Individual Differences in Skilled Reading and the Word Frequency Effect

Charlotte E. Lee[1], Hayward Godwin[1], Hazel I. Blythe[2], and Denis Drieghe[1]

[1] School of Psychology, University of Southampton, Southampton, United Kingdom,

[2] School of Psychology, Northumbria University, Newcastle, United Kingdom.

Corresponding author:

Charlotte E. Lee

School of Psychology

University of Southampton

Southampton SO17 1BJ

U.K.

Email: c.lee@soton.ac.uk

https://orcid.org/0000-0003-0319-5635

This experiment was not pre-registered.

Authors note:

The research for this article was funded by the UKRI Economic and Social Research Council South Coast Doctoral Training Partnership (Grant Number ES/P000673/1).

Open Practice Statement: All data and materials are available online at:

https://osf.io/fetw4/?view_only=166f33ba5c584b3aa10efd30972840eb

Word Count: 13005

Abstract

Variation in eye movement patterns can be considerable even within skilled readers. Here, individual differences and eye movements of 88 average-to-very-skilled readers were assessed to examine the reliability of previous observations of a reduced word frequency effect associated with skilled reading. Shorter fixation durations and higher skipping rates were observed for high frequency compared to low frequency words. Replicating the literature, high scores on reading ability and vocabulary knowledge tests predicted reduced frequency effects in gaze duration in models with single individual differences predictors. This reduced frequency effect expressed itself in faster reading of low frequency words in high scorers compared to low scorers. Next, a PCA grouped individual differences tests based on shared variance. High 'lexical proficiency' predicted shorter gaze durations, sentence reading times, and increased word skipping. 'Lexical proficiency' and the WIAT-II comprehension test predicted a reduced frequency effect in go past times and all tests apart from the NDRT comprehension test predicted a reduced frequency effect in sentence reading times. Data revealed surprising discrepancies in findings based on two subtests supposedly measuring comprehension (NDRT and WIAT-II), constituting an example of the *jingle fallacy*: the false assumption that two measures that share a name actually measure the same construct.

Keywords: individual differences, eye movements, word frequency, lexical proficiency, jingle fallacy

Statement of Public Significance:

This study highlights that lexical proficiency is important when predicting eye movement patterns in skilled reader populations, especially when it comes to the processing of low-frequency words. In addition, it demonstrates what Thorndike (1904) called a jingle fallacy whereby two often used offline tests for comprehension (NDRT and WIAT-II) share the same name but do not appear to measure the same underlying construct and as a result, should be carefully considered when selected as measures in future research.

Individual Differences in Skilled Reading and the Word Frequency Effect

In this study, we examine whether differences in the size of the word frequency effect on eye movements during reading are related to differences in a number of reading related skills as assessed by a large test battery. The current paper focusses on the word frequency effect during sentence reading in a controlled experiment, given that a comparatively small frequency effect during reading has been suggested as indicative of better reading ability (Ashby et al., 2005; Haenggi & Perfetti, 1994). The goal of this paper was to determine which skill (or skills) best predict individual differences in the word frequency effect on eye movement measures when controlling for the influence of other skills. We take a novel approach to address this question, with analyses separated into stages to first examine single skills allowing the comparison with previous research, and then deal with issues of multicollinearity by examining the skills assessed by our test battery simultaneously in a principal components analysis. This approach will also allow us to clarify inconsistencies reported in the literature based on research using different selected individual differences tests.

The gold standard when studying reading is to analyse eye movement patterns. These patterns are considered to reflect cognitive processes active during reading (Liversedge & Findlay, 2000; Rayner, 1998). The two basic types of eye movements made while reading are fixations, which are moments where the eyes remain relatively still to acquire new information, and saccades, which are sweeping movements during which visual input is mostly suppressed (see Rayner, 2009 for a review). Fixations typically last 225-250 *ms* and saccades on average move across 7-9 letters in English and in similar alphabetic languages (Rayner, 2009). A reader also makes regressions, where a backwards saccade is made to re-fixate a point in the text previously encountered. Variation in these eye movement patterns often reflect the characteristics of the written language or the difficulty of a text. Fixations are longer in duration when processing loads are high, for example when reading low frequency words compared to high frequency words (Inhoff & Rayner, 1986). Generally, as a text increases in difficulty, fixation durations increase, saccades become shorter and more regressions are made (Rayner, 1998).

