed a greater probability of making a plausible sound matched attempt as not while typically-reading 16-17-year-olds were more likely than not to make a plausible sound-match error. So while orthographic knowledge was relatively strong in the atypically-reading adults, application of that knowledge varied over time. This variability must impede the correct learning of correct spelling-sound relationships over time.

Nonword reading skill was the best predictor to effect word recognition across tasks that were congruent and incongruent with sublexical processing as optimal reading strategies. In the task where sublexical processing was sub-optimal as a reading strategy, the word naming task, the atypically-reading adults were lower in accuracy than the youngest readers in the sample.

The persistence of variables involved with sublexical processing as better predictors of outcomes leads to a conclusion that atypically-reading adults have many words that are of low lexical quality. The balance of words within the item sample that may be of high lexical quality is insufficient to push word reading skill forward as the recognition measure.

Low lexical quality indicates that sources of information are not integrated

with one another. They may activate at slightly different times and so mapping of repeated patterns is reduced in each learning opportunity. We have tentatively suggested that this integration or this mapping process may be the source of difficulty for atypically-reading adults because we observed stronger performance in letter search, equivalent performance on lexical decision and weaker performance on word naming. The word naming task is where the orthographical and phonological information must be integrated to be able to produce the correct pronunciation for the target item. It is here that we observed surprisingly low performance for the atypically-reading adults. As an exploratory study, we put this forward as an area for future studies to consider for confirmatory research involving an atypically-reading adult participant sample.

Across the sample, people of high processing speed made quicker responses. This is a surprising finding however previous research suggests that the RON measure could be a proxy for memory ability in older adults (Katz et al., 2012). Stanovich (1986) suggested that memory capacity may develop in low-literacy individuals as a consequence of reliance on memory for reading-by-rote. The absence of word reading skill in the models does not preclude lexical level processing operating, rather it suggests its is weaker, compared to nonword reading skill in the context of task demands. If RON is a proxy for memory, this may be expressed as a lexical level support. We put forward the inclusion of a more explicit measure of memory for consideration for future research design.

We introduced the division of labour hypothesis in section 3.1, and we can see that conditions for it to be useful as an explanatory mechanism are present: word recognition is most definitely supported by sources of phonological and semantic information for this sample. The quality of semantic information for atypically-reading adults has been addressed in previous studies, with several authors hypothesising that much of semantic knowledge for this atypically-reading adults arises from spoken language experience (Braze et al., 2007; Mellard et al., 2012b). As such it may be a weaker source of information compared to that garnered from print sources due to the lack of orthographic information input at the time of using the

word. Further investigation of this origin of semantic knowledge and resultant differences in strength of semantic information would be useful.

Our interpretation hinges upon a cross task difference observed under exploratory research circumstances. So, the findings and logically, the conclusions, are very tentative and need replication before any firm assertions as to a difference and then a site of difference is made. That being said, the realisation that under speeded conditions, an atypically-reading adult's accuracy for single word reading could be less accurate than a reader in the first year of their secondary school experience, may be of interest to persons involved in such education - learners and tutors. It may be grounds to motivate a further research endeavour with the explicit goal of confirming or disconfirming these findings. Upon confirmation of findings, the psycholinguistic predictors herein are markers of types of materials and ways of working with atypically-reading adults that could form the basis of intervention research to raise the lower levels of accuracy in single word recognition.

8.3 Limitations

Much has been made of a dominant reading strategy of sublexical processing for the atypically-reading adults. A strong test of this would have been the presence of a length effect for lexical decision and word naming outcome measures. We were conservative in our approach, due to an explicit aim of modelling multiple predictors simultaneously. When length was suggested as having an overly high VIF value, we decided to use the phonemes predictor as a proxy for length instead. Knowing what we now know, and looking at the phoneme predictor, the range of which may be restricted compared to that of length, we regret removing the length variable. We would either ensure that the phoneme variable gave equivalent coverage in future item samples or keep length as a predictor.

The interaction models that did converge all had higher IC values than the preferred models. However, some interaction models were still running at the time of writing. It is likely that the data from the present study were not sufficient to

successfully model interactions to the same extent as independent effects. Given the small and very small effects returned by independent predictor models, it is likely that much more information is needed to fully run, check and return an interaction model. This is especially true in the word naming and sentence reading tasks where accuracy was very high which adds to difficulties for estimation.

The rate of attrition across the atypically-reading adult and 16-17-year-old groups was the highest within the sample. The reduction in information on T3 experimental tasks, coupled with the high rate of accuracy in word naming and sentence reading tasks probably explains why the interaction models did not converge. In lexical decision and letter search, it may explain why the interaction models were not preferred. This remains an open question with a larger amount of data needed to be able to estimate small effects that exist, if they do at all. Not only does this indicate recruiting a larger sample, but over-recruiting to maintain optimal amount of information in the face of inevitable attrition across time.

In the letter search task, the number of trials in the letter position = "none" condition was equal to the amount of all the other letter position trials put together, merely to calculate the intercept of the model. Any future replication may consider rethinking the distribution of trials across the position variable so that it is more evenly balanced and more information is given to the estimation of effects across the variable. As a result of more information, other predictors may be estimated with greater precision and the inferences on the group predictor may be more conclusive.

Finally, the word naming and sentence reading accuracy levels were too high. A greater level of challenge needs to be added to the item set. Lowering the accuracy rate would possibly allow for more precise estimation of effects and a check on whether group differences in word naming accuracy are still indicated. For the word reading task, we would argue that the addition of nonwords would be the best choice of introducing a level of challenge, with presentation of word and nonwords in separate blocks to encourage use of both word and nonword reading skills for those whose strength of skills enables them to strategise this way. We would predict that this would be expressed through an interaction effect for group and word-reading skill

in a preferred model.

8.4 Future Directions

The most obvious need in the short term is replication of these findings, to ensure that this sample is not, in some unobserved way, unique and that the findings generalise to a new sample. Given the small population (relative to the school age population), the short window of time in which atypically-reading adults may be eligible for participation, and their rate of attrition in the present study, securing an optimal number of repeated sessions will likely involve a multi-site study of cohort design.

Extensions of the present work could be added to refine the measurements. Questions of a role for working memory are raised by the presence of RON effects. Including a measure of working memory would be desirable.

The prevalence of nonword reading and the array of ID measures that are indicated in the absence of a reliable word reading effect have been interpreted here as a preferential reading strategy at the sublexical level. HS99 approximated a mild form of developmental surface dyslexia in the HS99 PDP model by reducing connections between orthographic-phonological domain layers, inducing a sublexical reading strategy behaviour in the PDP model. Additionally, we have raised a question about integration of information in this population. HS99 suggests one site of difficulty.

Critical to the division of labour hypothesis is the strength of the interdependence between the two sources of semantic and phonology. The atypically-reading sample (both adult and 16-17-year-old) demonstrates a low level of vocabulary in the context of average phonological awareness skill. With the training protocol of HS99 and the action layer of D08 there are potential simulation environments that could explore the impact of different levels of interdependence (from both spoken and written semantic information) within a connectionist framework. This could guide or augment behavioural investigations.

The meta-analysis has been little mentioned in Chapter 8. With few exceptions, it was the models with the stronger priors that returned the lower IC

values and were consequently nominated as the preferred models. Information to form the strongly informative priors was lifted from the meta-analysis. The aim is to maintain the current project but also to develop it into a resource that can perform network meta-analyses across different sample constructs. Continuing on a bent of secondary analysis approaches, the sparsity of research and data for the adult-learner population means that every study is important and needs to work harder than a research field that is well populated. Using integrated-data-analysis techniques may mean that findings from the field can be augmented as single studies become part of a connected network of measures and yield secondary insights that may guide research design and foci.

8.5 Conclusion

The present study explored the single word recognition processes of a group of adult-learners. We assessed the participants on a series of reading-related individual difference measures to connect with previous studies. We have found that, like previous study findings with similar groups, the adult-learners have scores that emulate the performance of secondary school students.

Many of the previous studies compared adult-learners with younger readers of 11-12 years of age. We also included readers of 11-12 years of age and extended the reader sample to include 16-17 year old readers. We have found that, in respect of individual difference task scores, the adult-learners are positioned in between the typically-reading 11-12-years olds and 16-17-year-olds.

We are the first study that we know of to extend the exploration of adult-learners' word recognition processes into the experimental task landscape and estimate psycholinguistic predictor effects for an adult-learner population. From the experimental task outcomes, we have been able to capture cross-task differences that may suggest atypically-reading adults engage a sublexical reading strategy to effect word recognition. Most importantly, while appearing equivalent in skill in ID measures, the atypically-reading adults were reliably less accurate than the youngest

readers in speeded word naming. By including psycholinguistic predictors, we have found that phonological and semantic predictors assist nonword reading skills for single word recognition.

In terms of the lexical quality hypothesis, adult-learners look as if their triad of orthographical-phonological-semantic knowledge are either too weak to cohere with each other or are not sufficiently integrated to support fast and accurate word recognition. In terms of the division of labour hypothesis, the psycholinguistic predictors that support lexical decision and word naming suggest that this sample of readers use phonological and semantic information to help access word forms. But once more, for word recognition, the skills may be weak such that the interdependence upon which the division of labour hypothesis is based is not robust.

An intuitive solution (after confirmation of the present findings) is to work with adult learners to bring about a more flexible or whole word reading strategy. However the dominance of nonword reading, spelling and vocabulary across the models suggests that these sublexical predictors may reinforce a fractionated processing style, making it a very hard habit to break.

With the inclusion of two younger sets of readers, who are on a trajectory to possibly become adult-learners later in life, there may be opportunity to confirm if their reading styles show similar patterns. If this is confirmed, are the styles mutable, and if so at what stage and which are the best tools and methods? This is the stuff of further study.

8.6 Summary

This thesis reports the results of two studies: a wide ranging meta-analysis that aggregated study level effects for eight psycholinguistic predictors across groups of child and adult readers and a longitudinal study that describes a sample of atypically-reading adults and their similarities and differences with younger readers. Both studies contribute new knowledge to research on single word reading processes.

From the meta-analysis, we found that while individual studies were explicitly

designed to compare groups for effects of predictors, many were statistically underpowered for the estimation of these interaction effects. This lowers our confidence in the estimation of aggregated effects. To raise confidence in aggregated effects, we suggested building sample numbers through collaboration and networking between smaller studies. A move towards greater collaboration would enhance the robustness of effect estimation yet involve only small changes to research practice.

From the longitudinal study, we may have a better understanding of two things: within the atypically-reading adult sample, word and nonword reading skills show equivalent strength, such that word reading does not perform as the primary skill when individuals are exposed to single words. Atypically-reading adults do not appear to have a *lexical* approach to reading. Instead, nonword reading acts as the primary decoding predictor, supported by vocabulary and spelling knowledge. Atypically-reading adults appear to rely on this nexus of skills which are themselves under-developed and vary in their application over time. A trend for superior performance in identifying the presence of embedded letters in a letter string (as observed in the letter search task), may suggest that a lower accuracy in the single *word* reading task may be because atypically-reading adults are applying a sub-optimal reading strategy and that single word reading is performed using sublexical reading strategies.

At the outset of the study, we asked whether atypically-reading adults may enjoy a greater vocabulary knowledge than younger readers because of their older age and longer natural language exposure. As measured by the Shipley vocabulary scale in the present study, they do not. Furthermore, the presence of independent effects of vocabulary, rather than interaction effects in our statistical models, suggests that atypically-reading adults use their vocabulary skills in a similar way to younger readers. Spelling skill is relatively strong within the atypically-reading adult individual differences measures. Over time, however, observed spelling performance for the same item is highly variable in the atypically-reading adult sample.

In terms of psycholinguistic predictors, we found that independent effects of AoA, consistency, word-frequency, neighbourhood-size, bigram-frequency and

concreteness were reliably indicated as predictors of outcome variables across the four experimental tasks. Confidence intervals for the predictors of imageability, number of word meanings and semantic diversity fell on or just over zero, which we think means they should also be considered candidate predictors in future models. Future research should seek to replicate these models and confirm these results. When we compared the independent effect sizes of the longitudinal study with those of the meta-analysis, we found our study effects sizes to be generally smaller, except for the word-frequency effects on accuracy measures, which were larger.

The longitudinal study modelling strategy has demonstrated that in a moderately sized study, the parallel estimation of multiple predictors is possible. This differs from the much more conservative modelling strategy adopted by many of the included studies in the meta-analysis, where only one or two predictors were modelled in conjunction with each other. Given the theoretical reading models, we performed a stronger test of the viability and influence of psycholingustic predictors. This is likely to yield a more precise estimate of effects as a result of the more stringent testing conditions.

The findings from the present meta-analysis and empirical study will help to progress the methodological approach taken in the field. The meta-analysis revealed that it is clear the field is heavily invested in modelling interactions between group reading skills and psycholinguistic properties. However, statistical power is often low. In the empirical study, no interactions with time for any of the predictors were indicated. The lack of any effect for repeated measures over time suggests that the effects could be estimated with one round of data collection and that a correlational study design is appropriate.

Taken together, the findings suggest that correlational study design and collaboration across researchers would boost sample sizes and increase power to detect small interaction effects. Further, parallel estimation of multiple predictors would estimate effect sizes conditional upon the presence of a more representative set of predictors. This would facilitate greater precision in the estimation of predictor effects and identification of those predictors that show no influence when tested in

conjunction with the larger set. This will also progress the field in terms of constructing valid models that can speak to theory and cognitive models of single word reading.

Furthermore, the meta-analysis provides a contemporary dataset of research findings, available as an open resource to the research community. We chose to use the findings as strong priors in Bayesian Inference models that use linear-mixed-effects-models. Linear-mixed-effects-models offer greater flexibility, can incorporate different distributions of data and accommodate the correlated structure of observations that is integral to much of reading research empirical data. The additional flexibility of the Bayesian Inference paradigm arises from the Monte-Carlo simulation method which in this study showed a greater likelihood of model convergence over the frequentist linear-mixed-effect-models. While the ANOVA method was clearly the preferred method of analysis in the included studies of the meta-analysis, the present study is a demonstration that it need not be so, going forward.

The theoretical implications of our work are somewhat more clear at the end of the study. To begin, we underpinned our study with the lexical quality hypothesis (Perfetti & Hart, 2002) and the division of labour hypothesis (Plaut, 1996). In concluding our study, we think the division of labour hypothesis allows for a richer description of atypically-reading adult single word reading processes. The end state of the lexical quality hypothesis is the lexicalisation of a single word such that the orthographic code of the item is sufficient to access an item's pronunciation and meaning. The prevalence of nonword reading, spelling and vocabulary skills observed across all models (bar sentence reading reaction time) suggests that the lexical representations within an atypically-reading adult are likely of low quality. The lexical quality hypothesis has little to say about the mechanisms by which information for items of low lexical quality are accessed.

In contrast, the PDP model account describes cognitive mechanisms where sublexical processes contribute to the growth of stable lexical representations over time, without ever requiring lexicalisation of a whole word. Additionally, the division

of labour hypothesis, with its extensive modelling in both P96 and HS99, provides accounts of alternative pathways by which phonological, orthographic and semantic information may operate in interdependent ways that reflect a skilled and less-than skilled reading performance. Furthermore, the PDP framework is extendable in that it could be both computationally and empirically explored in future research for a sample of atypically-reading adults. With confirmation of findings through replication, the current study's data and effects could act as an initial benchmark by which to judge the output of parameters within an adapted PDP model.

The applied implications of the present study may also be helpful to educators of atypically-reading adults. Although generally, the atypically-reading adult shows a profile of an older secondary school student, the accuracy of the single word naming task for this group was reliably lower than that of first-year secondary-school students. This exploratory finding may be of immediate use to educators. Also, we have found that vocabulary and spelling skill appear to be supportive sources of information for atypically-reading participants. Critically, there is some suggestion in the literature that for atypical readers of college age and above, vocabulary may be accumulated through spoken language experience rather than print exposure, and as such may not contain the dual codes of orthographic and phonological information. Consequently, the source of vocabulary knowledge available to atypically-reading 16-17-years old and adult individuals may be a weak source of information.

This may have implications for reading practice in a secondary school or further education college classroom where online and immediate reading of novel text is often performed. Such episodes are not framed as reading practice per se, since the reading activity is secondary to the primary goal of understanding the content in order to be able to complete a task. Much of this immediate reading is performed independently and in silence. Any exposure to a novel printed word does not provide access to its acoustic signal and an opportunity to experience the dual codes of print and sound contiguously is lost. This may have important implications for the development of vocabulary and spelling skill over time.

To conclude, our beginning aim was to understand how similar or different

atypically-reading adults were to cohorts of younger readers. We may tentatively say that atypically-reading adults' individual differences profile for component skills that contribute to successful single word naming look to be very similar to atypically-reading older secondary school students. Crucially, however, while similar skills may appear to be in place, when applied, the atypically-reading adults are lower in accuracy performance in single word reading tasks. We observed patterns of results across letter-search and word naming tasks that suggest atypically-reading adults may apply sublexical reading strategies for single word naming, Over time, this strategy can act as a barrier to the formation of strong links between letters and sounds and slow the consolidation of predictable relationships that contributes to the development of fluent and accurate reading. We can assume from this, that the reading strategy was in place during their school years and may explain why, given opportunity and exposure, these individuals' reading skills developed more slowly than their peers. This is a testable hypothesis and with consideration of simulation studies of the division of labour hypothesis using computational models, could form the basis for future research.

Appendix A

Meta-Analysis: Systematic Search Strategy

Updated Search Strategy 2020

Terms for search in March 2020 were developed with the advice of Lancaster University librarian, Jonathan Barbrook.

EbscoHost Search – using Psycholinguistic as a key term:

TI ((psycholinguistic AND (predictor* OR variable* OR effect)) OR "age of acquisition" OR "contextual diversity" OR "word frequency" OR "word familiarity" OR "imageability" OR "concreteness" OR "word length" OR " (neighborhood OR neighbourhood) size" OR "consistency" OR "semantic diversity" OR "sensory experience" OR "valence" OR "regularity" OR "bigram frequency" OR "regular spelling patterns") OR AB ((psycholinguistic AND (predictor OR variable* OR effect*)) OR "age of acquisition" OR "contextual diversity" OR "word frequency" OR "word familiarity" OR "imageability" OR "concreteness" OR "word length" OR " (neighborhood OR neighbourhood) size" OR "consistency" OR "semantic diversity" OR "sensory experience" OR "valence" OR "regularity" OR "bigram frequency" OR "regular spelling patterns")

AND

"word naming" OR "word-naming" OR "word recognition" OR "lexical decision"

→ S1 AND S2 = s3 = 3957 records

"english as a second language" OR "L2"

aphasia OR dysphasia OR "communication disorder"

"alzheimer* disease" OR dementia

→ s4 OR s5 OR s6 = s7 = 355,455 records

→ s3 NOT s7 = s8 = 3666 records

(typical OR normal OR good OR "non disabled" OR "non-disabled" OR nondisabled) AND readers

(old* AND young*) AND adults

(old* AND young*) AND children

→ s9 OR s10 OR s11 = s12 = 172003 records

→ s8 AND s12 = 233 records

EbscoHost through PsychInfo:

- Academic Search Ultimate
- APA PsycInfo
- APA PsycArticles

Appendix B

Meta-Analysis: Included Articles

Table 1

Articles Included in the Meta-Analysis

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Allen et al., 1991	Article	USA	English	High	U	Adult	48	LD	Exp.	Freq., Length	Acc, RT
Allen et al., 1993 (Study 1 & 3)	Article	USA	English	High	U	Adult	40	LD	Exp.	Freq., Length	Acc, RT
Allen et al., 2002 (Study 1: single task condition)	Article	USA	English	High	U	Adult	40	LD	Exp.	Freq.	Acc, RT
Allen et al., 2004	Article	USA	English	High	U	Adult	193	LD	Exp.	Freq.	Acc, RT
Allen et al., 2011 (Exp 1 & 2)	Article	USA	English	High, Unclear	C, U	Adult	40	WN	Exp.	Cons., Freq.	Acc, RT
Araujo et al., 2014	Article	Portugal	Portugese	Unclear	School	Child	37	LD	Ab.	Freq., Length	Acc, RT
Backman et al., 1984	Article	Canada	English	Low	School	Child	112	WN	Ab., Exp.	Cons.	Acc, RT
Baddeley et al., 1982 (Exp 3)	Article	UK	English	Low	School	Child	30 / 45	WN	Ab., Age, Exp.	Image	Acc
Baddeley et al., 1988	Article	UK	English	Unclear	School	Child	32 – 64	WN	Ab., Age, Exp.	AoA, Cons., Freq., Image, Length,	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Balota & Ferraro, 1996	Article	USA	English	High	U	Adult	96	LD	Exp.	Freq.	Acc, RT
Balota et al., 2004	Article	USA	English	Low	C, U	Adult	60	LD, WN	Exp.	Cons., Freq., Image, Length, N-size, Synset	Acc, RT
Barber 2009	Thesis	Canada	English	High	U	Adult	40	LD, WN	Ab.	Freq.	Acc, RT
Barca et al., 2006	Article	Italy	Italian	Low	School	Child	82	WN	Ab.	Freq.	Acc, RT
Barry et al., 2006 (Priming stage)	Article	UK	English	High	C, U	Adult	20 / 19	LD, WN	Exp.	AoA	Acc, RT
Beech & Harding, 1984	Article	UK	English	High	School	Child	92 / 79	WN	Ab.	Cons.	Acc, RT
Ben-Dror et al., 1991	Article	USA	English	Low	U	Adult	38	WN	Ab.	Cons.	Acc, RT
Bosman et al., 2006	Article	Netherlands	Dutch	Low	School	Child	69	LD	Ab., Age, Exp.	Cons., Freq.	Acc, RT
Brown, 1997	Article	UK	English	Low	School	Child	20	WN	Age	Cons.	Acc

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Bruck 1988 (Task 2)	Article	Canada	English	Low	Clinic, School	Child	34	WN	Age	Cons., Freq.	Acc, RT
Bruck 1990 (Section 2)	Article	Canada	English	Low	School, U	Adult	55	WN	Ab.	Cons., Freq.	RT
Burani et al., 2002 (Exp 1 & 2)	Article	Italy	Italian	Low	School	Child	90	LD, WN	Exp.	Freq., Length	Acc, RT
Butler & Hains, 1979	Article	Canada	English	Unclear	U	Adult	12	LD, WN	Ab.	AoA, Freq., Length	RT
Cohen-Shikora & Balota, 2016	Article	USA	English	Low	Database	Adult	148	LD, WN	Exp.	Cons., Freq., Image, Length, Valence	Acc, RT
Coltheart et al., 1988 (Exp 1)	Article	England	English	Low	School	Child	47	WN	Ab.	AoA, Image	Acc
Compton, 1993	Thesis	USA	English	Unclear	School	Child	22	LD, WN	Age	Cons., Freq.	Acc, RT
Davies et al., 2013	Article	Spain	Spanish	Low	School	Child	29	WN	Ab.	Freq., Length	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Davies et al., 2017	Article	UK	English	Low	School, U, C	Adult	219 / 117	LD, WN	Exp.	AoA, BF, Cons., Freq., Image, Length, N-size,	RT
Davies, Cuetos & Glez-Seijas, 2007	Article	Spain	Spanish	Unclear	School	Child	66	WN	Ab., Age, Exp.	Freq., Length, N-size	Acc, RT
De Luca et al., 2008	Article	Italy	Italian	High	School	Child	51	WN	Ab.	Length	RT
De Luca et al., 2010 (Test 5)	Article	Italy	Italian	Low	School	Child	54	WN	Ab.	Length	RT
De Luca et al., 2017	Article	Italy	Italian	Low	School	Adult, Child	76	WN	Ab., Exp.	Freq., Length	RT
Defior et al., 1996	Article	Spain	Spanish	High	School	Child	140	WN	Ab.	Freq., Length	Acc
Deyne & Storms, 2007	Article	Belgium	French	Unclear	School, U	Adult	41	LD	Exp.	AoA	RT
Di Filippo et al., 2006	Article	Italy	Italian	Unclear	School	Child	63	LD	Ab., Exp.	Length	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
DiBenedetto et al., 1983	Article	USA	English	Low	School	Child	40 / 60	WN	Ab., Age, Exp.	Cons.	Acc
Dorot & Mathey, 2010	Article	France	French	High	U	Adult	85	LD	Exp.	AoA, Freq.	RT
Dujardin et al., 2011	Article	France	French	Low	U	Adult	52	LD	Ab.	Freq.	Acc, RT
Dunabeitia & Vidal-Abarca, 2008	Article	Spain	Spanish	Low	School	Child	262	LD	Exp.	N-size	Acc, RT
FRiLL, 2012	Database	UK	English	Unclear	School	Child	61 – 93	WN	Ab., Age, Exp.	Cons.	Acc
Gottardo et al., 1999	Article	USA	English	Low	School	Child	112	WN	Ab., Age	Cons.	Acc
Hautala et al., 2013	Article	Finland	Finnish	Low	School	Child	28	LD, WN	Ab.	Length	Acc, RT
Holligan & Johnston, 1988 (Exp 4)	Article	UK	English	Low	School	Child	40	WN	Age	Cons., Freq.	Acc
Horn & Manis, 1985 (Exp 2)	Article	USA	English	Low	School	Child	36	LD	Ab., Age, Exp.	Freq.	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Hsiao & Nation, 2018 (Exp 1 -3)	Article	UK	English	Low, Unclear, Low	School	Child	35 / 114 / 350	LD, WN	Ab.	AoA, Freq., Length, Semantic diversity,	Acc, RT
Ishaik, 2003	Thesis	Canada	English	High	School	Child	77	WN	Ab.	Cons.	Acc
Jimenez Gonzalez & Hernandez Valle, 2000	Article	Spain	Spanish	High	School	Child	118	LD, WN	Ab., Age, Exp.	Freq., Length	Acc, RT
Johnston et al., 1990	Article	UK	English	Unclear	School	Child	40	WN	Ab., Age, Exp.	Cons., Freq.	Acc
Jorm, 1977 (Exp 2)	Article	Australia	English	Low	School	Child	48	WN	Ab.	Freq., Image, Length	Acc
Jorm, 1981	Article	Australia	English	Low	School	Child	38	WN	Ab.	Cons.	Acc
Keating, 1987 (Exp 8)	Thesis	UK	English	Low	School	Child	60	WN	Exp.	Cons.	Acc
Kitzan et al., 1999 (Exp 1)	Article	USA	English	Low	C, U	Adult	88	LD	Exp.	Synset	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Lavidor et al., 2006	Article	England	English	Unclear	U	Adult	22	LD	Ab.	N-size	Acc, RT
Laxon et al., 1988	Article	UK	English	Low	School	Child	47	LD, WN	Ab.	N-size	Acc
Laxon et al., 1991	Article	UK	English	Low	School	Child	87	WN	Ab.	Cons.	Acc
Laxon et al., 1994	Article	UK	English	Low	School	Child	40	WN	Exp.	Cons., N-size	Acc
Laxon et al., 2002 (Exp 1)	Article	UK	English	Low	School	Child	94	WN	Ab.	N-size	Acc
Leach, 1984 (Task 3)	Thesis	USA	English	Low	FE College	Adult	36	WN	Ab.	Length	Acc, RT
Lewellen et al., 1993 (Exp 1 & 2)	Article	USA	English	Unclear, Low	U	Adult	30 / 70	LD, WN	Ab.	Freq., N-size	Acc, RT
Lovett, 1987	Article	Canada	English	High	Clinic, School	Child	96	WN	Ab.	Cons., Freq.	Acc, RT
Luque et al., 2013	Article	Spain	Spanish	Low	School	Child	158	LD	Ab., Age, Exp.	Freq.	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Macdonald, 2013 (Exp 2 & 3)	Thesis	Canada	English	Low	C, U	Adult	63	LD	Exp.	Semantic Density, Semantic N-size	Acc, RT
Mahe et al., 2012	Article	France	French	High	C, U	Adult	31	LD	Ab.	Freq.	Acc, RT
Mahe et al., 2018	Article	France	French	Low	U	Adult	42	WN	Ab.	Cons.	Acc, RT
Marcolini et al., 2011	Article	Italy	Italian	Low	School	Child	63	WN	Ab.	Freq.	Acc, RT
Marinelli et al., 2011	Article	Italy	Italian	Low	School	Child	65	LD, WN	Ab.	Freq.	Acc, RT
Marinelli et al., 2013	Article	Italy	Italian	Low	School	Child	66	WN	Ab.	Freq., N-size	Acc, RT
Marinelli et al., 2014	Article	Italy	Italian	High	School	Child	41	LD	Ab.	Freq., Length	Acc, RT
Marinus & de Jong, 2010a	Article	Netherlands	Dutch	Low	School	Child	72	WN	Ab., Age, Exp.	Freq., Length, N-size	RT
Marinus & de Jong, 2010b	Article	Netherlands	Dutch	High	School	Child	63	WN	Ab., Age, Exp.	N-size	Acc, RT
Martens & de Jong, 2006	Article	Netherlands	Dutch	High	School	Child	66	LD	Ab., Age, Exp.	Length	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Martens & de Jong, 2008	Article	Netherlands	Dutch	High	School	Child	64	WN	Ab., Exp.	Length	RT
Martens, 2006ba	Thesis	Netherlands	Dutch	High	School	Child	43	WN	Ab., Age, Exp.	Length	Acc, RT
Martin et al., 2010 (Exp 2)	Article	France	French	Unclear	U	Adult	30	WN	Ab.	Length	Acc, RT
Mason, 1978 (Exp 1, 2 & 3)	Article	USA	English	Low	U	Adult	24	WN	Ab.	Cons., Length	Acc, RT
McKoon & Ratcliff, 2016	Article	USA	English	Low	FE College, U	Adult	180	LD	Ab.	Freq.	Acc, RT
Morrison et al., 2002 (Exp 1a & 1b)	Article	England	English	Low	Database, U	Adult	60	WN	Exp.	AoA, Freq., Image	RT
Morrison et al., 2003 (Exp 2a & 2b)	Article	England	English	High	Database, U	Adult	60	WN	Exp.	AoA, Freq., Image, Length	RT
Murphy & Pollatsek, 1994	Article	USA	English	Unclear	School	Child	82 / 130	WN	Ab., Age, Exp.	Cons.	Acc
Murphy et al., 1988	Article	USA	English	Low	School	Child	28	WN	Ab.	Cons.	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Navarro-Pardo et al., 2013	Article	Spain	Spanish	High	U	Adult	80	LD	Exp.	Freq.	RT
Nazir et al., 2003	Article	France	French	High	School	Child	75 / 30	LD	Exp.	AoA	Acc, RT
Olson et al., 1985 (Task D & E)	Article	USA	English	Unclear, High	School	Child	281	WN	Ab., Exp.	Cons.	Acc
Paizi et al., 2013 (Exp 1 – 4)	Article	Italy	Italian	Unclear	School	Child	34	LD, WN	Ab.	Freq.	Acc, RT
Parrila et al., 2007	Article	Canada	English	High	U	Adult	55	WN	Ab.	Cons.	Acc, RT
Perea et al., 2016	Article	Spain	Spanish	Unclear	C	Adult	32	LD	Exp.	Freq.	Acc, RT
Perfetti & Hogaboam, 1975 (Grade 3)	Article	USA	English	Low	School	Child	30	WN	Ab.	Freq.	RT
Provazza et al., 2019	Article	UK	English	Unclear	U	Adult	36	WN	Ab.	Freq., Length	Acc, RT
Raman & Baluch, 2001 (Exp 2)	Article	Turkey	Turkish	Unclear	U	Adult	44	WN	Ab.	Cons., Freq., Image	Acc, RT
Raman, 2000	Article	Turkey	Turkish	Unclear	U	Child	40	WN	Ab.	Image	Acc
Raman, 2011	Article	Turkey	Turkish	Unclear	U	Adult	30	WN	Ab.	AoA	RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Ratcliff et al., 2004	Article	USA	English	Unclear	C, U	Adult	98 / 94	LD	Exp.	Freq.	Acc, RT
Ratcliff et al., 2010b	Article	USA	English	Low	C, U	Adult	85	LD	Exp.	Freq., Length	Acc, RT
Robert & Duarte, 2016	Article	France	French	Low	U	Adult	50	LD	Exp.	No. of features	RT
Roderigo Lopez & Jimenez Gonzalez, 1999	Article	Spain	Spanish	Unclear	School	Child	132	WN	Ab.	Freq., Length	Acc
Roderigo Lopez & Jimenez Gonzalez, 2000	Article	Spain	Spanish	Unclear	School	Child	132	WN	Ab.	Freq., Length	RT
Romani et al., 2008	Article	UK	English	Unclear	C, U	Adult	64	WN	Ab.	Cons.	Acc, RT
Schroter & Schroeder, 2017	Article	German	German	Unclear	School, C	Adult, Child	46	LD, WN	Exp.	AoA, Arousal, Freq., Image, Length, N-size, Valence	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Schwanenflugel et al., 1994 (Exp 2a & 2b)	Article	USA	English	Low, Unclear	School, U	Child	32 / 32	LD	Exp.	Image	Acc, RT
Seidenberg et al., 1985	Article	Canada	English	Low	Clinic, School	Child	51	WN	Ab.	Cons.	Acc, RT
Seymour, 1987b	Article	UK	English	Low	School	Child	22	WN	Ab.	Freq.	Acc, RT
Siegel & Ryan, 1988	Article	Canada	English	Unclear	School	Child	56 / 79 / 66	WN	Ab., Age, Exp.	Cons.	Acc
Spinelli et al., 2005 (Study 1)	Article	Italy	Italian	High	School	Child	84	WN	Ab.	Length	RT
Stanovich et al., 1988	Article	USA	English	Unclear	School	Child	64	WN	Age	Cons.	Acc
Steacy et al., 2017	Article	USA	English	Low	School	Child	170	WN	Ab.	Freq., Image, Length, N-size	Acc
Strain & Herdman, 1999	Article	Canada	English	High	U	Adult	60	WN	Ab.	Cons., Image	Acc, RT
Suarez-Coalla & Cuetos, 2012	Article	Spain	Spanish	Unclear	Clinic, School	Child	38	WN	Ab.	AoA	RT
Suarez-Coalla & Cuetos, 2015	Article	Spain	Spanish	High	C	Adult	60	LD, WN	Ab.	Freq., Length	Acc, RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Szeszulski & Manis, 1987	Article	USA	English	Unclear	School	Child	51 / 34	WN	Ab., Age, Exp.	Cons.	Acc
Tainturier et al., 1989	Article	Canada	English	High	C	Adult	39	LD	Exp.	Freq.	RT
Tainturier, Tremblay & Lecours, 1992	Article	France	French	High	C	Adult	39	LD	Ab.	Freq.	RT
Traficante et al., 2014	Article	Italy	Italian	Low	School	Child	54	WN	Ab.	Freq., Length	Acc, RT
Treiman & Hirsh-Pasek, 1985	Article	USA	English	High	Clinic, School	Child	74	WN	Age	Cons.	Acc
Treiman et al., 1995	Article	USA	English	Low	School	Child	40	WN	Exp.	Cons.	Acc
Verhoeven & Keuning, 2018	Article	Netherlands	Dutch	Unclear	School	Child	3157	WN	Ab., Exp.	Length	Acc, RT
Waters et al., 1985	Article	Canada	English	Low	School	Child	36	WN	Ab.	Cons.	Acc, RT
Weekes et al., 2006: (Exp 1)	Article	UK	English	Low	School	Child	40	WN	Exp.	Cons., Freq.	Acc
Whiting et al., 2003	Article	USA	English	Unclear	C	Adult	24	LD	Exp.	Freq.	RT

Table 1

Articles Included in the Meta-Analysis (continued)

						Sample					
Study Name	Format	Country	Language	RoB	Setting	Type	n	Task	Contrast	Predictors	Outcome
Willcutt, 2008a	Thesis	USA	English	High	C, U	Adult	60	LD	Ab.	Length	RT
Ziegler et al., 2003	Article	France	English, German	Low	School	Child	149	WN	Ab., Age, Exp.	Length, N-size	Acc, RT
Ziegler et al., 2008	Article	France	French	Low	Clinic, School	Child	48	WN	Ab.	Cons.	Acc, RT
Zoccolotti et al., 2005	Article	Italy	Italian	Low	School	Child	37	WN	Ab.	Length	RT
Zoccolotti et al., 2009	Article	Italy	Italian	High	School	Child	503	WN	Exp.	Freq., Length	RT

Note:

Values in the *n* column indicate total sample numbers for each study, separated by a forward slash where appropriate. RoB = Risk of bias. U = University; C = Community; WN = Word naming; LD = Lexical Decision; Ab. = Ability; Exp. = Experience; AoA = age of acquisition; BF = Bigram Frequency; Cons. = Consistency; Freq = Frequency; Image = Imageability; N-size = Neighbourhood-size; Acc = Accuracy; RT = Response time.

Appendix C

Meta-Analysis: Confidence Judgement Process

The summary of findings is comprised of several parts that contribute to an adjudication of confidence for each meta-analysis estimate. In conducting our evaluations for confidence, we were guided by the GRADE process (Cochrane: https://training.cochrane.org/introduction-grade) and their five domains of imprecision, indirectness, inconsistency, publication bias and risk of bias at the outcome level. We explain our operationalisation of these domains and their presentation in the summary of findings next. Data for each of these domains is available at the project webpage.

Imprecision of Summary Effects

Imprecision was assessed using three points of information: (1) RDF results of < 0.8 to replicate d(rep) and the magnitude error > 2; (2) after dividing total participant sample and total number of items, each sample < 40 per group / condition; (3) confidence intervals of the estimated effect sizes cross more than two effect size categories. Two matches in any three contributed to a lowering of confidence by one level.