Importantly, variability in eye movement patterns also corresponds to reader skill. A skilled reader typically reads more quickly, makes shorter fixations, longer saccades and makes fewer regressions than beginner, less skilled or dyslexic readers (Rayner, 1998; Ashby et al., 2005). However, Underwood et al. (1990) demonstrated that there was no direct relationship between speed and reading skill when the latter was defined as the ability to efficiently extract information. Highly skilled readers were not necessarily those who demonstrated faster overall reading rates. A speed-accuracy trade-off exists whereby increasing reading speed eventually results in a decrease in comprehension (Rayner et al., 2016). It is important to note in this context that in eye movement research, a *skilled reader* is usually described as one who does not experience difficulties and the term is often used interchangeably with *average reader.*

Differences in eye movement behaviour can occur even within groups of skilled readers (see Andrews, 2012; 2015 for reviews). Fixation durations vary between 50 - 600 *ms* (sometimes longer), and saccade lengths vary between 1 - 20 letter spaces (or more) (see Rayner, 2009). Indeed, fixation durations, saccade lengths and frequency of regressions all vary widely between readers (Rayner, 1998). Differences are assumed to be quantitative rather than qualitative whereby all skilled readers use similar reading processes with varying efficiency (Ashby et al., 2005). Computational models of eye movement control, such as the E-Z Reader (Reichle et al., 1998; 1999), SWIFT (Engbert et al., 2005), OB1-Reader (Snell et al., 2018) and Über-Reader (see Reichle, 2020), each assume that readers use comparable word recognition processes. If qualitative processing differences were to be found between average and highly skilled readers, models would need to be adapted.

In general, low frequency words (words that are less commonly occurring within a language), are read more slowly and skipped over less often than more common, high frequency words (Angele et al., 2014; Rayner & Fischer, 1996; Rayner et al., 1996; Rayner et al., 2004). This indicates that individuals have greater familiarity with words that are more common within a language and can be explained by implicit learning (e.g., Dijkstra & Van Heuven, 2002; McClelland & Rumelhart, 1981; Morton, 1970). Repeated exposure to common words increases the baseline activation of that item in the lexicon, which means it will be processed and recognised more quickly (e.g., Monsell,1991). Some individuals will be more familiar with words that are less common within a language and as mentioned previously, differences in the size of this effect have been linked to individual differences in reading skill (Ashby et al, 2005) and are the main focus of this paper.

In the next section we will discuss the cognitive skills that will be examined in this paper starting with overall reading ability measures, followed by vocabulary, spelling, print exposure, rapid automatized naming (RAN) and finally working memory capacity (WMC). For these skills we will discuss relevant findings, theory where appropriate, and also mention any research that additionally examined the word frequency effect during reading in combination with testing these skills.

Firstly, reading ability as a whole can be assessed with standardised reading tests which usually consist of comprehension and vocabulary or other related sub-skills (e.g., word decoding and naming speed). Ashby et al. (2005) grouped readers into high and average proficiency based on the Nelson Denny Reading Test (NDRT; Brown et al., 1993) and examined differences in eye movements during silent reading. They found that highly skilled readers exhibited significantly shorter fixation durations than average readers and found a trend towards smaller word frequency effects for highly skilled readers in first and single fixation durations (the duration of fixations on the target word when the readers fixated the target word exactly once during the first pass) and a marginally significant interaction in the same direction for gaze duration (total time spent fixating a word on the first pass). These researchers did not observe much difference in fixation times on high frequency words when comparing the groups of readers, but there was a much greater difference in the fixation times on low frequency words. Less skilled readers showed a proportionately greater cost to reading times for low frequency words. Ashby et al. (2005) suggested that this may be due to the skilled readers' greater familiarity with less common words. They also found that readers spent more time on the next fixation following a low frequency word (increased spillover effects) compared to a high frequency word when the word is also unpredictable. Though the findings from Ashby et al. (2005) were based on small groups of participants (22 per group) and the interactions reported were marginal, the researcher's claims have been validated by further research in the field.