Power and Magnitude of Error

Power values from the RDF analyses are presented in the summary findings. Sign and magnitude of error data are available at the project webpage. Of the 127 estimated effects, 124 summary estimates showed power levels for replication that were below

80% (RE: 71; FE 52), with 59 of these having power of below 10%. Mean power across all estimated effects is 20.53% (sd = 21.8; RE: 29.05%, sd = 24.52 and FE: 8.24%, sd = 6.38). Ninety-six estimates (RE: 46; FE: 50) showed a magnitude error bigger than twice the size of d(rep), indicating a greater instablity of the meta-analysis finding.

Sample Size

Number of participants, studies and items for each summary estimate is presented in each figure. One hundred and three estimates had discrete groups of less than 40 participants, of which 28 had less than 40 stimuli per predictor condition. In contrast, of the 127 estimated effects, only 32 summary effects show less than 40 items per predictor condition, of which 28 show low participant numbers; of the remaining 95 studies showing adequate levels of item stimuli, 75 also show lower than desired participant numbers. Clearly, studies are more likely to be adequately powered in the item sample than the participant sample.

Confidence Interval Span

Point estimates, with their 95% confidence intervals are presented in each plot. Vertical dashed lines are drawn to indicate values of 0, 0.2, 0.5 and 0.8 to allow readers to judge reliability and the size category for each estimate. Ninety-nine of the 127 estimates' confidence intervals spanned more than two effect size categories (RE: 49; FE: 50). Essentially, a combination of low power, particularly from the low numbers for participant samples, contributed to some very wide confidence intervals.

Indirectness of Summary Effects

Credible intervals estimate predictions of effects for out of sample findings. If credible intervals cross zero, we should exercise caution as to the generalisability of the estimated effect. Credible intervals that cross zero reduce our confidence in the estimate by one level. Credible intervals are estimated at the same level as 95%

confidence intervals for fixed effects models. Of the 75 RE models, 32 of the estimate credible intervals crossed zero, meaning we feel unable to generalise the estimates outside of the present sample. Credible interval data is available at the project webpage.

Inconsistency Within Summary Effects

Heterogeneity values are presented for each summary estimate. We use residual heterogeneity values to define inconsistency and retained the thresholds used throughout the analyses. Consequently, inconsistency was adjudicated as present if, after sensitivity analyses and outlier removal, I^2 values remained high (i.e. $> 75\%$) for a summary estimate. I^2 values are not relevant for FE model estimates.

None of the 75 RE models displayed high inconsistency. Five studies showed low consistency (I^2 values between 25 - 50%), with the remaining 70 showing very low values (i.e. below 25%) and Cochrane's Q p-values greater than .05, indicating random sampling variation as a competing source of variability within sample.

Publication Bias

Egger's Test and Begg's Rank Correlation Test were performed on all full, subgroup samples where there were three or more study effects (n = 53). If either test returned a result $p < .1$, publication bias was indicated as present and confidence was lowered. We present the lowest p-value in the summary findings. Five out of 53 summary estimates showed evidence of publication bias in at least one test, two of the five showed significant p-values for both tests.

Risk-of-Bias Outcome Level Judgements

Risk-of-Bias (RoB) judgements for each estimated effect are presented in the summary findings. We performed a simple counting of RoB judgements across the

individual studies for each subgroup meta-analysis. As noted in the methods section, low and unclear RoB adjudications did not lead to a lowering of confidence, but an adjudication of high RoB did. Very few of the summary effects obtained a "high" RoB adjudication (n = 8). The majority of summary effects were estimated as "unclear" (n = 67), with the remaining 52 studies adjudicated as "low" risk of bias.

Further to this counting we inspected the comments attributed to low, unclear and high judgements within each RoB domain and briefly describe some features of high judgements given to study level evidence. High RoB for the selection domain was often suggested by unequal participant samples at the recruitment phase and then employing an analysis of variance strategy. In the performance domain, a high RoB judgement was given for confounds of ability with age or vice versa between the participant groups that were not corrected for by adding covariates, missingness of randomisation or fixed order of presentation for item samples and instructions to participants that could introduce greater workload for one of the sample groups. Indicators of potential high RoB for the detection domain were mainly around outlier analysis, unequal data trimming practices and removal of participants (sometimes with replacement) or items with no clear indication of how that affected the data sample before analysis for inference. In the reporting domain, notwithstanding changes in reporting standards over time, high RoB was given for selective reporting of results where initial hypotheses and study design explicitly prescribed their inclusion, with no explanation why, or in a few cases, additional analyses with the introduction of a new variable.

Confidence Ratings within Summary Effects

We present our confidence rating for each estimate within the summary findings plot. The GRADE process recommends beginning at higher levels of confidence and using the domain evidence to lower confidence. High levels of confidence indicate a convergence of the analysis estimate with the notion of a "true" effect size. As confidence is lowered, the distance between the estimated effect size and any "true"

effect size increases. Though the GRADE process has four ratings for confidence - high, moderate, low and very low - we operated only three of them as the “high” rating is recommended to be reserved for studies using randomised controlled designs or with blinding mechanisms at the assignment, assessment and analysis stages. No studies use randomised assignment of participants to groups within this sample. Consequently, each summary effect began with a “moderate” rating and the above evidence was taken into account when making a final decision. Of the 127 summary estimates calculated at the time of writing, only 28 estimates retained their moderate confidence rating. Eighteen were reduced one level to “low” confidence and 81 were reduced by two levels to “very low” confidence. Data for this process is available at the OSF project page

Appendix D

Longitudinal Study: Item List Construction

The process of generating four lists of approximate equivalence is described below:

In the first instance, ratings databases were downloaded and merged together to form a composite list of 1903 words that contained measures for all variables of interest arising from the pilot study. This set of words was sorted in an Excel spreadsheet by length and Zipf frequency (SUBTLEX_UK) and words labelled from one to four at the beginning of the low frequency scale and the end of the high frequency scale such that the four lists contained 50 words of three-seven letters. A second list was constructed that was sorted and labelled as a function of length and Contextual Diversity (SUBTLEX_UK). Both of these ratings are from the SUBTLEX corpus. Each set was further decomposed into four sets of 50 words (Zipf1 - Zipf4; CD.1 - CD.4) - each with 25 low and 25 high frequency ratings - as proposed stimuli sets for the word naming and lexical decision tasks across four time points in the longitudinal study. The equivalence of each list for frequency needed to be evaluated.

Lexical Decision and Word Naming Items

In order to test that the distributions of low and high frequency words are distinct within word lists, frequency ratings were partitioned at a Zipf value of 3.5 with values below categorised as low and ratings above categorised as high. To perform a statistical check of difference, a two-sample Kolmogorov-Smirnov test was used (CD list: $p < .001$; Zipf list: $p < .001$). The difference between the low and high frequency values within lists by both frequency scales is signficant. We tested Zipf values across CD and Zipf scale lists using a independent samples t-tests. For both low and high

frequency values, across lists, there were no differences (all ps > 0.07) for both the CD and Zipf lists.

We tested for the equivalence of variance for Zipf values across the CD200 and Zipf200 list. There were no statistically significant differences for variance values across lists 1-4 for low or high Zipf Frequency values (all ps $> .31$).

We repeated the process for contextual diversity values across CD and Zipf ranked lists. To perform a statistical check of difference, a two-sample Kolmogorov-Smirnov test was used (CD list: $p < .001$; Zipf list: $p < .001$). The difference between the low and high frequency values within lists by both frequency scales is signficant. We tested CD values across CD and Zipf scale lists using a independent samples t-tests. For both low and high frequency values, across lists 1-4, there were no differences (all ps $> .37$).

We tested the equivalence of variance values for sets of contextual diversity ratings across lists 1 -4 in the CD ranked list. These results showed that there were statistically significant differences between some of the low lists (1 vs 2 $p < .001$; 1 vs 4 $p = .002$; 2 vs 3 $p < .001$; 2 vs 4 $p < .001$). There were no significant differences for variance of high frequency values across the lists (all ps $> .86$).

Given that the low contextual diversity ratings show a difference in variance, while the Zipf sorted list shows equivalence across lists for means and variances, the Zipf sorted list is chosen as the list to prepare as stimuli for the main study.

Letter Search

Kolmogorov Smirnoff tests for the frequency value distributions are significant and Wilcoxon test aslo confirm that the means between the two categories of frequency are distinct (all ps $< .05$). Within the sentence reading set, there is some overlap between the low and high frequency distributions. Kolmogorov Smirnoff test demonstrates that the distributions are distinct and the Wilcoxon Test demonstrates that the means between the two categories are distinct (all ps $< .05$).

T-tests for low frequency values across the lists confirm that the stimuli items

are drawn from the same population and have equivalent means. The variance within the low frequency distribution is also equivalent. High frequencies across the lists also appear to be normally distributed. The same is true for words with high frequency values.

Sentence Reading

Both low and high frequency density plots display similar trends across lists but bimodal distributions across both categories. KS-tests and Wilcoxon tests revealed that distributions for the sentence reading stimuli were equivalent across lists; the mean values also appear to be similar. Variance values were tested for similarity across the lists and shown to be equivalent.

Appendix E

Longitudinal Study: Item Stimuli

Table 2

Items for Letter Search Task

		List		
Condition	Frequency	1	2	3
word	Low	horde	cleft	twine
		clang	mauve	sneer
		wring	waive	nymph
		scorn	gleam	purge
		scoff	mulch	wreak
		suave	crypt	slant
		shunt	shawl	lunge
		bulge	crepe	plush
		quilt	wield	tract
		snore	shrug	speck
	High	voice	press	doubt
		track	shape	sleep
		force	style	blood
		drive	dream	proud
		trust	stick	earth
		fight	catch	build
		queen	check	board

Table 2

Items for Letter Search Task (continued)

Condition	Frequency	1	2	3
		guess	quick	tough
		prime	stage	third
		state	break	white
non-word		qsmgt	crezb	mcodj
		lpevg	cjfng	joitx
		wqsme	xrizu	tlurc
		hivzt	dntwe	ngfal
		qgahn	gemvr	ykqvd
		sberp	tfyks	gxlnf
		hyueg	owajq	idtzr
		jtyvx	phskw	tnkzg
		uryla	uvkcx	lrpoa
		xqinm	yisce	mhail
		meahf	jwaxs	ktbqz
		yhkqo	inpdb	wdfge
		mnhjd	bsyjf	xyrju
		ntibc	mdfbz	qjzwh
		eghui	kozbg	aerzp
		tvufr	hvsqx	uhdlo
		acous	hvogf	vbcht
		byxwc	lmndb	hqsym
		jqcgb	rvoeq	ymtzd
		yxver	nryjk	gkujx

Table 3

Items for Lexical Decision and Word Naming Tasks

			No. of Letters				
List	Condition	Frequency	3	4	5	6	7 or 8
1	word	Low	sag	curt	scowl	squall	
			jot	slur	taunt	wheeze	
			hoe	lewd	smock	clique	
			pox	moot	drape	crutch	
			oar	lint	gruff	shriek	
				yank	bathe	screech	
		High	bra	bike	phone	drawer	glimpse
			top	team	watch	spring	stretch
			big	play	south	strike	strange
			put	nice	small	search	
			say	year	wrong	choose	
			see	know	right	change	
	non-word		hoa	bign	bidst	baphed	chaides
			oir	brai	culct	drawps	chooged
			seu	jodd	knohm	gruess	clitsch
				mout	lebbs	phoink	crurghs
				pohl	lixth	riqued	drawped
				pufo	nintz	scorde	screuch
				salb	plawp	smayes	searned
				sato	shraik	smonde	spriscs
				tohl	slurg	sourth	squayer
					teapt	taphed	stribid
					yaubs	waides	wheefed
					yeaux	wrorne	glirched

Table 3

Items for Lexical Decision and Word Naming Tasks (continued)

List	Condition	Frequency	3	4	5	6	7 or 8
							strawped
							strelfth
2	word	Low	keg	wisp	snide	broach	
			pry	hick	spunk	hoarse	
			lax	daze	braid	thrift	
			fad	reek	whine	quench	
			pew	husk	smirk	sphinx	
						scourge	
						fraught	
		High	flu	trot	midst	scheme	scratch
			pop	must	stove	speech	thought
			job	mean	price	square	
			try	tell	north	church	
			man	make	close	street	
			day	like	place		
			let		three		
			one				
3	non-word		pri	dalc	darcs	branst	broafed
			tra	fadv	hibid	cloilt	churghs
				fluv	hulct	misles	hoathes
				joif	lixth	nowths	quevved
				kegm	madze	plarcs	schewts
				laxe	meaut	pridth	speethe
				legg	murke	smilns	sphiscs
				malb	reeze	snixth	squayes

Table 3

Items for Lexical Decision and Word Naming Tasks (continued)

List	Condition	Frequency	3	4	5	6	7 or 8
				onde	telte	spuess	thrixth
				pekg	threo	stoifs	frarques
				poif	tronj	struet	scounced
					wicbm	whidst	scraphed
							thouache
	word	Low	orb	gush	knoll	scorch	
			hag	lisp	snarl	clench	
			wad	kink	taint	clothe	
			sob	snub	girth	soothe	
			coy	brig	trite	squint	
						cleanse	
						breadth	
		High	lab	pail	depth	prayer	breathe
			may	week	grand	threat	strength
			old	four	piece	bridge	
			new	sure	heart	bright	
			two	find	whole	league	
			can	need	round	health	
				time	world		
	non-word			caye	briud	delfth	brinxed
				corb	filst	gitsch	britzed
				haxe	foubt	grayer	clelfth
				lalc	gulge	headth	clondes
				maif	kilst	knorne	heansed
				necd	lixth	pieled	learled

Table 3

Items for Lexical Decision and Word Naming Tasks (continued)

List	Condition	Frequency	3	4	5	6	7 or 8
				nued	parcs	rourns	praults
				onde	snufo	snarcs	scoynes
				onne	suoys	taults	sooched
				sohl	tibid	tricbm	squilge
				twoz	weeze	whorde	threamt
				wadv		worpse	brearled
							brearths
							clearled
							strerthed

Table 4

Items for Sentence Reading Task

		No. of Letters			
List	Frequency	3	4	5	6
1	Low		mule	leash	blouse
			shin	crate	bruise
					thorns
	High	bun	ship	plane	shrimp
		dam	card	clock	mousse
		gun	dish		sponge
		hat			
		leg			
2	Low	ape	tire	latch	spleen

Table 4

Items for Sentence Reading Task (continued)

List	Frequency	3	4	5	6
			clam	moose	stripe
					stairs
	High	gum	meat	paint	thread
		arm	ring	dress	branch
		bus			breast
		cat			
3	Low	mop	yarn	sloth	wrench
		fig	crib	shack	fleece
				prune	sleeve
	High	rug	wood	bread	script
		van	bird	fruit	throat
		pot	bear		
			foot		

Appendix F

Longitudinal Study: Missing Data Process

We used random regression imputation (Gelman et al., 2021) to impute values for the 22 spelling and 27 vocabulary values. Missing data is especially problematic within this dataset as a repeated measures design. Any empty cell for one missing ID score is propogated across each observation for each item in the experimental tasks, creating multiple rows with missing data. Where the computer algorithm detects a missing datum, it will silently drop the trial level observation from the analysis, resulting in loss of all information for that trial.

Solutions are to drop entire variables or entire participants from the analyses. This results in a deleterious loss of information and loss of statistical power. Rather than omit these observations, we inspected the patterns of missingness and, given the percentages of missing data found, used data imputation to replace missing data values to give a complete set of ID measure scores.

Our approach was the following: First, we considered the possible mechanism for the missingness of the data. We plotted the pattern for the missingness of the data and found that it followed a connected and general pattern (van Buuren, 2018). This means that scores from other ID measures are present by which missing data can be estimated. Since all other information for these observations are fully recorded, we can assume that the mechanism for the missingness is 'missing at random'. Second, since these data are longitudinal by design, they are structured by time. This means the order of missing data imputation is critically important. If later data is dependent upon earlier data, then missingness on earlier data needs to be inspected first before treating later instances of missingess (Enders, 2010).

Below we describe our treatment for missing data according to the rate of

missingness at each time point. After treatment, we conducted *t*-tests between raw and imputed ID measure data across all three time points and found no statistically significant differences between the two sets (all *p*s > .7).

Imputing Missing Values for T1

Missing data is present for spelling (n = 22 / 218; 10%) and vocabulary scores (n = 27 / 218; 12%). This is due to an error on the administration of the tests to a subset of classes by the class teachers and participant absence on the day of testing. We used random regression imputation (Gelman et al., 2021) to impute values for the missing values.

Spelling scores were regressed upon predictors for gender, institution, word reading accuracy, nonword reading accuracy, processing speed, phonological awareness, type of group, skill and age. Predictions from that model were generated, with the error term from the predictions being added to the imputed values to give back some uncertainty to the values, in an attempt to mitigate bias. This procedure was repeated for vocabulary missing values.

Imputing Missing Values for T2 and T3

At T2, one participant (1 / 191; 0.5% per ID measure) completed experimental tasks but not ID measures due to a fire drill during the session. A further participant did not complete the spelling test on the day of testing due to absence from school. At T3, one participant (1 / 173; 0.5%) did not complete ID measures at T3 due to timing difficulties on the day of testing. Five participants (5 / 173; 2.8%) did not complete the spelling measure and four participants (4 / 173; 2.3%) did not complete the vocabulary measure. This is due to absence from class on the day that the measures were administered to the 11-12 year olds participants. At T2 and T3 the missing data rates per ID measure are all below 5%.

Due to the low rate of missingness and the mechanism of missingness still being categorised as missing at random, we use single value random sampling to

impute values for these participants (as described in Gelman et al., 2021). Using the `sample()` function, we randomly sampled from the range of values within each variable at T2 for missing data values at T2, repeating the procedure for variable values at T3 for missing data values at T3. As mentioned above, when we tested for differences between the data with missing values and the data with imputed values, there were no statistically significant differences between the datasets (all ps $> .7$). We present visualisations and results from the imputed dataset from this point forward.

Appendix G

Longitudinal Study: Spelling Error Analysis

Spelling Error Analysis

The lexical quality hypothesis says that words that are consistently spelled correctly index high lexical quality, with the inverse being true. Martin-Chang et al. (2014) measured standard spelling in the traditional correct / incorrect sense but also the variability of a person's spelling errors over repeated assessments. They found those words that were incorrectly spelled but consistently (mis)spelled within a participant, were named faster than those inaccurately spelled words that varied in the type of spelling errors. The faster response times for in incorrect word suggests that an incorrect spelling when believed to be correct may still have high lexical quality within an individual, relative to other words in their vocabulary.

We took the errors from the data and explored the nature of the recorded answers. We looked at which group was most likely to leave out an spelling item, rather than make an attempt. We looked at the occurence of real word substitutions for answers and whether the probabilty of supplying a real word was related to the target word's status as a homophone. Irrespective of homophone status, we also explored whether the supplied answer was an attempt to sound like the target word.