When standardised reading ability tests are included for research purposes, it has been noted that different cognitive abilities are addressed depending on which specific reading ability test is selected (Keenan et al., 2008; Kendeou et al., 2012). As a result, inconsistencies in research based on reading ability measures may arise dependent on whether the selected test includes specific sub-skills (e.g., word decoding) in addition to comprehension and vocabulary (see Mézière et al., 2023; Cutting & Scarborough, 2006). This is not an uncommon issue in psychological research, with a large range of different tests available to assess similar constructs, making it difficult to determine what the optimal test is to study these constructs (Flake & Fried, 2020). We included overall reading ability measures to examine trends reported by Ashby et al. (2005). Given the different tests available to measure reading ability, we made the rather unusual decision to include both the NDRT, which is often used in the US and Australia and the Wechsler Individual Achievement Test (WIAT-II UK; Wechsler, 2005), a normed test based on a UK adult sample, in the current study. It is important to remember that these tests are developed to diagnose reading difficulties. There are some notable format differences between these tests. For example, the NDRT features self-paced silent reading comprehension of non-fiction passages and a multiple-choice vocabulary test whereas the WIAT-II features a varied reading comprehension texts (featuring single line reading, fiction, and non-fiction paragraph reading, with the option to read aloud) alongside additional subtests (pseudoword decoding and word reading). Given the differences between these two reading ability tests and the potential of these differences impacting our findings, we decided to include both the NDRT and the WIAT-II.

In the context of our second cognitive skill, vocabulary, and some of the cognitive skills we discuss below, it is important to first discuss the Lexical Quality Hypothesis (Perfetti, 1992; 2007; Perfetti & Hart, 2002). Perfetti suggested that skilled readers develop high-quality lexical representations of words that in turn facilitate fast, automatic word recognition. A high-quality representation is defined as one with precision, redundancy and coherence (Perfetti, 2007). This means that it is represented by precise orthographic information that has little interference from similarly spelled words, it consistently triggers phonology of the correct word and is related to specific and redundant semantic information that defines the word. Highly skilled readers have high quality representations of most words (Andrews, 2008; 2012). Good and poor readers may, therefore, be differentiated due to the precision of their lexical representations (Perfetti, 2007; Perfetti & Hart, 2002).

Lexical quality can be divided into orthographic and phonological representations, of which spelling and decoding skills are a proxy, and semantic representations, measured by vocabulary and reading comprehension (Andrews, 2015; Perfetti, 2007). There is an indirect link from reading skills via lexical quality that may lead to a reduced word frequency effect: Individuals with good spelling and decoding skills, a larger vocabulary, better reading comprehension and more reading experience will develop higher quality lexical representations. In turn, they should have greater familiarity with less common words than the general population and require less time to process them than individuals who have less experience or are not as proficient in these cognitive skills.

These predictions about vocabulary size were confirmed in a study of monolingual and bilingual readers reading an entire novel (Cop et al., 2015). For the purpose of this paper, we focus on their examination of monolingual readers. Cop and colleagues found that vocabulary knowledge scores (measured by the LexTALE test; Lemhöfer & Broersma, 2012) indicated that readers with a larger vocabulary exhibited smaller frequency effects compared to those with a smaller vocabulary. This smaller frequency effect was mostly due to comparatively shorter single fixation durations for low frequency words, similar to the pattern based on NDRT scores from Ashby et al. (2005) which includes vocabulary and comprehension.

We now turn to the third cognitive skill of spelling. Good spelling in an opaque language such as English may be a good indicator of precise lexical representations. However, this is not likely to be the case in transparent languages such as Norwegian or Swedish where phonology maps onto orthography more directly, meaning unknown words can be more easily spelled (e.g., Furnes & Samuelsson, 2010). Within a study by Veldre et al. (2017), better spelling, but not reading ability (measured by the NDRT) in native English readers, was associated with a reduced effect of word frequency in gaze duration and total duration of fixations on target words. They found that, in general, higher spelling ability predicted longer saccades, more word skipping and landing positions that were further into the target word than low spelling ability but did not affect fixation times. Skipping probabilities are thought to reflect reaching an advanced stage of word recognition during parafoveal processing through the processing of the upcoming word viewed in the parafovea during a fixation on the previous word (e.g., E-Z Reader; Reichle et al., 1998; 1999). Precise lexical representations would therefore lead to rapid word identification during parafoveal processing, negating the need for the reader to directly fixate an already close-to-being identified word, and result in an increased likelihood of word skipping. Increased word skipping predicted by high spelling scores here is consistent with the idea that good spelling reflects high quality lexical representations which can be processed efficiently in the parafovea.

In the same study, Veldre et al. (2017) found that higher reading ability scores also predicted a reduction in the number of fixations and sentence reading times, smaller fixation durations and fewer regressions compared to lower reading ability scores but did not influence the size of the frequency effect. Even though reading and spelling abilities were moderately correlated in this study, spelling ability predicted differences in eye movements measures that were not predicted by reading ability, indicating that these skills influence different aspects of eye movement control (see also Slattery & Yates, 2017). This is consistent with differential effects of reading comprehension and spelling ability seen in parafoveal processing, (see Veldre & Andrews 2015a; 2015b). Spelling ability is a stronger and more reliable index of lexical precision than comprehension tests since readers who have poor lexical representations can compensate with context-based strategies, whereas spelling requires specific word form knowledge (Perfetti, 1992; Andrews, 2012).