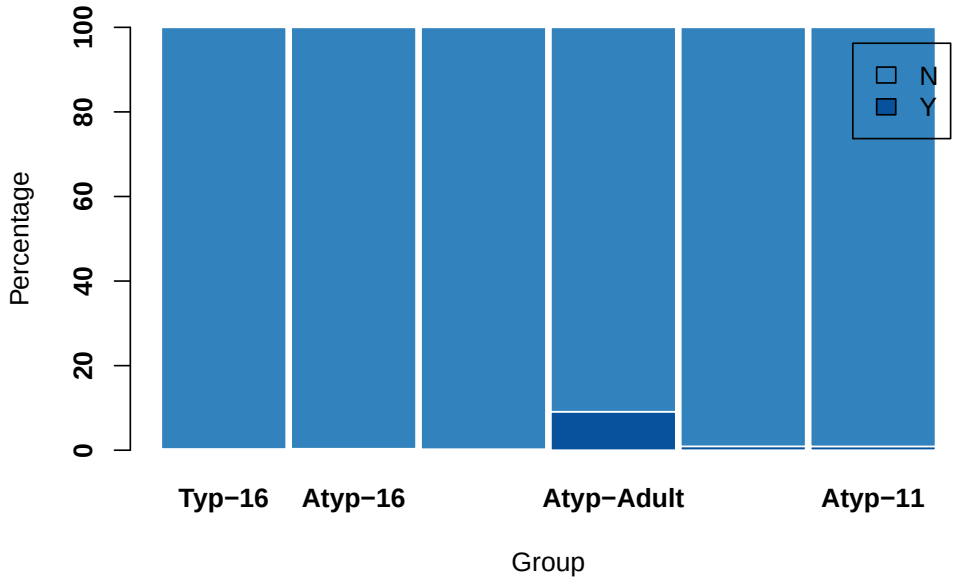
Finally, we charted errors across time. Where errors were made on every occasion, we explored how similar those errors were compared to the target word and also how similar they were to each other.

Omitted Answers. We observed 6455 spelling errors across all participants for three time points. Of those 6455 observations, 116 were missing answers. We ran a

logistic regression with missing status as a binary outcome variable (Yes / No) and group as the dependent variable (reference level = typical 16 year olds). Figure 1 displays the probabilities of choosing to omit an answer by group. Only the atypically-reading adults showed significantly lower odds for supplying an answer than the typically-reading 16-17-year-old participants ($\beta = 3.54$, SE = 0.72, $p < .001$). On a probability scale, there is a probability of 0.29% that typically-reading 16-17-year-olds will omit an answer compared to a 9.1% probability for atypically-reading adults.

Figure 1

Omitted Answers (yes) vs Not Omitted Answers (no) by Group



Real-Word Substitutions. Of the 6339 observed spelling errors, 1041 were real-word substitutions for the target word. We coded the errors as either real-words (Yes) or not real-words (No) and used this variable in a logistic regression to estimate the odds of real-word substitutions as errors as a function of group (reference level = typically-reading 16-17-year-olds). We also constructed a variable that categorised the target word as either a homophone (Yes) or not a homophone (No) to model whether the odds of a real-word being substituted as an answer was related to the target word having multiple spellings that sounded the same.

The best fitting model contained independent effects of group and homophone status plus the interaction term between group and homophone status (AIC = 4286.9; see Table 5). Typically-reading 16-17-year-olds showed a 6.9% probability of substituting the target word with a real-word when the word was not a homophone. The only group that showed a statistically significant difference from this rate of errors was the typically-reading adults (2.9%, $p = .015$). Essentially, most groups were just as likely as each other to mistakenly write real-words for target words that were not homophones.

The change in probability of substituting a real-word once the target word was categorised as a homophone was substantial. Target words that were homophones had higher odds of a real-word substitution being given as an answer than non-homophones. Typically-reading 16-17-year-olds were 67.5% more likely to substitute a real-word if the target word was categorised as a homophone ($p < .001$). Also, the change from non-homophones to homophones did see some group differences. Atypically-reading adults were 54.7% more likely ($p = .061$) and 16-17-year-olds were 51.5% more likely ($p = .026$) to (incorrectly) give a real-word as an answer. The change in atypically-reading adult rates is non-significant which implies that that, overall, atypically-reading adult spelling behaviour for real-word substitution is more similar to the typically-reading 16-17-year-old readers spelling behaviour. Typically-reading adults moved from 2.9% to 75.6% for real-word spelling errors ($p = .005$). Typically-reading 11-12-year-olds moved from 6.1% to 41.8% ($p = .001$) and atypically-reading 11-12-year-olds moved from 6.1% to 34.5% more likely ($p = <.001$) to give real-word spelling error answers if the target word was a homophone. The rates for the 11-12-year-old readers are significantly lower than the rates of real-word substitutions for homophonic target words for typically-reading 16-17-year-old readers.

Figure 2 shows the percentages of errors by group for nonword errors (NA), and the split between homophones (Y) and non-homophones (N) for real-word substitutions as errors.

Errors that are similar in sound to the the target word. Steacy et al. (2017b)

Table 5

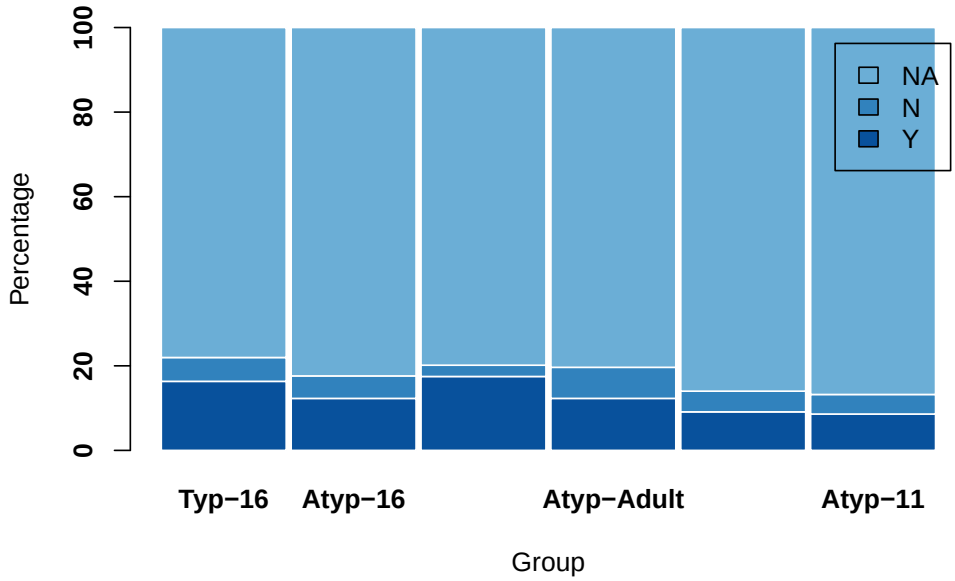
Summary of Estimates for Group and Their Likelihood of Making Real-Word Substitutions as a Function of Whether the Word is a Homophone

Term	lOR	sd	t	p
Intercept	-2.60	0.17	-15.04	0.000
Atypical-16	-0.01	0.22	-0.04	0.971
Typical-adult	-0.92	0.38	-2.43	0.015
Atypical-adult	0.06	0.23	0.28	0.783
Typical-11	-0.13	0.21	-0.63	0.531
Atypical-11	-0.14	0.21	-0.65	0.515
Homophone	3.33	0.24	14.02	0.000
Atypical-16:Homophone	-0.66	0.30	-2.22	0.026
Typical-adult:Homophone	1.32	0.48	2.78	0.005
Atypical-adult:Homophone	-0.60	0.32	-1.87	0.061
Typical-11:Homophone	-0.93	0.29	-3.27	0.001
Atypical-11:Homophone	-1.23	0.28	-4.33	0.000

Note:
lOR = log-odds ratio

Figure 2

Percentage of Non-Real-Word (NA) and Real-Word Substitutions, Conditioned on Their Status as a Homophone of the Target Item (Y / N) by Group

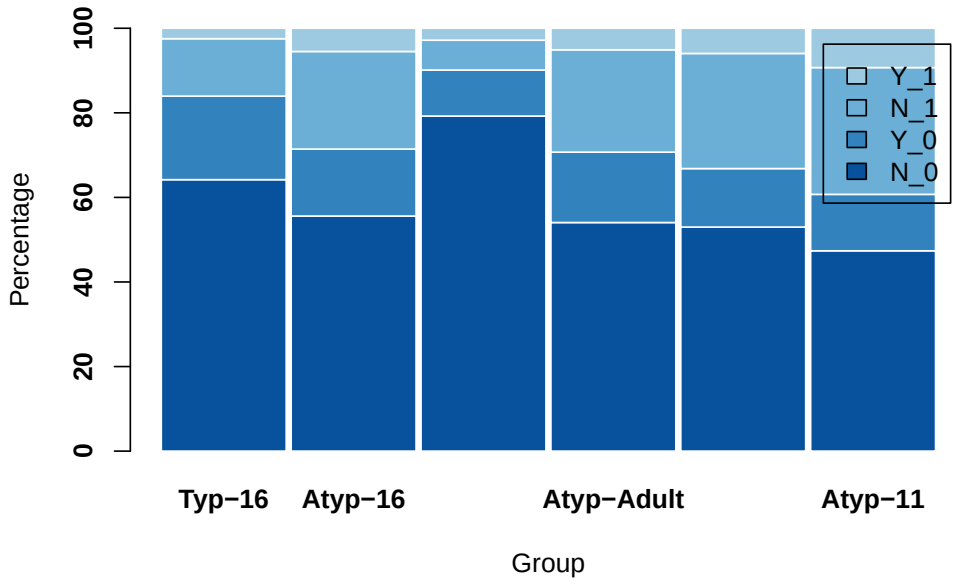


found that spelling (orthographic choice task) + a string distance measure of spelling answers to correct spelling forms was implicated in their sample of 11-year-old readers showing signs of late emerging reading difficulties. We analysed errors using sound as a matching criterion. The function `soundex` from the `stringdist` package (van der Loo, 2014) constructs the phonetic code of the target word and the error and returns `0` for a match between the two phonetic codes and `1` for a non-match. We then tabulated the errors by the homophone status of the target word (Yes / No) and whether soundex code for the error matched the soundex code for the target word (0 / 1). Figure 3 shows the observed percentages of the split across errors by group. From the figure it is clear that both the typical-16-year-old readers and adult readers are more likely to produce sound matching errors than non-sound mathcing errors, irrespective of whether the word is a homophone or not. Both the atypical-16-year-old reading groups and the atypically-reading adults are reduced in their split for sound matched errors, with approximately a third of their errors being scored as a non-match for sound with the target word. This distribution looks similar to that of the 11-12 year old groups, who we have already discussed as being very naive for this measure. Given their greater exposure, the ability of the atypical-16-year-olds and adults to approximate the sound form of the target word looks under-developed. Coupled with the findings from the real-word substitutions, this may suggest a particular phonetic code weakness for the atypically-reading adults, relative to their real-word substitutions, as above they showed equivalent performance with the typical-16-year-old real-word spelling error rates, here they are weaker.

Across Time. We looked at errors across data collection sessions to explore whether error spellings were consistent or not. Due to attrition, 25 participants were excluded from this analysis because they completed one time point only. We further separated participants who completed only two data collection points ($n = 33$) from those that completed three ($n = 160$). We also removed any observation that included an omitted answer. We scored each time point for correct or incorrect answers and then removed observations where a correct answer was present, leaving only

Figure 3

Percentage of Errors that Match the Target Word for Sound (0) as a Function of Whether the Target Word is a Homophone (Yes)



observations that included errors across each time point. We compared spelling errors for orthographic and phonetic (as measured by `soundex`) similarity.

Orthographic similarity. Across two time points, 33 participants generated 459 errors. Where answers were omitted, the observation was removed ($n_{T1} = 27$; $n_{T2} = 5$). Of the remaining 427 observations, 318 spellings were incorrect across both times.

Across three time points, 160 participants made 2121 errors. Omitted answers ($n_{T1} = 4$; $n_{T2} = 5$; $n_{T3} = 1$) were further removed. Of the remaining 2111 observations, 1251 were incorrect across the three time points.

Figure 4 plots the percentage data for groups that repeatedly spelled words incorrectly over two time points (left) and three time points (right), and whether the spelling of those errors was consistent or inconsistent with each other across each occasion.

In these figures, the height of the dark blue portion reflects the percentage of errors that were inconsistently spelled across time points. Due to their naivety, we

would expect the 11-12-year-old groups to have a higher proportion of different spellings. When looking at the atypically-reading adults bar, they are the same height as the 11-12-year-old group bars, reinforcing the suggestion of spelling skills being under-developed. With respect to the lexical quality hypothesis, atypically-reading adults are showing under-specified representations of words for the orthographical dimension of word properties.

Figure 4

Matched and Non-Matched Spelling Errors Across Two (left) and Three (right) Occasions by Group.

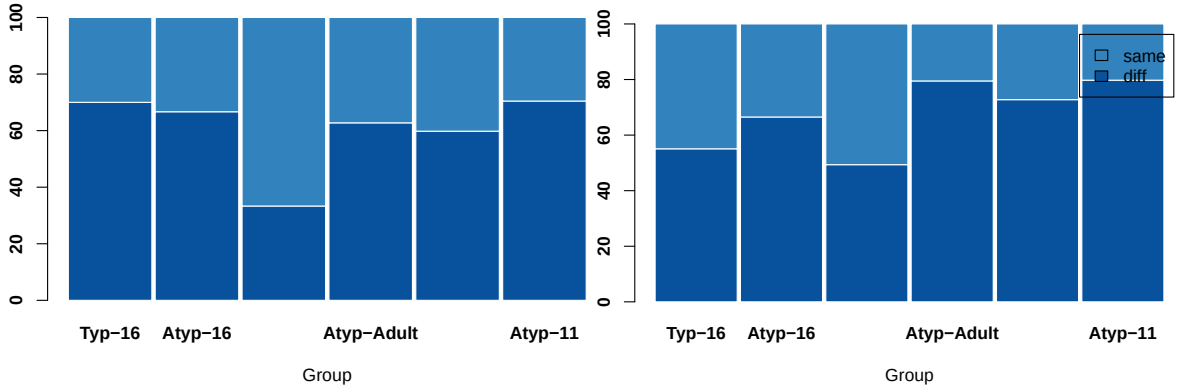
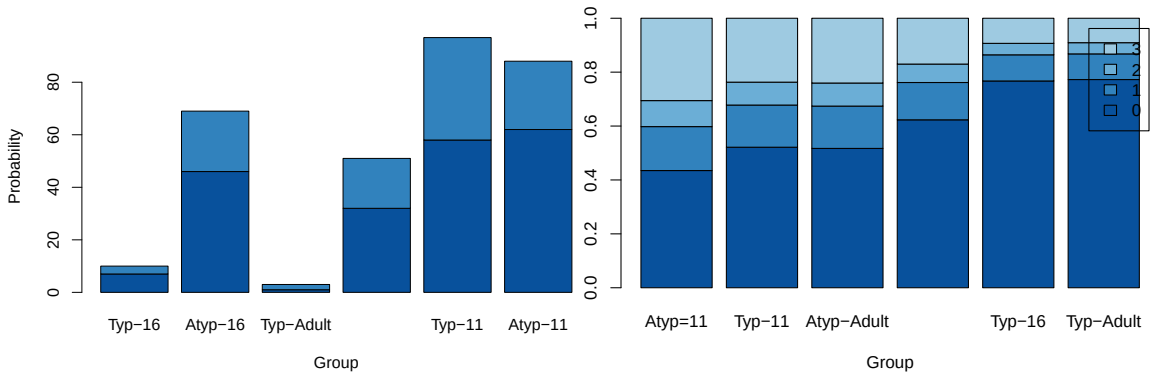


Figure 5

Matched and Non-Matched Values for Phonological Similarity of Errors to Target Word Across Two (left) and Three (right) Occasions by Group.



Phonological similarity. We summed `soundex` values across time points. Consequently, a value of 0 means that each error maintained fidelity with the target word phonetic code; a value of 1, 2 or 3 the sum of how many times any error differed in phonetic code from the target word over a maximum of three time points. We removed the typically-reading adult group from the two time point data as their error rate ($n = 3$) was extremely low, and each observation comprised a sound match. An ordinal regression with summed `soundex` values (min = 0, max = 2) regressed onto group found no significant differences across groups for the log-odds of making two errors that sound different from the target word. The change in log-odds between 0 and 1 sound difference to two sound differences approached significance (log-odds = 1.37, se = 0.70, $p = .051$).

The ordinal regression for `soundex` values across three time points (min = 0, max = 3) returned significant estimates for differences across groups and also changes in log-odds for number of differences in sound across the time points. Only the typically-reading adults showed a non-significant estimate compared to the typically-reading 16-17-year-olds. All other groups differed significantly. The probability of errors sharing the same sound code as the target word across all three time points was ~76%. Of the remaining categories, there was a 10% probability of one or all three of the errors sounding different; there was a 4% probability of at least two of the errors sounding different. Figure 5 shows the differences in probabilities across the number of sound codes generated by group. The plots are ordered so that as you move from left to right, within each plot, there is higher fidelity of sound between the errors and the target word. The atypically-reading adults resemble the younger readers very closely. The typically-reading adults and the typically-reading 16-17-year-old readers are also very similar in the distribution of their sound values and show a much smaller proportion of their errors as generating inconsistent sound codes across time.

Taken together, the atypically-reading adults appear to be very like the youngest participants when they do not know how to spell a given word. Although they scored significantly higher than the youngest readers when they are correct, their

use of orthographic knowledge and phonological skills to approximate a spelling for an unfamiliar word resembles that of relatively inexperienced readers.

Appendix H

Longitudinal Study: Modelling Strategy and Information Criterion Values for Accuracy and Reaction Time Models

Table 6

Overview of Modelling Strategy for Generalised Linear Mixed Models for Accuracy Outcomes

		Effect			lme4		brms			
Model	Nested Model	Fixed	Subjects	Items	AIC	AIC:Group	LOOIC-s	LOOIC-s:Group	LOOIC-w	LOOIC-w:Group
Letter Search										
Base-RI		Position + Days + ID	Intercept	Intercept	7459	**7453**	7282.9	7279.6	7283.4	7280.1
Base-RIS	Base-RI		+ ID	+ ID	8048	7820	7159.8	**7158.8**	7162.3	7161.3
Additive-RI	Base-RIS	+ PV			7472	7466	7285.1	7283.4	7285.6	7283.8
Additive-RIS	Additive-RI		+ PV	+ PV	8233	8719	7181.2	7180.4	7187.2	7186.2
Interaction-RI	Additive-RIS	+ ID * PV			7677	NC	7529.2	9775.1	7532.6	10718.9
Interaction-RIS	Interaction-RI				NC	NC				
Lexical Decision										
Base-RI		Days + ID	Intercept	Intercept	23929	23912	23074.1	23064	23075.4	23065
Base-RIS	Base-RI		+ ID	+ ID	23771.87	23826.81	21971.5	21975	21977.7	21982
Additive-RI	Base-RIS	+ PV			23684	23667	23053.3	23041	23054.1	23043
Additive-RIS	Additive-RI		+ PV	+ PV	NC	NC	**21730.1**	21733.4	21741.6	21744.6
Interaction-RI	Additive-RIS	+ ID * PV			**23339**	NC	21881.4	21898.8	21898.8	21881.4
Interaction-RIS	Interaction-RI				NC	NC	22727.7	23727.2	22729.7	23875.4
Word Naming										
Base-RI		Days + Onsets + ID	Intercept	Intercept	11928	11897	11370.8	11359	11374.2	11362
Base-RIS	Base-RI		+ ID	+ ID	12032.11	11907.11	10879.8	10876	10891.3	10887
Additive-RI	Base-RIS	+ PV			11806	**11774**	11354.7	11342	11357.1	11344.6
Additive-RIS	Additive-RI		+ PV	+ PV	NC	NC	10758.6	**10757.7**	10775.6	10773.5
Interaction-RI	Additive-RIS	+ ID * PV			11797.47	NC	11366.3		11371.7	
Interaction-RIS	Interaction-RI				NC	NC	10959.5		10993.3	
Sentence Reading										
Base-RI		Context + Days + Onsets + ID	Intercept	Intercept	7347	7337	7064.8	7066.3	7066.4	7068
Base-RIS	Base-RI		+ ID	+ ID	7238.6	**7160.3**	6669.9	6673.8	6678	6682.1
Additive-RI	Base-RIS	+ PV			7314	NC	7064.7	7066.4	7066.4	7068.1
Additive-RIS	Additive-RI		+ PV	+ PV	7408.4	4797	**6532.9**	6538.2	6559.3	6565.4
Interaction-RI	Additive-RIS	+ ID * PV			NC	NC	7860.5		7945.7	
Interaction-RIS	Interaction-RI				NC	NC				

Note:
RI = Random intercepts models. RIS = Random intercepts and slopes models. ID = Individual difference measures. PV = Psycholinguistic variables. AIC = Akaike information criterion. LOOIC-s = PSIS-Leave-one-out information criterion for strongly informative priors models. LOOIC-w = PSIS-Leave-one-out information criterion for weakly informative priors models. Group = information criterion for the models that include the planned contrast variables for Group. NC = Non-convergence.

Table 7

Overview of Modelling Strategy for Linear Mixed Models for Reaction Time Outcomes

		Effect			lme4		brms			
Model	Nested Model	Fixed	Subjects	Items	AIC	AIC:Group	LOOIC-s	LOOIC-s:Group	LOOIC-w	LOOIC-w:Group
Letter Search										
Base-RI		Position + Days + ID	Intercept	Intercept	24855	**24816**	23560.2	23554.1	23561.5	23555.5
Base-RIS	Base-RI		+ ID	+ ID	25733	26336	**23286.6**	23288.8	22847.7	22851.7
Additive-RI	Base-RIS	+ PV			24927	24888	23561.4	23555.3	23562	23556.1
Additive-RIS	Additive-RI		+ PV	+ PV	28482	28977	22861.3	22867.3	22873.7	22879.3
Interaction-RI	Additive-RIS	+ ID * PV			27009	34348	23719.3		23725.3	
Interaction-RIS	Interaction-RI				30286.71	NC	23055.7		23087.5	
Lexical Decision										
Base-RI		Days + ID	Intercept	Intercept	81885	81859	77296.5	77292.2	77297	77292.8
Base-RIS	Base-RI		+ ID	+ ID	84734	85012	75662.7	75669.4	75667.1	75673.7
Additive-RI	Base-RIS	+ PV			81672	**81645**	77271.6	77267.1	77272	77267.6
Additive-RIS	Additive-RI		+ PV	+ PV	90693.75	91870.12	**75313.8**	75320.2	75327.2	75334.3
Interaction-RI	Additive-RIS	+ ID * PV			83792	92426				
Interaction-RIS	Interaction-RI				NC	NC				
Word Naming										
Base-RI		Days + Onsets + ID	Intercept	Intercept	98989	98991	88272.3	88272.8	88270.7	88271.3
Base-RIS	Base-RI		+ ID	+ ID	98288.06	**98182.61**	84223.2	84231.5	84225.5	84234.6
Additive-RI	Base-RIS	+ PV			98939	98941	88265.7	88266.3	88263.5	88264.1
Additive-RIS	Additive-RI		+ PV	+ PV	107709.6	105184.8	83347.1	83376	**83344.2**	83373.9
Interaction-RI	Additive-RIS	+ ID * PV			100346	108727				
Interaction-RIS	Interaction-RI				NC	NC				
Sentence Reading										
Base-RI		Context + Days + Onsets + ID	Intercept	Intercept	**87074**	87077	81965.5	81962	81965.7	81962.2
Base-RIS	Base-RI		+ ID	+ ID	88112	88057	80087	80088.2	80088.5	80089.6
Additive-RI	Base-RIS	+ PV			87131	87134	81964.2	81962.1	81964.4	81962.3
Additive-RIS	Additive-RI		+ PV	+ PV	99340.03	99138.69	**79926.7**	79929.2	79932.9	79935.2
Interaction-RI	Additive-RIS	+ ID * PV			92914	115123				
Interaction-RIS	Interaction-RI				NC	NC				

Note:
RI = Random intercepts models. RIS = Random intercepts and slopes models. ID = Individual difference measures. PV = Psycholinguistic variables. AIC = Akaike information criterion. LOOIC-s = PSIS-Leave-one-out information criterion for strongly informative priors models. LOOIC-w = PSIS-Leave-one-out information criterion for weakly informative priors models. Group = information criterion for the models that include the planned contrast variables for Group. NC = Non-convergence.

Appendix I

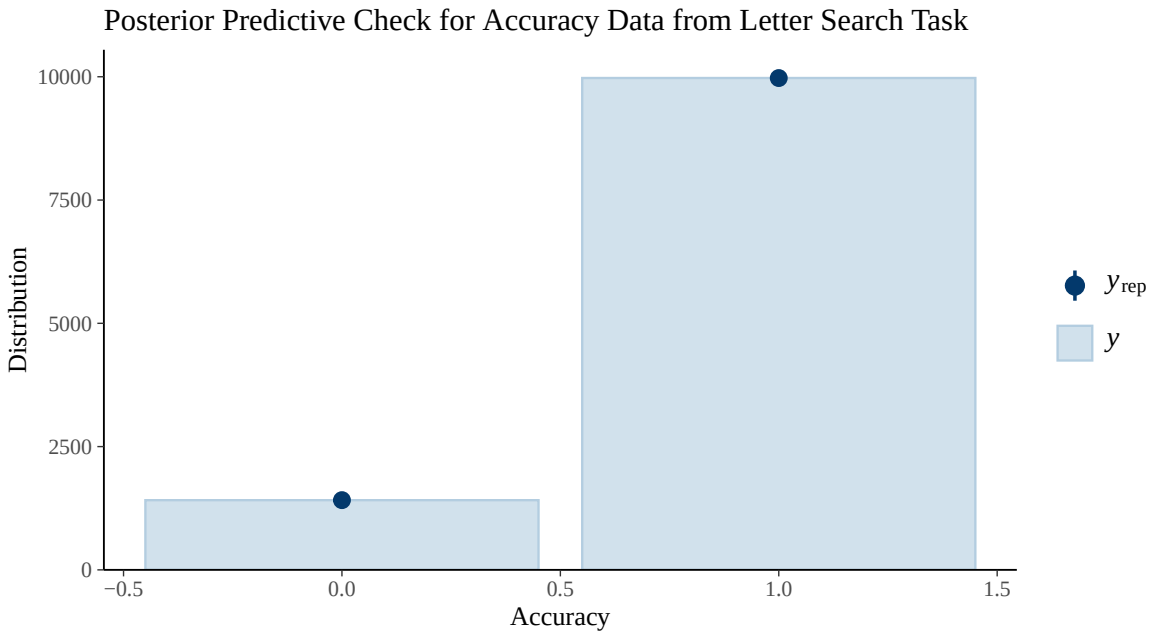
##Longitudinal Study: Model Diagnostic Information {-}

Letter Search

Accuracy. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. Figure 6 shows how the posterior predictive check (PPC) estimates reflected the empirical data. The model implied estimates for accurate and inaccurate responses (blue dots at the top of each bar) are in line with the observed rates of accurate and inaccurate responses (light blue bars).

Figure 6

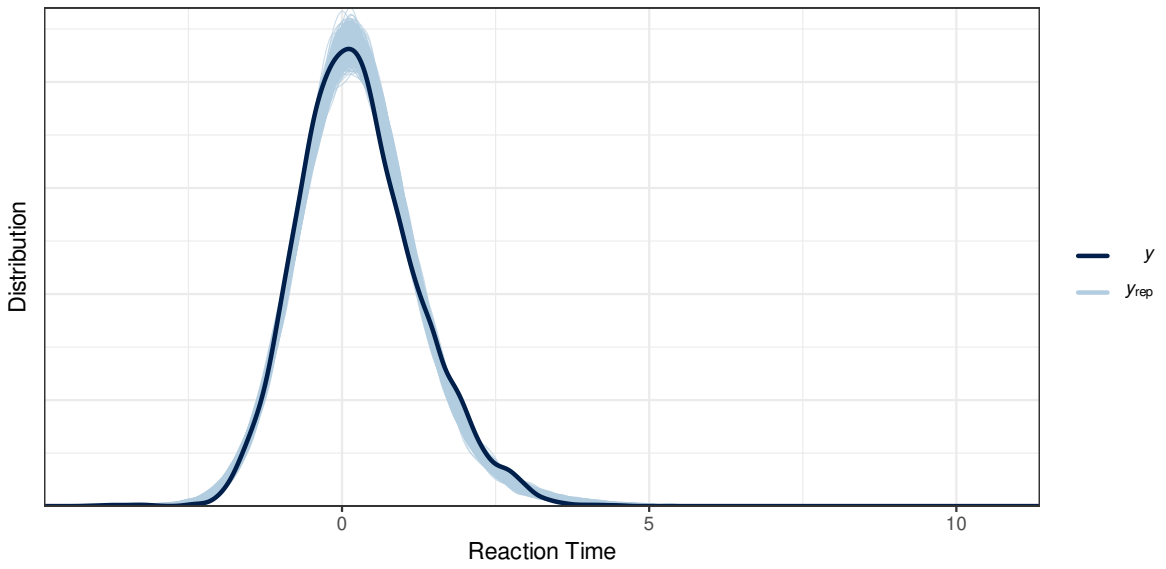
Posterior Predictive Check for Accuracy Data in the Letter Search Task



Reaction Time. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. The PPC for the model displayed a satisfactory fit of the posterior distribution estimates compared to the observed data (see Figure 7). The grey lines resulting from the markov chain sampling process clearly follow the darker line of the observed reaction time distribution.

Figure 7

Posterior Predictive Check for Reaction Time Data in the Letter Search Task

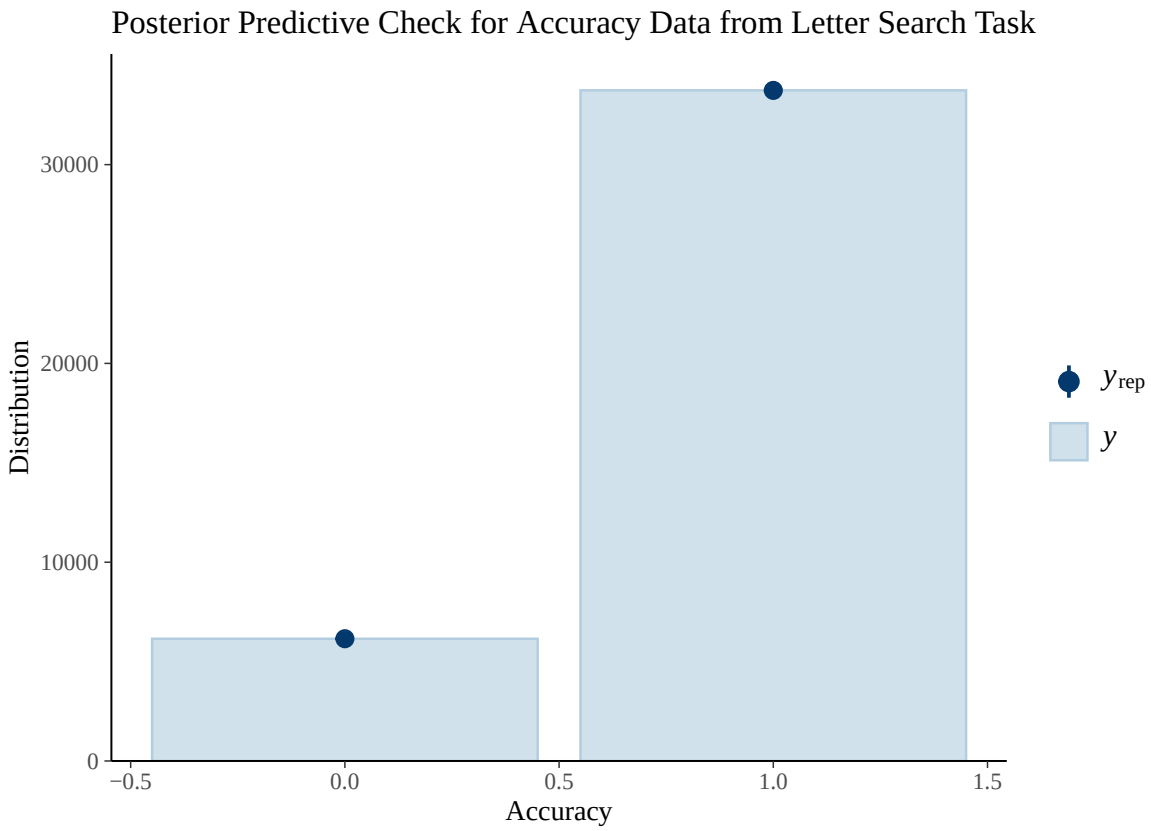


Lexical Decision

Accuracy. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. Figure 8 shows how the PPC estimates reflected the empirical data. The model implied estimates from the PPC; accurate and inaccurate responses are in line with the observed rates of accurate and inaccurate responses.

Figure 8

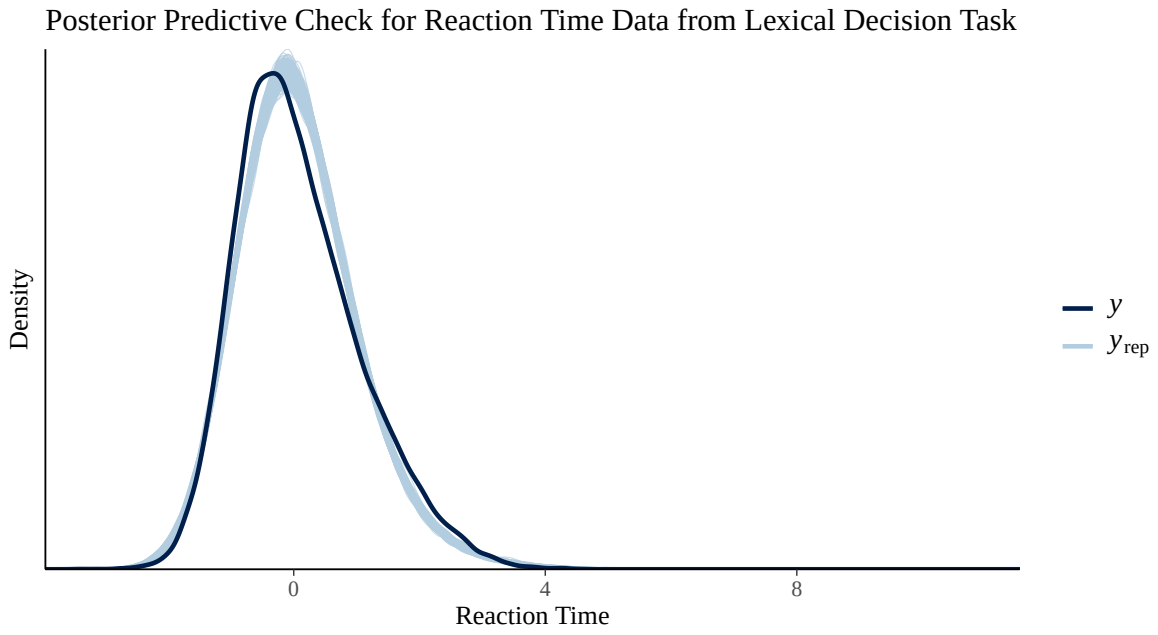
Posterior Predictive Check for Accuracy Data in the Lexical Decision Task



Reaction Time. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. The PPC for the model slightly underestimates the observed data. Figure 9 shows the grey lines of the markov-chain sampling process shifted slightly to the right of the leading edge of the curve for the observed data, and slightly to the left in the tail of the distribution. Essentially, the posterior distribution of the model is estimated with a narrower spread of values than those observed.

Figure 9

Posterior Predictive Check for Reaction Time Data in the Lexical Decision Task

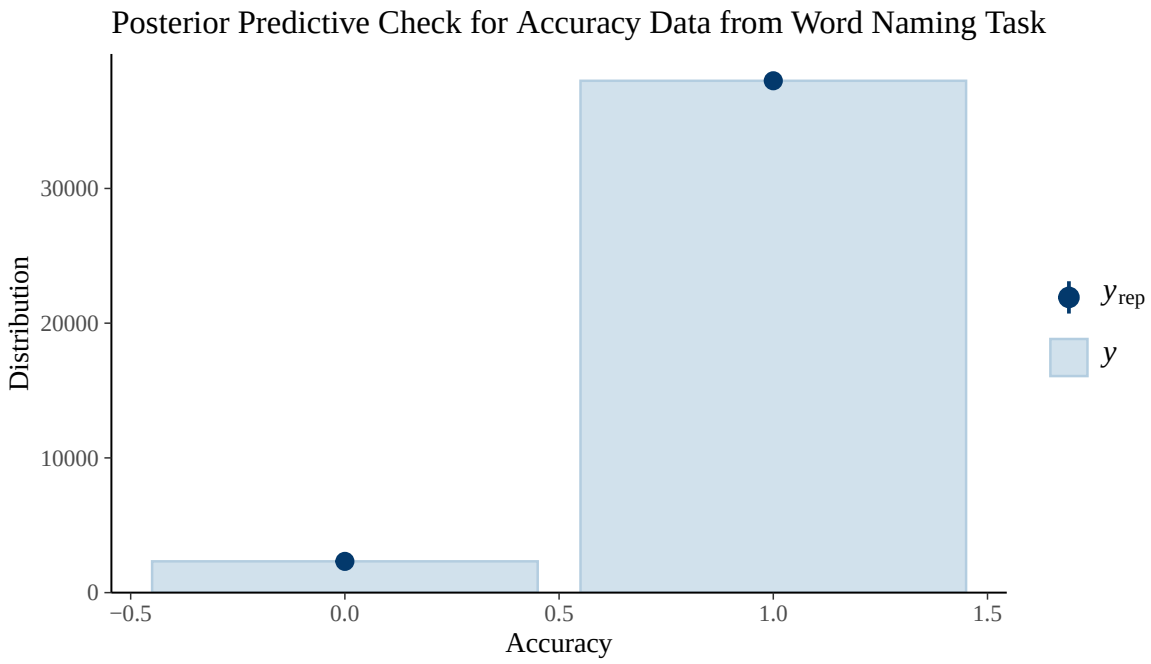


Word Naming

Accuracy. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. Figure 10 shows that the model implied estimates for accurate and inaccurate responses from the PPC are in line with the observed rates of accurate and inaccurate responses.