Our fourth cognitive skill concerns print exposure (how much experience an individual has of reading novels). Greater reading experience enables readers to develop high-quality lexical representations which facilitate fast and efficient word recognition processes in skilled reading (Perfetti, 1985). Chateau and Jared (2000) studied university students with similar comprehension scores (using the NDRT comprehension subtest) but who had different levels of print exposure (as measured by the Author Recognition Test (ART; Stanovich & West, 1989) and found that students with higher levels of print exposure were more efficient in naming tasks and in lexical decision tasks than those with lower levels. They suggested that since the two groups were matched on comprehension, differences between the groups were due to differences in print exposure rather than general reading ability. Chateau and Jared (2000) suggested that print exposure increases general knowledge and enhances word recognition processes as well as contributing to vocabulary size.

Our fifth cognitive skill is Rapid Automatized Naming (RAN; Denckla & Rudel, 1974). RAN tasks measure an individual's ability to name a sequence of repetitive and visually familiar stimuli (letters, digits, colours or images) quickly and accurately. They are often considered good measures for predicting reading development. Alphanumeric RAN tasks (letters and digits) in particular, are often found to be predictive of early reading difficulties compared to RAN objects and colours (Bowey, 2005). Performance on a RAN task has been found to predict reduced word frequency effects in single sentence reading for readers who were not college-educated (Kuperman & Van Dyke, 2011). Kuperman and Van Dyke (2011) also found that faster RAN predicted smaller fixation durations and a decreased likelihood of regressions and refixations in a sample of low-to-average-skill readers.

However, when included in models alongside ART scores, Gordon et al. (2020) found that faster RAN did not influence the word frequency effect. Higher ART scores were associated with an increased likelihood of word skipping, shorter gaze durations, and a reduced effect of word frequency on gaze durations, consistent with Chateau and Jared (2000), whereas faster RAN times were instead associated with fewer regressions and rereading times and did not modulate the effect of word frequency in any measures. Gordon and colleagues suggested that the RAN and ART influence different aspects of the word recognition process, with faster RAN indicating efficient perceptual-motor coordination and attentional processes during reading. It has been suggested that the relationship between RAN and reading reflects phonological processes due to the oral reading aspect of the task (e.g., Wagner et al., 1994; Torgesen et al., 1997; Clarke et al., 2005; Ziegler et al., 2010). However, alternative ideas suggest that these observations may be due to rapid visual processing (originally suggested by Denckla, 1988), general processing speed (e.g., Cutting & Denkla, 2001; Powell et al., 2007; Kail & Hall, 1994; Savage et al., 2007; Wolf & Bowers, 1999), or, more recently, automaticity of lexical retrieval (see Jones et al., 2016). Indeed, there is little consensus about why a relationship between RAN and reading ability may appear (see Kirby et al., 2010 for a review). The present study includes RAN tasks to investigate their influence on eye movement patterns when considered alone and alongside other measures.

Finally, a "common cause" variable in the relationship between RAN and reading may be an individual's working memory capacity (WMC) (Papadopoulos et al., 2016). Daneman and Carpenter (1980) suggested that poor readers are characterised by inefficient processing and storage in WMC. One theory of WMC in reading, the Capacity Theory of Comprehension (Just & Carpenter, 1992), suggests that each reader has limited resources for storage and processing of information, and the capacity differs between individuals. When the resources needed to process information are more than an individual's capacity allows, processing efficiency is reduced, and information is lost. Therefore, when processing demands are high or capacity is very low, comprehension is diminished. Alternatively, Separate-Sentence-Interpretation-Resource (SSIR) Theory (Waters & Caplan, 1996), suggest that WMC is modular, with one module dedicated to syntactic processing, and the other dedicated to post-parsing sentences. The theory suggests that individuals differ only in the capacity of the second module, not the first. Therefore, individual differences in WMC should appear when looking at late eye movement measures, such as sentence reading times and wrap-up effects, when information is integrated towards the end of a sentence, especially when sentences are complex. Indeed, readers with smaller WMC have been found to spend more time re-reading ambiguous regions of a text than readers with larger WMC when sentences are complex (Clifton et al., 2003). In contrast, Joseph et al. (2015) did not find a direct influence of WMC on reading times and regression probabilities and number of studies have shown that the influence of WMC on reading behaviours is most likely because of shared variance with other cognitive skills (Hamilton et al, 2013; Traxler et al., 2012; Van Dyke, et al., 2014). Interestingly, Long and Freed (2021) found that readers with a greater WMC display smaller differences in gaze durations and sentence reading times for high and low frequency words than readers with a smaller WMC when considered alongside a large test battery of individual differences. Further research is therefore needed to assess these findings when considered alongside individual variation in other cognitive skills. A digit span task was selected for the current study on the basis that it would be less likely to correlate highly with other reading skills than a reading span task would. This approach was taken to best differentiate between working memory capacity and linguistic tasks.

Summarising, we selected the most common tasks that prior research had suggested may influence eye movement patterns; measures of reading ability overall; the NDRT (Brown et al., 1993) and the WIAT-II (Wechsler, 2005), the LexTALE test of vocabulary knowledge (Lemhöfer & Broersma, 2012), spelling (Andrews & Hersch, 2010), print exposure measured by the ART (Acheson et al., 2008), alphanumeric RAN (Denkla & Rudel, 1974), and WMC measured by a backwards digit span test (Gathercole et al., 2004). For simplicity, we will refer to the combination of reading skills and more general cognitive skills tests as cognitive skills.

As mentioned, the approach of the current paper was to examine the skills discussed here first separately, to provide clear comparisons with previous research. For brevity, only one eye movement measure was selected. We assessed individual skills as predictors of gaze duration. Gaze duration has been suggested as the eye movement measure most likely to reflect word recognition (Rayner, 1998). Next, skills were grouped via principal components analysis to reduce multicollinearity when assessing multiple skills simultaneously. We predicted that, where cognitive tests are considered in separate models, our findings would replicate previous research that linked high overall reading ability (Ashby et al., 2005), better spelling (Veldre et al., 2017), large vocabularies (Cop et al., 2015), high print exposure (Chateau & Jared, 2000; Gordon et al., 2020), faster RAN (Kuperman & Van Dyke, 2011) and larger WMC (Long & Freed, 2021) to shorter gaze durations and a reduced word frequency effect. We also expected skills related to lexical quality to be grouped together by PCA (Andrews, 2015; Perfetti, 2007). We expected lexical quality to be a good predictor of a reduced word frequency effect as seen in previous research about related skills (spelling; Veldre et al., 2017; vocabulary; Cop et al, 2015; print exposure; Chateau & Jared, 2000; Gordon et al.,2020). Since previous observations of the predictive value of RAN were associated with less skilled populations (e.g. Kuperman & Van Dyke, 2011), where RAN often predicted reading difficulties (Bowey, 2005) and was sometimes not found when included alongside measures of print exposure (Gordon et al., 2020), we hypothesised that the predictive value of RAN times may be secondary to lexical quality in our sample of average-to-very-skilled adult readers. We made similar predictions about WMC, due to suggested associations of WMC and RAN (Papadopoulos et al., 2016).

For completeness, we mention three studies that have examined individual differences within skilled reading in somewhat similar ways. The first was conducted by Kuperman and Van Dyke (2011) who found that individual differences in RAN and word identification tests influenced the magnitude of the word frequency effects on fixation times. They studied a different population of readers, who were non-college-bound participants from low to average reading skill, and took a different methodological approach using a corpus of natural sentence reading as opposed to an experimental design with tightly controlled materials. Kuperman et al. (2018) took an alternative approach to assessing both reader and text-level characteristics in passage reading using a Random Forests non-parametric regression technique to predict eye movement patterns and comprehension accuracy. We also note that different reading tasks were selected for individual differences measures in both of these studies. Most recently, a study was reported by Long and Freed (2021) who used test battery data collected by Freed et al. (2017) to predict eye movement patterns and found that language experience, decoding, and WMC interacted with word frequency to influence eye movement measures. A key difference in their approach is that besides a slightly different choice in the individual differences tests they used, they analysed eye movement data using structural equation modelling whereas this paper uses linear mixed models on individual trial data. Importantly, separate models were created by Long and Freed (2021) for the effects of word length and frequency, and predictability was not controlled for. In our design, words are matched on word length and predictability is controlled, which allows for a more controlled examination of the word frequency effect.