Figure 10

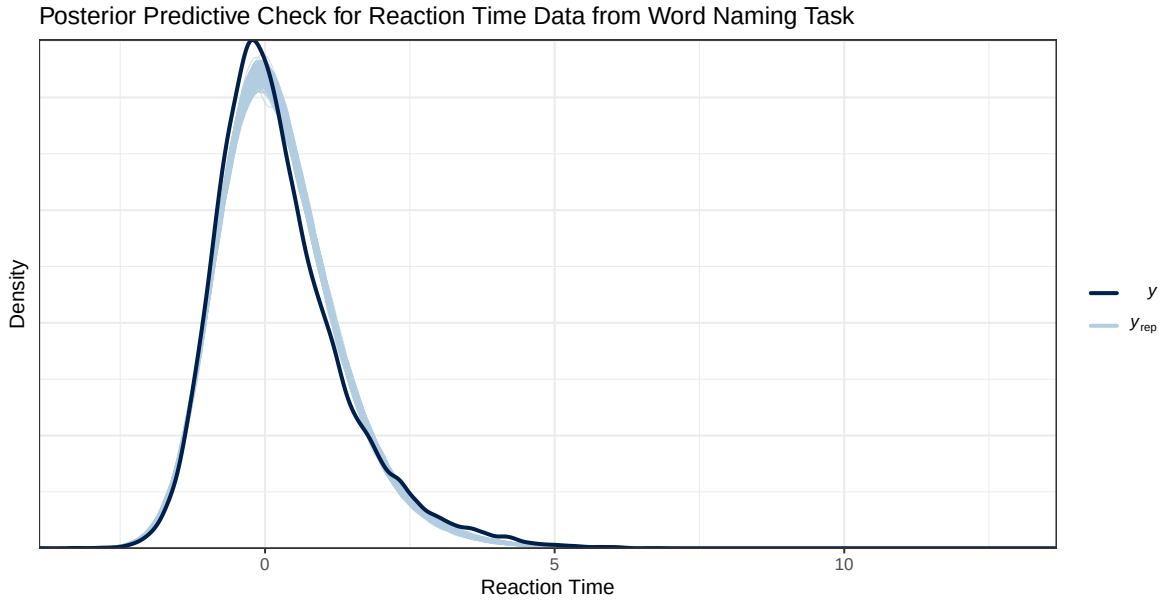
Posterior Predictive Check for Accuracy Data in the Word Naming Task



Reaction Time. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. The PPC shows that the posterior distribution of the model is estimated with a narrower spread of values than those observed. Figure 11 shows the grey lines of the markov-chain sampling process shifted slightly to the right of the leading edge of the curve for the observed data, and slightly to the left in the tail of the distribution.

Figure 11

Posterior Predictive Check for Reaction Time Data in the Word Naming Task

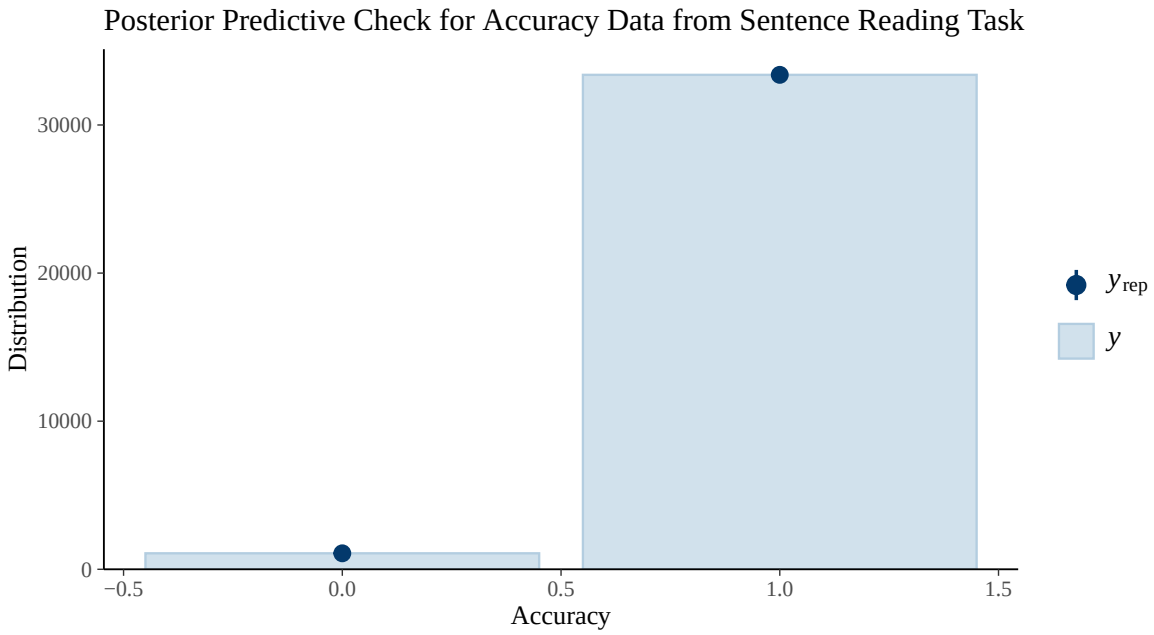


Sentence Reading

Accuracy. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. See Figure 12 for the PPC estimates. The model implied estimates for accurate and inaccurate responses are in line with the observed rates of accurate and inaccurate responses.

Figure 12

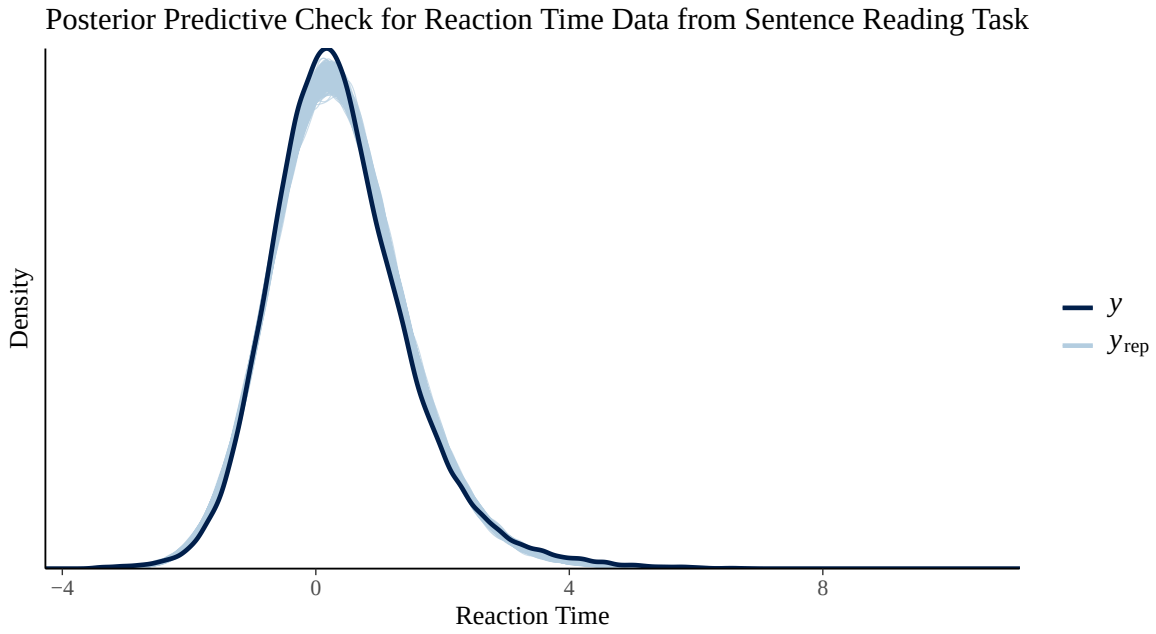
Posterior Predictive Check for Accuracy Data in the Sentence Reading Task



Reaction Time. Rhat and ESS diagnostic values were within limits and graphical inspection of the MCMC trace plots showed good mixing of MCMC chains during the sampling process. The PPC for the model slightly underestimates the observed data. Figure 13 shows the grey lines of the markov-chain sampling process shifted slightly to the right of the leading edge of the curve for the observed data, and slightly to the left in the tail of the distribution. Essentially, the posterior distribution of the model is estimated with a narrower spread of values than those observed.

Figure 13

Posterior Predictive Check for Reaction Time Data in the Sentence Reading Task



Bibliography

Adelman, J. S. and Brown, G. D. (2007). Phonographic Neighbors, Not Orthographic Neighbors, Determine Word Naming Latencies. *Psychonomic Bulletin and Review*, 14(3):455–459.

Adelman, J. S., Brown, G. D. A., and Quesada, J. F. (2006). Contextual Not Word Diversity , Frequency , Determines and Lexical Decision Times. *Psychological Science*, 17(9):814–823.

Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., and Estes, Z. (2014). Individual Differences in Reading Aloud: A Mega-Study, Item Effects, and Some Models. *Cognitive Psychology*, 68(0):113–160.

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26.

Allen, P. A., Bucur, B., Grabbe, J., Work, T., and Madden, D. J. (2011). Influence of Encoding Difficulty, Word Frequency, and Phonological Regularity on Age Differences in Word Naming. *Experimental Aging Research*, 37(3):261–292.

Allen, P. A., Lien, M. C., Murphy, M. D., Sanders, R. E., Judge, K. S., and McCann, R. S. (2002). Age Differences in Overlapping-Task Performance: Evidence for Efficient Parallel Processing in Older Adults. *Psychology and Aging*, 17(3):505–519.

Allen, P. A., Madden, D. J., and Crozier, L. C. (1991). Adult Age Differences in Letter-Level and Word-Level Processing. *Psychology and Aging*, 6(2):261–271.

Allen, P. A., Madden, D. J., and Slane, S. (1995). Visual word encoding and the effect of adult age and word frequency. In Allen, P. A. and Th. R. Bashore, editors, *Age Differences in Word and Language Processing*, chapter 2, pages 30–72. Elsevier Science B.V., London.

Allen, P. A., Madden, D. J., Weber, T. A., and Groth, K. E. (1993). Influence of Age and Processing Stage on Visual Word Recognition. *Psychology and Aging*, 8(2):274–282.

Allen, P. A., Murphy, M. D., Kaufman, M., Groth, K. E., and Begovic, A. (2004). Age Differences in Central (Semantic) and Peripheral Processing: The Importance of Considering Both Response Times and Errors. *The Journals of Gerontology*, 59B(5):210–9.

Andrews, S. (1989). Frequency and Neighborhood Effects on Lexical Access: Activation or Search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):802–814.

Andrews, S. (1992). Frequency and Neighborhood Effects on Lexical Access: Lexical Similarity or Orthographic Redundancy? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18(2):234–254.

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

Andrews, S. (2012). Individual Differences in Skilled Visual Word Recognition and Reading: The Role of Lexical Quality. In Adelman, J. S., editor, *Visual Word Recognition: Volume 2*, pages 151–172. Psychology Press, Hove.

Andrews, S. and Hersch, J. (2010). Lexical Precision in Skilled Readers: Individual Differences in Masked Neighbor Priming. *Journal of Experimental Psychology : General*, 139(2):299–318.

Andrews, S. and Lo, S. (2012). Not All Skilled Readers Have Cracked the Code: Individual Differences in Masked Form Priming. *Journal of Experimental Psychology: Learning Memory and Cognition*, 38(1):152–163.

Andrews, S. and Lo, S. (2013). Is Morphological Priming Stronger for Transparent than Opaque Words? It Depends on Individual Differences in Spelling and Vocabulary. *Journal of Memory and Language*, 68(3):279–296.

Andrews, S., Veldre, A., and Clarke, I. E. (2020). Measuring Lexical Quality: The Role of Spelling Ability. *Behavior Research Methods*, 52(6):2257–2282.

Araujo, S., Reis, A., Petersson, K. M., and Faisca, L. (2015). Rapid Automatized Naming and Reading Performance: A Meta-Analysis. *Journal Of Educational Psychology*, 107(3):868–883.

Åvall, M., Wolff, U., and Gustafsson, J.-E. (2019). Rapid automatized naming in a developmental perspective between ages 4 and 10. *Dyslexia (Chichester, England)*, (July):1–14.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical data base on CD-ROM. *Linguistic Data Consortium*, (January 1995).

Backman, J., Bruck, M., Hebert, M., and Seidenberg, M. S. (1984a). Acquisition and Use of Spelling-Sound Correspondences in Reading. *Journal of Experimental Child Psychology*, 38:114–133.

Backman, J. E., Mamen, M., and Ferguson, H. B. (1984b). Reading Level Design: Conceptual and Methodological Issues in Reading Research. *Psychological bulletin*, 96(3):560–568.

Baddeley, A. D., Logie, R. H., and Ellis, N. C. (1988). Characteristics of Developmental Dyslexia. *Cognition*, 29(3):197–228.

Baguley, T. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences*. Palgrave MacMillan.

Bakhtiari, D., Greenberg, D., Patten-Terry, N., and Nightingale, E. (2015). Spoken Oral Language and Adult Struggling Readers. *Journal of Research and Practice for Adult Literacy*, 4(1):9–20.

Balota, D. A. and Chumbley, J. I. (1984). Are Lexical Decisions a Good Measure of Lexical Access? The Role of Word Frequency in the Neglected Decision Stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10(3):340–357.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, 133(2):283–316.

Balota, D. A. and Ferraro, F. R. (1993). A Dissociation of Frequency and Regularity Effects in Pronunciation Performance across Young Adults, Older Adults, and Individuals with Senile Dementia of Alzheimer Type. *Journal of Memory and Language*, 32:573–592.

Balota, D. A. and Ferraro, F. R. (1996). Lexical, Sublexical, and Implicit Memory Processes in Healthy Young and Healthy Older Adults and in Individuals with Dementia of the Alzheimer Type. *Neuropsychology*, 10(1):82–95.

Balota, D. A., Yap, M. J., and Cortese, M. J. (2006). Visual Word Recognition: The Journey from Features to Meaning (A Travel Update). In Traxler, M. and Gernsbacher, M., editors, *Handbook of Psycholinguistics*, pages 285–375. Elsevier Inc., London, 2nd editio edition.

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., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Barber, H. A., Otten, L. J., Kousta, S.-T., and Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain Lang*, 125(1):47–53.

Barber, J. (2009). *The Development of Word Reading Automaticity in Connected Text of Adults with Reading Disabilities*. PhD thesis, University of Alberta.

Barca, L., Burani, C., Di Filippo, G., and Zoccolotti, P. (2006). Italian Developmental Dyslexic and Proficient Readers: Where Are the Differences? *Brain and Language*, 98(3):347–351.

Baron, J. and Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 2(3):386–393.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, 4:1–2.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beech, J. R. and Harding, L. M. (1984). Phonemic Processing and the Poor Reader from a Developmental Lag Viewpoint. *Reading Research Quarterly*, 19(3):357–366.

Begg, C. B. and Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias Author (s): Colin B . Begg and Madhuchhanda Mazumdar Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2533446. *Biometrics. Journal of the International Biometric Society*, 50(4):1088–1101.

Beidas, H., Khateb, A., and Breznitz, Z. (2013). The Cognitive Profile of Adult Dyslexics and Its Relation to Their Reading Abilities. *Reading and Writing*, 26(9):1487–1515.

Bell, L. C. and Perfetti, C. A. (1994). Reading Skill: Some Adult Comparisons. *Journal of Educational Psychology*, 86(2):244–255.

Ben-Dror, I., Pollatsek, A., and Scarpati, S. (1991). Word Identification in Isolation and in Context by College Dyslexic Students. *Brain and Language*, 40(4):471–490.

Bertelson, P., de Gelder, B., Tfouni, L. V., and Morais, J. (1989). Metaphonological abilities of adult illiterates: New evidence of heterogeneity. *European Journal of Cognitive Psychology*, 1(3):239–250.

Binder, K. S., a. Snyder, M., Ardoin, S. P., and Morris, R. K. (2011). Dynamic Indicators of Basic Early Literacy Skills: An Effective Tool to Assess Adult Literacy Students? *Adult Basic Education & Literacy Journal*, 5(3):150–160.

Bone, R. B., Cirino, P., Morris, R. D., and Morris, M. K. (2002). Reading and Phonological Awareness in Reading-Disabled Adults. *Developmental Neuropsychology*, 21(3):306–320.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, Chichester, UK.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., and Rothstein, H. R. (2017). Basics of Meta-Analysis: I 2 Is Not an Absolute Measure of Heterogeneity. *Research Synthesis Methods*, (September 2016).

Bosman, A. M. T., Vonk, W., and Van Zwam, M. (2006). Spelling Consistency Affects Reading in Young Dutch Readers with and without Dyslexia. *Annals of Dyslexia*, 56(2):271–300.

Braze, D., Tabor, W., Shankweiler, D. P., and Mencl, W. E. (2007). Speaking Up for Vocabulary: Reading Skill Differences in Young Adults. *Journal of Learning Disabilities*, 40(3):226–243.

Breznitz, Z. and Misra, M. (2003). Speed of processing of the visual-orthographic and auditory-phonological systems in adult dyslexics: The contribution of "asynchrony" to word recognition deficits. *Brain and Language*, 85(3):486–502.

Brown, G. D. and Deavers, R. P. (1999). Units of Analysis in Nonword Reading: Evidence from Children and Adults. *Journal of Experimental Child Psychology*, 73(3):208–242.

Bruck, M. (1988). The Word Recognition and Spelling of Dyslexic Children. *Reading Research Quarterly*, 23(1):51–69.

Bruck, M. (1990). Word-Recognition Skills of Adults with Childhood Diagnoses of Dyslexia. *Developmental Psychology*, 26(3):439–454.

Brysbaert, M. (2019). How many words do we read per minute ? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109(July):104047.

Brysbaert, M. and Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, 1(1):1–20.

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):441–458.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–11.

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological methods & research*, 33(2):261–304.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munaf'o, M. R. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Castles, A. and Coltheart, M. (1993). Varieties of Developmental Dyslexia. *Cognition*, 47:149–180.

Castles, A. and Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition*, 91(1):77–111.

Castles, A., Rastle, K., and Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, 19(1):5–51.

Cattell, J. M. (1886). The Time It Takes To See and Name Objects. *Mind; a quarterly review of psychology and philosophy*, os-XI(41):63–65.

Chang, Y.-N. and Monaghan, P. (2018). Quantity and Diversity of Preliteracy Language Exposure Both Affect Literacy Development: Evidence from a Computational Model of Reading. *Scientific Studies of Reading*, 00(00):1–19.

Chen, Y. (2008). A Statistical Associative Account of Vocabulary Growth in Early Word Learning. *Language Learning and Development*, 4(1):32–62.

Chetail, F. (2017). What Do We Do with What We Learn? Statistical Learning of Orthographic Regularities Impacts Written Word Processing. *Cognition*, 163:103–120.

Chumbley, J. I. and Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Mem Cognit*, 12(6):590–606.

Citron, F. M. M., Weekes, B. S., and Ferstl, E. C. (2014). How are affective word ratings related to lexicosemantic properties? Evidence from the Sussex Affective Word List. *Applied Psycholinguistics*, 35(02):313–331.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, New York, 2 edition.

Cohen-Shikora, E. R. and Balota, D. A. (2016). Visual Word Recognition across the Adult Lifespan. *Psychology and Aging*, 31(5):488–502.

Colenbrander, D., Miles, K. P., and Ricketts, J. (2019). To See or Not to See: How Does Seeing Spellings Support Vocabulary Learning? *Language, speech, and hearing services in schools*, 50(4):609–628.

Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychology*, 13(6):749–762.

Coltheart, M., Davelaar, E., Jonasson, J. T., and Besner, D. (1977). Access to the Internal Lexicon. In Dornic, S., editor, *Attention and Performance VI*, chapter 25, pages 535–556. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological review*, 108(1):204–56.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York, 2 edition.

Cortese, M. J. and Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 36(3):384–387.

Cumming, G. and Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and psychological measurement*, 61(4):532–574.

Davies, R., Arnell, R., Birchenough, J. M. H., Grimmond, D., and Houlson, S. (2017). Reading through the Lifespan: Individual Differences in Psycholinguistic Effects. *Journal of experimental psychology. Learning, memory, and cognition*.

Davies, R., Cuetos, F., and Glez-Seijas, R. M. (2007). Reading Development and Dyslexia in a Transparent Orthography: A Survey of Spanish Children. *Annals of Dyslexia*, 57(2):179–198.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37(1):65–70.

de Vaan, L., Schreuder, R., and Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon*, 2(1):1–24.

DfE (2023). The reading framework. Technical Report January.

Dilkina, K., McClelland, J. L., and Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2):136–164.

Dujardin, T., Etienne, Y., Contentin, C., Bernard, C., Largy, P., Mellier, D., Lalonde, R., and Rebaï, M. (2011). Behavioral performances in participants with phonological dyslexia and different patterns on the N170 component. *Brain and Cognition*, 75(2):91–100.

Dunabeitia, J. A. and Vidal-Abarca, E. (2008). Children like Dense Neighborhoods: Orthographic Neighborhood Density Effects in Novel Readers. *Spanish Journal of Psychology*, 11(1):26–35.

Ellis, A. W. and Lambon Ralph, M. A. (2000). Age of Acquisition Effects in Adult Lexical Processing Reflect Loss of Plasticity in Maturing Systems: Insights from Connectionist Networks. *Journal of experimental psychology. Learning, memory, and cognition*, 26(5):1103–23.

Ellis, N. C. (2002). FREQUENCY EFFECTS IN LANGUAGE PROCESSING A Review with Implications for. *SSLA*, 24:143–188.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK.

Eme, E., Lambert, E., and Alamargot, D. (2014). Word reading and word spelling in French adult literacy students: The relationship with oral language skills. *Journal of Research in Reading*, 37(3):268–296.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

Estes, Z. and Adelman, J. S. (2008). Automatic Vigilance for Negative Words Is Categorical and General. *Emotion (Washington, D.C.)*, 8(4):453–457.

Forster K., I. and Forster J., C. (2003). DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1):116–124.

Fox, J. and Weisburg, S. (2019). *An R Companion to Applied Regression*. Thousand Oaks, CA, 3 edition.

Fracasso, L. E., Bangs, K., and Binder, K. S. (2016). The Contributions of Phonological and Morphological Awareness to Literacy Skills in the Adult Basic Education Population. *Journal of learning disabilities*, 49(2):0022219414538513–.

Frost, R. (1998). Toward a Strong Phonological Theory of Visual Word Recognition: True Issues and False Trails. *Psychological Bulletin*, 123(1):71–99.

Frost, R. (2012). Towards a Universal Model of Reading. *Behavioral and Brain Sciences*, 35(5):263–279.

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect.

Gelman, A. and Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 9(6):641–51.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition, 3rd Edition*. 3rd edition. edition.

Gelman, A., Hill, J., and Vehtari, A. (2021). *Regression and Other Stories*. Cambridge University Press, Cambridge, UK.

Gelman, A. and Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. 60(4):328–331.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2):256–281.

Glushko, R. J. (1979). The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5(4):674–691.

Gottardo, A., Chiappe, P., Siegel, L. S., and Stanovich, K. E. (1999). Patterns of Word and Nonword Processing in Skilled and Less-Skilled Readers. *Reading and Writing*, 11(5-6):465–487.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11(4):323–343.

Graham, S. and Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and Writing*, 27(9):1703–1743.

Grainger, J. and Jacobs, A. M. (1996). Orthographic Processing in Visual Word Recognition: A Multiple Read-Out Model. *Psychological Review*, 103(3):518–565.

Graves, W. W., Binder, J. R., Desai, R. H., Humphries, C., Stengel, B. C., and Seidenberg, M. S. (2014). Anatomy Is Strategy: Skilled Reading Differences Associated with Structural Connectivity Differences in the Reading Network. *Brain and Language*, 133:1–13.

Greenberg, D., Ehri, L. C., and Perin, D. (1997). Are Word-Reading Processes the Same or Different in Adult Literacy Students and Third–Fifth Graders Matched for Reading Level? *JOURNAL OF EDUCATIONAL PSYCHOLOGY*, 89(2):262–275.

Hair Jr., J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2017). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage Publications, Ltd., London, 2 edition.

Hall, R., Greenberg, D., Laures-Gore, J., and Pae, H. K. (2014). The Relationship between Expressive Vocabulary Knowledge and Reading Skills for Adult Struggling Readers. *Journal of Research in Reading*, 37(S1):S87–S101.

Harm, M. W. and Seidenberg, M. S. (1999). Phonology, Reading Acquisition, and Dyslexia: Insights from Connectionist Models. *Psychological review*, 106(3):491–528.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor between Visual and Phonological Processes. *Psychological review*, 111(3):662–720.

Harrer, M., Cuijpers, P., and Ebert, D. (2019a). *Doing Meta-Analysis in R*.

Harrer, M., Cuijpers, P., Furukawa, T., and Ebert, D. D. (2019b). *Dmetar: Companion R Package for the Guide 'Doing Meta-Analysis in R'*.

Hart, L. and Perfetti, C. (2008). Learning Words in Zekkish: Implications for Understanding Lexical Representation. In Grigorenko, E. L. and Naples, A. J., editors, *Single-Word Reading: Behavioral and Biological Perspectives.*, New Directions in Communication Disorders Research: Integrative Approaches, pages 107–128. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Hartigan, J. A. and Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of statistics*, 13(1):70–84.

Higgins, J. P. T., Altman, D. G., and Sterne, J. A. C. (2011). Assessing risk of bias in included studies. In Higgins, J. P. T. and Green, S., editors, *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 8. John Wiley & Sons, Ltd, London, version 5. edition.

Hino, Y. and Lupker, S. J. (1996). Effects of Polysemy in Lexical Decision and Naming: An Alternative to Lexical Access Accounts. *Journal of Experimental Psychology: Human Perception and Performance*, 22(6):1331–1356.

Hock, M. F. (2012). Effective literacy instruction for adults with specific learning disabilities: Implications for adult educators. *Journal of learning disabilities*, 45(1):64–78.

Hoffman, P., McClelland, J. L., and Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, 125(3):293–328.

Hoffman, P. and Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of experimental psychology. Human perception and performance*, 41(2):385–402.

Hofmann, M. J., Stenneken, P., Conrad, M., and Jacobs, A. M. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behavior Research Methods*, 39(3):620–629.

Horn, C. C. and Manis, F. R. (1985). Normal and Disabled Readers' Use of Orthographic Structure in Processing Print. *Journal of Reading Behavior*, 17(2):143–161.

Hsiao, Y. and Nation, K. (2018). Semantic Diversity, Frequency and the Development of Lexical Quality in Children's Word Reading. *Journal of Memory and Language*, 103(August):114–126.

Hulslander, J., Olson, R. K., Willcutt, E. G., and Wadsworth, S. J. (2010). Longitudinal Stability of Reading-Related Skills and Their Prediction of Reading Development. *Scientific Studies of Reading*, 14(2):111–136.

Jared, D. (1997). Spelling – Sound Consistency Affects the Naming of High-Frequency Words. *Journal of Memory and Language*, 36:505–529.

Jared, D., McRae, K., and Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, 29(6):687–715.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive psychology*, 13(2):278–305.

Jastrzembski, J. E. and Stanners, R. F. (1975). Multiple word meanings and lexical search speed. *Journal of verbal learning and verbal behavior*, 14(5):534–537.

Jimenez, J. E., Garcia, E., and Venegas, E. (2008). Are phonological processes the same or different in low literacy adults and children with or without reading disabilities? *Reading and Writing*, 23(1):1–18.

Jimenez Gonzalez, J. E. and Valle, I. H. (2000). Word Identification and Reading Disorders in the Spanish Language. *Journal of Learning Disabilities*, 33(1):44–60.

Jorm, A. F. (1981). Children With Reading and Spelling Retardation: Functioning of Whole-Word and Correspondence-Rule Mechanisms. *Journal of Child Psychology and Psychiatry*, 22(2):171–178.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131(5):684–712.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., and Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology*, 64(9):1683–1691.

Kassambara, A. (2023). *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests*.

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., and Whalen, D. H. (2012). What lexical decision and naming tell us about reading. *Read Writ*, 25(6):1259–1282.

Keating, G. C. (1987). *The Effects of Word Characteristics on Children's Reading*. PhD thesis, City of London Polytechnic.

Kessler, B., Treiman, R., and Mullennix, J. (2002). Phonetic Biases in Voice Key Response Time Measurements. *Journal of Memory and Language*, 47(1):145–171.

Keuleers, E. and Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Q J Exp Psychol (Hove)*, 68(8):1457–1468.

Keuleers, E. and Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Kirby, J. R., Georgiou, G. K., Martinussen, R., Parrila R. Bowers, P., and Landerl, K. (2010). Naming Speed and Reading: From Prediction to Instruction. *Reading Research Quarterly*, 45(3):241–362.

Klose, A. E., Schwartz, S., and Brown, J. W. (1983). The Imageability Effect in Good and Poor Readers. *Bulletin of the Psychonomic Society*, 21(6):446–448.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., and Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

Kucera, H. and Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Brown University Press, Providence, RI.

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. *Frontiers in Psychology.*

Kuperman, V., Estes, Z., Brysbaert, M., and Warriner, A. B. (2014). Emotion and Language: Valence and Arousal Affect Word Recognition. *J Exp Psychol Gen*, 143(3):1065–1081.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Kuperman, V. and Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of experimental psychology. Human perception and performance*, 39(3):802–23.

Kwok, R. K. W. and Ellis, A. W. (2014). Visual Word Learning in Adults with Dyslexia. *Frontiers in Human Neuroscience*, 8(May):264–264.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44:701–710.

Lau, J., Ionnidis, J. P. A., Terrin, N., Schmid, C. H., and Olkin, I. (2006). The Case of the Misleading Funnel Plot. *BMJ : British Medical Journal*, 333(7568):597–600.

Laver, G. D. and Burke, D. M. (1993). Why Do Semantic Priming Effects Increase in Old Age? A Meta-Analysis. *Psychology and Aging*, 8(1):34–43.

Laxon, V., Masterson, J., and Coltheart, V. (1991). Some Bodies Are Easier to Read: The Effect of Consistency and Regularity on Children's Reading. 43(4):793–824.

Laxon, V. J., Coltheart, V., and Keating, C. (1988). Children Find Friendly Words Friendly Too: Words with Many Orthographic Neighbours Are Easier to Read and Spell. *British Journal of Educational Psychology*, 58(1):103–119.

Leisch, F. (2006). A toolbox for K -centroids cluster analysis. *Computational statistics & data analysis*, 51(2):526–544.

Leitch, S. (2006). Prosperity for All in the Global Economy-World Class Skills Final Report. Technical report, HM Treasury, London.

Lichacz, F. M., Herdman, C. M., Lefevre, J. O., and Baird, B. (1999). Polysemy Effects in Word Naming. *Canadian Journal of Experimental Psychology*, 53(2):189–193.

Lima, S. D., Hale, S., and Myerson, J. (1991). How general is general slowing? Evidence from the lexical domain. *Psychology and Aging*, 6(3):416–425.

Long, J. D. (2011). *Longitudinal Data Analysis for the Behavioral Sciences Using R*. SAGE, London.

Lovett, M. W. (1987). A Developmental Approach to Reading Disability : Accuracy and Speed Criteria of Normal and Deficient Reading Skill. *Child Development*, 58(1 PG - 234-260):234–260.

MacArthur, C. A., Greenberg, D., Mellard, D. F., and Sabatini, J. P. (2010). Introduction to the Special Issue on Models of Reading Component Skills in Low-Literate Adults. *Journal of Learning Disabilities*, 43(2):99–100.

Mahé, G., Pont, C., Zesiger, P., and Laganaro, M. (2018). The electrophysiological correlates of developmental dyslexia: New insights from lexical decision and reading aloud in adults. *Neuropsychologia*, 121(October):19–27.

Marcolini, S., Traficante, D., Zoccolotti, P., and Burani, C. (2011). Word Frequency Modulates Morpheme-Based Reading in Poor and Skilled Italian Readers. *Applied Psycholinguistics*, 32(03):513–532.

Marinelli, C. V., Traficante, D., and Zoccolotti, P. (2014). Does Pronounceability Modulate the Letter String Deficit of Children with Dyslexia? A Study with the Rate and Amount Model. *Frontiers in Psychology*, 5(DEC):1–16.

Marinelli, C. V., Traficante, D., Zoccolotti, P., and Burani, C. (2013). Orthographic Neighborhood-Size Effects on the Reading Aloud of Italian Children With and Without Dyslexia. *Scientific Studies of Reading*, 17(5):333–349.

Marinus, E. and de Jong, P. F. (2010). Size Does Not Matter, Frequency Does: Sensitivity to Orthographic Neighbors in Normal and Dyslexic Readers. *Journal of Experimental Child Psychology*, 106(2-3):129–144.

Martin-Chang, S., Ouellette, G., and Madden, M. (2014). Does Poor Spelling Equate to Slow Reading? The Relationship between Reading, Spelling, and Orthographic Quality. *Reading and Writing*, 27(8):1485–1505.

Mason, M. (1978). From Print to Sound in Mature Readers as a Function of Reader Ability and Two Forms of Orthographic Regularity. *Memory & Cognition*, 6(5 PG - 568-581):568–581.

Masterson, J., Laxon, V., Lovejoy, S., and Morris, V. (2007). Phonological Skill, Lexical Decision and Letter Report Performance in Good and Poor Adult Spellers. *Journal of Research in Reading*, 30(4):429–442.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.

McElreath, R. (2020). *Statistical Rethinking : A Bayesian Course with Examples in R and Stan*. Boca Raton, second edition. edition.

McKoon, G. and Ratcliff, R. (2016). Adults with Poor Reading Skills: How Lexical Knowledge Interacts with Scores on Standardized Reading Comprehension Tests. *Cognition*, 146:453–469.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, 34(2):103–115.

Melby-Lervag, M., Lyster, S. A. H., and Hulme, C. (2012). Phonological Skills and Their Role in Learning to Read: A Meta-Analytic Review. *Psychological Bulletin*, 138(2):322–352.

Mellard, D., Anthony, J., and Woods, K. (2012a). Understanding oral reading fluency among adults with low literacy: Dominance analysis of contributing component skills. *Reading and Writing*, 25(6):1345–1364.

Mellard, D., Woods, K., and Md Desa, Z. D. (2012b). An Oral Reading Fluency Assessment for Young Adult Career and Technical Education Students. *Learning Disabilities Research & Practice*, 27(3):125–135.

Mellard, D. F., Fall, E., and Woods, K. L. (2010). A Path Analysis of Reading Comprehension for Adults With Low Literacy. *Journal of Learning Disabilities*, 43(2):154–165.

Mellard, D. F. and Patterson, M. B. (2008). Contrasting Adult Literacy Learners With and Without Specific Learning Disabilities. *Remedial and Special Education*, 29(3):133–144.

Mellard, D. F., Woods, K. L., and Lee, J. H. (2016). Literacy profiles of at-risk young adults enrolled in career and technical education. *Journal of Research in Reading*, 39(1):88–108.

Metsala, J. L., Stanovich, K. E., and Brown, G. D. A. (1998). Regularity Effects and the Phonological Deficit Model of Reading Disabilities: A Meta-Analytic Review. *Journal Of Educational Psychology*, 90(2):279–293.

Meyer, M. S., Wood, F. B., Hart, L. A., and Felton, R. H. (1998a). Longitudinal Course of Rapid Naming in Disabled and Nondisabled Readers. In *Annals of Dyslexia*, volume 48, pages 91–114.

Meyer, M. S., Wood, F. B., Hart, L. A., and Felton, R. H. (1998b). Selective Predictive Value of Rapid Automatized Naming in Poor Readers. *Journal of Learning Disabilities*, 31(2):106–117.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Mimeau, C., Ricketts, J., and Deacon, S. H. (2018). The Role of Orthographic and Semantic Learning in Word Reading and Reading Comprehension. *Scientific Studies of Reading*, 22(5):384–400.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta Analyses: The Prisma Statement. *Bristish Medical Journal*, 339(7716):332–336.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., and Stewart, L. A. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. *Systematic Reviews*, 4(1):1–9.

Mol, S. E. and Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, 137(2):267–296.

Monaghan, P. and Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language*, 63(4):506–525.

Morais, J., Bertelson, P., Cary, L., and Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, 24(1):45–64.

Morais, J., Cary, L., Alegria, J., and Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, 7(4):323–331.

Morrison, C. M., Hirsh, K. W., Chappell, T., and Ellis, A. W. (2002). Age and Age of Acquisition: An Evaluation of the Cumulative Frequency Hypothesis. *European Journal of Cognitive Psychology*, 14(4):435–459.

Morrison, C. M., Hirsh, K. W., and Duggan, G. B. (2003). Age of Acquisition, Ageing, and Verb Production: Normative and Experimental Data. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 56 A(4):705–730.

Muncer, S. J., Knight, D., and Adams, J. W. (2014). Bigram Frequency, Number of Syllables and Morphemes and Their Effects on Lexical Decision and Word Naming. *Journal of Psycholinguistic Research*, 43(3):241–254.

Nakagawa, S. and Santos, E. S. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, 26(5):1253–1274.

Nanda, A. O., Greenberg, D., and Morris, R. (2010). Modeling Child-Based Theoretical Constructs with Struggling Adult Readers. *Journal of Learning Disabilities*, 43(2):139–153.

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *npj Science of Learning*, 2(1):0–1.

Nation, K. and Castles, A. (2017). Putting the Learning into Orthographic Learning. In Cain, K., Compton, D. L., and Parrila, R. K., editors, *Theories of Reading Development*, pages 147–168. John Benjamins Publishing Co., Amsterdam.

Nelson, J. R., Balass, M., and a. Perfetti, C. (2005). Differences between Written and Spoken Input in Learning New Words. *Written Language & Literacy*, 8(2):25–44.

Nilssen-Nergård, T. and Hulme, C. (2014). Developmental dyslexia in adults: Behavioural manifestations and cognitive correlates. *Dyslexia (Chichester, England)*, 20(3):191–207.

Olson, R. K., Kliegl, R., Davidson, B. J., and Foltz, G. (1985). Individual and Developmental Differences in Reading Disability. *Reading research: Advances in theory and practice, Vol. 4.*, (December 2013):1–64.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science (New York, N.Y.)*, 349(6251):aac4716–aac4716.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, 372.

Paivio, A. (1991). Dual Coding Theory: Retrospect and Current Status. *Canadian journal of experimental psychology*, 45(3):255.

Palmer, T. M., Peters, J. L., Sutton, A. J., and Moreno, S. G. (2008). Contour-enhanced funnel plots for meta-analysis. *Stata Journal*, 8(2):242–254.

Parrila, R., Georgiou, G., and Corkett, J. (2007). University Students with a Significant History of Reading Difficulties: What Is and Is Not Compensated? *Exceptionality Education Canada*, 17(2):195–220.

Patterson, K. E. and Marcel, A. J. (1977). Aphasia, Dyslexia and the Phonological Coding of Written Words. *Quarterly Journal of Experimental Psychology*, 29(2):307–318.

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, 11(4):357–383.

Perfetti, C. (2011). Phonology is critical in reading: But a phonological deficit is not the only source of low reading skill. In Brady, S. A., Braze, D., and Fowler, C. A., editors, *Explaining Individual Differences in Reading: Theory and Eveidence*, chapter 8, pages 153–171. Psychology Press, New York.

Perfetti, C. and Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading*, 18(1):22–37.

Perfetti, C. A. and Hart, L. (2002). The Lexical Quality Hypothesis. In Verhoeven, L., Elbro, C., and Reitsma, P., editors, *Precursors of Functional Literacy*, pages 189–213. John Benjamins Publishing Co., Amsterdam.

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested Incremental Modeling in the Development of Computational Theories: The CDP+ Model of Reading Aloud. *Psychological review*, 114(2):273–315.

Pigott, T. D. (2012). *Advances in Meta-Analysis*. Springer Science+Business Media, London.

Plaut, D. C. (1997). Structure and Function in the Lexical System: Insights from Distributed Models of Word Reading and Lexical Decision. *Language and Cognitive Processes*, 12(5-6):765–806.

Plaut, D. C., Mcclelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding Normal and Impaired Word Reading : Computational Principles in Quasi-Regular Domains. *Psychological Review*, 103(1):56–115.

Pritchard, S. C., Coltheart, M., Marinus, E., and Castles, A. (2018). A Computational Model of the Self-Teaching Hypothesis Based on the Dual-Route Cascaded Model of Reading. *Cognitive Science*, 42(3):722–770.

Protopapas, A. (2007). Check Vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX - Lancaster University (Alma). *Behaviour Research Methods*, 39(4):859–862.

Protopapas, A., Mitsi, A., Koustoumbardis, M., Tsitsopoulou, S. M., Leventi, M., and Seitz, A. R. (2017). Incidental orthographic learning during a color detection task. *Cognition*, 166:251–271.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rack, J. P., Snowling, M. J., and Olson, R. K. (1992). The Nonword Reading Deficit in Developmental Dyslexia : A Review. *Reading research quarterly*, 27(1):28–53.

Ramscar, M., Hendrix, P., Love, B., and Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The mental lexicon*, 8(3):450–481.

Ratcliff, R., Thapar, A., Gomez, P., and McKoon, G. (2004). A Diffusion Model Analysis of the Effects of Aging in the Lexical-Decision Task. *Psychology and aging*, 19(2):278–89.

Ratcliff, R., Thapar, A., and McKoon, G. (2010). Individual Differences, Aging, and IQ in Two-Choice Tasks. *Cognitive Psychology*, 60(3):127–157.

Re, A. C. D. (2013). *Compute.Es: Compute Effect Sizes*.

Reis, A., Araújo, S., Morais, I. S., and Faísca, L. (2020). Reading and reading-related skills in adults with dyslexia from different orthographic systems: A review and meta-analysis. *Annals of Dyslexia*.

Ricketts, J., Bishop, D. V., and Nation, K. (2009). Orthographic facilitation in oral vocabulary acquisition. *Quarterly Journal of Experimental Psychology*, 62(10):1948–1966.

Ricketts, J., Bishop, D. V., Pimperton, H., and Nation, K. (2011). The role of self-teaching in learning orthographic and semantic aspects of new words. *Scientific Studies of Reading*, 15(1):47–70.

Ricketts, J., Davies, R., Masterson, J., Stuart, M., and Duff, F. J. (2016). Evidence for Semantic Involvement in Regular and Exception Word Reading in Emergent Readers of English. *Journal of Experimental Child Psychology*, 150:330–345.

Rohr, M. and Wentura, D. (2021). Degree and Complexity of Non-conscious Emotional Information Processing – A Review of Masked Priming Studies. *Frontiers in Human Neuroscience*, 15.

Romani, C., Di Betta, A., Tsouknida, E., and Olson, A. (2008). Lexical and Nonlexical Processing in Developmental Dyslexia: A Case for Different Resources and Different Impairments. *Cognitive Neuropsychology*, 25(6):798–830.

Rosenthal, J. A. (1996). Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research*, 21(4):37–59.

Sabatini, J. P. (2002). Efficiency in Word Reading of Adults : Ability Group Comparisons. *Scientific Studies of Reading*, 6(3):267–298.

Sabatini, J. P., Sawaki, Y., Shore, J. R., and Scarborough, H. S. (2010). Relationships among Reading Skills of Adults with Low Literacy. *Journal of learning disabilities*, 43(2):122–138.

Samara, A. and Caravolas, M. (2014). Statistical learning of novel graphotactic constraints in children and adults. *Journal of Experimental Child Psychology*, 121(1):137–155.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, 48:115–136.

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110:104038.

Schröter, P. and Schroeder, S. (2017). The Developmental Lexicon Project: A Behavioral Database to Investigate Visual Word Recognition across the Lifespan. *Behavior Research Methods*, 49(6):2183–2203.

Schunemann, H. J., Oxman, A. D., Vist, G. E., Higgins, J. P. T., Deeks, J. J., Glasziou, P., and Guyatt, G. H. (2011). Chapter 12: Interpreting Results and Drawing Conclusions. In Higgins, J. P. T. and Green, S., editors, *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, 2011, version 5. edition.

Schwanenflugel, P. J., Harnishfeger, K. K., and Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of memory and language*, 27(5):499–520.

Sedlmeier, P. and Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, 105(2):309–316.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition*, 19(1):1–30.

Seidenberg, M. S., Bruck, M., Fornarolo, G., and Backman, J. (1985). Word recognition processes of poor and disabled readers: Do they necessarily differ? *Applied Psycholinguistics*, 6(2):161–180.

Seidenberg, M. S. and McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological review*, 96(4):523–68.

Seidenberg, M. S. and Plaut, D. C. (1998). Evaluating Word-Reading Models at the Item Level: Matching the Grain of Theory and Data. *Psychological Science*, 9(3):234–237.

Seidenberg, M. S. and Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In Andrews, S., editor, *From Inkmarks to Ideas: Current Issues in Lexical Processing*, pages 25–49. Psychology Press, Hove, UK.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., and Tanenhaus, M. K. (1984). When Does Irregular Spelling or Pronunciation Influence Word Recognition? *Journal of Verbal Learning and Verbal Behavior*, 23(3):383–404.

Shadish, W. R. and Haddock, C. K. (2009). Combining Estimates of Effect Sizes. In Cooper, H., Hedges, L. V., and Valentine, J. C., editors, *The Handbook of Research Synthesis and Meta-Analysis*, pages 257–277. Russell Sage Foundation, New York, 2 edition.

Shanahan, T. and Lonigan, C. J. (2010). The National Early Literacy Panel: A summary of the process and the report. *Educational Researcher*, 39(4):279–285.

Share, D. (1999). Phonological Recoding and Orthographic Learning: A Direct Test of the Self-Teaching Hypothesis. *Journal of Experimental Child Psychology*, 72:95–129.

Share, D. L. (1995). Phonological Recoding and Self-Teaching - Sine-Qua-Non of Reading Acquisition. *Cognition*, 55(2):151–218.

Share, D. L. (2004). Orthographic Learning at a Glance: On the Time Course and Developmental Onset of Self-Teaching. *Journal of Experimental Child Psychology*, 87(4):267–298.

Share, D. L. (2021). Is the Science of Reading Just the Science of Reading English? *Reading Research Quarterly*, 56(1):S391–S402.

Shipley, W. C. (1940). A Self-Administering Scale for Measuring Intellectual Impairment and Deterioration. *The Journal of Psychology*, 9(2):371–377.

Siegel, L. S. and Ryan, E. B. (1988). Development of Grammatical-Sensitivity, Phonological, and Short-Term Memory Skills in Normally Achieving and Learning Disabled Children. *Developmental Psychology*, 24(1):28–37.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General*, 143(2):534–547.

Singer, J. D., Willett, J. B., and Online, O. S. (2003). *Applied Longitudinal Data Analysis Modeling Change and Event Occurrence*. Oxford University Press, New York.

Snefjella, B. and Kuperman, V. (2016). It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition*, 156:135–146.

Snowling, M. J. and Melby-Lervag, M. (2016). Oral Language Deficits in Familial Dyslexia: A Meta-Analysis and Review. *Psychological Bulletin*, 142(5):498–545.

Spieler, D. H. and Balota, D. A. (1997). Bringing Computational Models of Word Naming down to the Item Level. *Psychological Science*, 8(6):411–416.

Spieler, D. H. and Balota, D. A. (2000). Factors Influencing Word Naming in Younger and Older Adults. *Psychology and Aging*, 15(2):225–231.

Stanovich, K. E. (1986). Matthew Effects in Reading : Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly*, 21(4):360–407.

Steacy, L. M., Elleman, A. M., and Compton, D. L. (2017a). Opening the "Black Box" of Learning to Read: Inductive Learning Mechanisms Supporting Word Acquisition Development with a Focus on Children Who Struggle to Read. In Cain, K., Compton, D. L., and Parrila, R. K., editors, *Theories of Reading Development*, pages 99–124. John Benjamins Publishing Co., Amsterdam.

Steacy, L. M., Kearns, D. M., Gilbert, J. K., Compton, D. L., Cho, E., Lindstrom, E. R., and Collins, A. A. (2017b). Exploring Individual Differences in Irregular Word Recognition Among Children with Early-Emerging and Late-Emerging Word Reading Difficulty. *Journal of Educational Psychology*, 109(1):51–69.

Steyvers, M. and Tenenbaumb, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78.

Strain, E. and Herdman, C. M. (1999). Imageability Effects in Word Naming: An Individual Differences Analysis. *Canadian Journal of Experimental Psychology*, 53(4):347–359.

Suarez-Coalla, P. and Cuetos, F. (2015). Reading Difficulties in Spanish Adults with Dyslexia. *Annals of Dyslexia*, 65(1):33–51.

Swanson, H. L. (1994). Short-Term Memory and Working Memory: Do Both Contribute to Our Understanding of Academic Achievement in Children and Adults with Learning Disabilities? *J Learn Disabil*, 27(1):34–50.

Swanson, H. L. and Hsieh, C. J. (2009). Reading disabilities in adults: A selective meta-analysis of the literature. *Review of Educational Research*, 79(4):1362–1390.

Swanson, H. L., Trainin, G., and Necoechea, D. M. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. 73(4):407–440.

Tainturier, M. J., Tremblay, M., and Lecours, A. R. (1989). Aging and the Word Frequency Effect: A Lexical Decision Investigation. *Neuropsychologia*, 27(9):1197–1203.

Talwar, A., Greenberg, D., and Li, H. (2018). Does memory contribute to reading comprehension in adults who struggle with reading? *Journal of research in reading*, 41(S1):S163–S182.

Tamura, N., Castles, A., and Nation, K. (2017). Orthographic Learning, Fast and Slow: Lexical Competition Effects Reveal the Time Course of Word Learning in Developing Readers. *Cognition*, 163:93–102.

Taylor, J. S., Davis, M. H., and Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. *Journal of Experimental Psychology: General*, 146(6):826–858.

Thompkins, A. C. and Binder, K. S. (2003). A comparison of the factors affecting reading performance of functionally illiterate adults and children matched by reading level. *Reading Research Quarterly*, 38(2):236–258.

Thorndike, E. L. (1944). *The Teacher's Word Book of 30,000 Words.* Columbia university.

Tighe, E. L. and Schatschneider, C. (2016). Examining the Relationships of Component Reading Skills to Reading Comprehension in Struggling Adult Readers: A Meta-Analysis. *Journal of learning disabilities*, 49(4):395–409.

Torgesen, J. K., Wagner, R. K., and Rashotte, C. A. (2012). Test of Word Reading Efficiency 2. Technical report, Pro-ed, Austin: Texas.

Treiman, R. (2018). Statistical Learning and Spelling. *Language, Speech, and Hearing Services in Schools*, 49(3S):644–652.

Treiman, R., Goswami, U., and Bruck, M. (1990). Not All Nonwords Are Alike: Implications for Reading Development and Theory. *Memory & Cognition*, 18(6):559–567.

Treiman, R. and Hirsh-Pasek, K. (1985). Are There Qualitative Differences in Reading Behavior between Dyslexics and Normal Readers? *Memory & Cognition*, 13(4):357–364.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., and Richmond-Welty, E. D. (1995). The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography. *Journal of Experimental Psychology: General*, 124(2):107–136.

Van Aert, R. C. M., Wicherts, J. M., and Van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p-values: Reservations and recommendations for applying p-uniform and p-curve. (1925).

van Buuren, S. (2018). *Flexible Imputation of Missing Data.* CRC Press, Boca Raton, Florida, 2 edition.

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1):111–122.

Van Heuven, W., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A New and Improved Word Frequency Database for British English. *Quarterly journal of experimental psychology.*, 67(6):1176–1190.

van Ijzendoorn, M. H. and Bus, A. G. (1994). Meta-Analytic Confirmation of the Nonword Reading Deficit in Developmental Dyslexia. *Reading Research Quarterly*, 29(3):266–275.

Vasishth, S. and Nicenboim, B. (2016). Statistical Methods for Linguistic Research: Foundational Ideas - Part I. *Language and Linguistics Compass*, 10(8):349–369.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. 27(5):1413–1432.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., and Chen, R. (2007). Components of Reading Ability: Multivariate Evidence for a Convergent Skills Model of Reading Development. *Scientific Studies of Reading*, 11(1):3–32.

Venezky, R. L. (1970). The Structure of English Orthography. In *The Structure of the English Orthography*, chapter 7, pages 1–162. Mouton, reprint 20 edition.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the Metafor Package. *Journal of Statistical Software*, 36(3):1–48.

Wagner, R., Torgesen, J., Rashotte, C., and Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing*. Pro-Ed, Austin: Texas, 2 edition.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45 VN - r(4):1191–1207.

Waters, G. S. and Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition*, 13(6):557–572.

Waters, G. S., Seidenberg, M. S., and Bruck, M. (1984). Children's and adults' use of spelling-sound information in three reading tasks. *Memory & cognition*, 12(3):293–305.

Wechsler, D. (2001). *Wechsler Individual Achievement Test*. Pearson, New York, NY, 2 edition.

Westfall, J. (2016). Designing multi-lab replication projects: Number of labs matters more than number of participants.

Westfall, Jacob, Kenny, D. A., and Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5):2020–2045.

Wild, H., Kyrolainen, A.-J., and Kuperman, V. (2022). How representative are student convenience samples ? A study of literacy and numeracy skills in 32 countries. *PLoS ONE*, 17(7):1–22.

Woollams, A. M. (2005). Imageability and Ambiguity Effects in Speeded Naming: Convergence and Divergence. *J Exp Psychol Learn Mem Cogn*, 31(5):878–890.

Woollams, A. M., Lambon Ralph, M. A., Madrid, G., and Patterson, K. E. (2016). Do You Read How I Read? Systematic Individual Differences in Semantic Reliance amongst Normal Readers. *Frontiers in Psychology*, 7(NOV):1–16.

Yang, C. L., Perfetti, C. A., and Schmalhofer, F. (2005). Less Skilled Comprehenders' ERPs Show Sluggish Word-to-Text Integration Processes. *Written Language and Literacy*, 8(2):233–257.

Yap, M. J., Balota, D. A. W. U., Sibley, D. E., and Ratcliff, R. (2012). Individual Differences in Visual Word Recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology-Human Perception and Performance*, 38(1):53–79.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., and Rueckl, J. (2015). Responding to Nonwords in the Lexical Decision Task: Insights from the English Lexicon Project. *Journal of experimental psychology. Learning, memory, and cognition*, 41(3):597–613.

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

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45(e1):1–78.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, 15(5):971–979.

Yurovsky, D., Fricker, D. C., Yu, C., and Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, 21(1):1–22.

Zevin, J. D. and Balota, D. A. (2000). Priming and Attentional Control of Lexical and Sublexical Pathways in Naming. *Journal of Experimental Psychology: Learning Memory and Cognition*, 26(1):121–135.

Zevin, J. D. and Seidenberg, M. S. (2006). Simulating Consistency Effects and Individual Differences in Nonword Naming: A Comparison of Current Models. *Journal of Memory and Language*, 54:145–160.

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F.-X., and Perry, C. (2008). Developmental Dyslexia and the Dual Route Model of Reading: Simulating Individual Differences and Subtypes. *Cognition*, 107:151–178.

Ziegler, J. C. and Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, 131(1):3–29.

Ziegler, J. C., Stone, G. O., and Jacobs, A. M. (1997). What's the pronunciation for –ough and the spelling for /u/? A database for computing feedforward and feedback inconsistency in English. *Behavior Research Methods, Instruments, & Computers*, 29(4):600–618.

Zoccolotti, P., de Luca, M., Di Filippo, G., Judica, A., and Martelli, M. (2009). Reading Development in an Orthographically Regular Language: Effects of Length, Frequency, Lexicality and Global Processing Ability. *Reading and Writing*, 22(9):1053–1079.