Method

Transparency and Openness

We report our sample size and how it was selected, all data exclusions based on individual differences and eye movements, all manipulations, and all measures in the study following JARS (Kazak, 2018). Data were analysed in R (version 4.1.1; R Core Team, 2022) using the in-built function ‘prcomp’ for the PCA and the lme4 package (version 1.1-27.1; Bates et al., 2015) for linear mixed models. The study design and analysis were not pre-registered. All data, analysis code and materials are available at https://osf.io/fetw4/?view_only=166f33ba5c584b3aa10efd30972840eb

Participants

One hundred participants consisting of students and staff from the University of Southampton as well as members of the wider community (20 Males, mean *age* = 22.54, *SD* = 9.78, range = 18-72) took part in the study. Participants were all native English speakers with normal or corrected to normal vision and no known reading difficulties. Student participants were sampled from various courses across the university including psychology, humanities and health sciences. Participants received either course credits (for psychology undergraduate students) or £9 for completing the study. Sample size was chosen so that the statistical power in this experiment would be comparable to previous experiments where 100 participants is often the norm for studies interested in individual differences. In our lab, we currently adopt the procedure of collecting data from 10 participants and running power analyses on that limited dataset to determine the desired number of participants. We calculated the power based on data from the first 10 participants using the simR package (Green &

MacLeod, 2016) which indicated that a sample size of 80 participants would be sufficient to achieve 80 % statistical power for an interaction between word frequency and LexTALE scores with the effect size extracted from the initial model. The LexTALE was selected as a measure likely to predict differences in the word frequency effect for skilled readers based on findings from Cop et al., (2015). Data collection exceeded this target, with a sample of 100 participants. Twelve participants' data were removed due to eye tracking errors and/or extreme low scores in reading tasks that may indicate reading difficulties (i.e., scores that corresponded to a "borderline" categorical reading level on the WIAT-II), meaning that 88 datasets (20 Males, mean *age* = 22.94, *SD* = 10.35, range = 18-68) were therefore included in analyses, still above the target sample size indicated by the power analysis. This research was reviewed and approved through the University of Southampton, Faculty of Environmental and Life Sciences Ethics Committee on 19/2/2019 (ERGO Ref. 47773).

Apparatus

Participants completed all tests and questionnaires via a web browser running Qualtrics on a 14-inch Dell Laptop, with the exception of the WIAT-II and digit span tests. The digit span test was administered using Inquisit on a 19-inch DELL monitor (1024 × 768-pixel resolution). For the WIAT-II (Wechsler, 2005), researchers used the testing flip-pad, scoring sheets and word/pseudoword cards included in the test pack and a stopwatch to record reading times.

Participants' eye movements were recorded during the eye tracking task using an SR Research EyeLink1000 eye tracker (sampling rate = 1000 Hz, max 0.5° calibration error). Sentences were presented on an ASUS HD monitor (1024 × 768-pixel resolution) at a viewing distance of roughly 73 cm to ensure 3 characters equated about 1° of visual angle. Stimuli were presented in Courier New font. Participants' head movements were minimised using a chinrest.

Materials

The study included a battery of tests to assess cognitive skills. Participants were first asked two questions about their reading behaviour: "How often do you read for work?" and "How often do you read for leisure?" (Response options were Not at all/ Not very often/ Occasionally / Quite often / Very often).

Reading Ability Tests

Nelson Denny Reading Test (Brown et al., 1993). The vocabulary subtest presented on the screen was comprised of 80 short descriptions of target words. Participants were asked to select the most appropriate word or phrase from five options to describe the key word's meaning. The reading comprehension subtest required participants to read passages silently before reading and answering multiple choice comprehension questions (with 5 possible answers) that appeared on the same page below the text. Test reliability for the NDRT ranges from .89 to .98 (Brown et al., 1993). This subtest was timed allowing participants 10 minutes to complete it, this is half of the standard time given for this test. This speeded procedure has been used previously and has proven to produce more normally distributed data than the full-timed procedure and can therefore obtain increased discrimination between more proficient readers (Andrews et al., 2020).

Wechsler Individual Achievement Test (WIAT-II UK; Wechsler, 2005). The word reading subtest required participants to read aloud a list of real words presented on paper that increased in difficulty in pronunciation, the test began with "the" and ended with "hierarchical". Participant's progress was recorded, and testing was discontinued when the participant made six sequential errors. The pseudoword decoding subtest required participants to read aloud a list of orthographically legal nonsense words, for example "flimp". Participants' progress was recorded and again testing was discontinued after the participant made six sequential errors. Participants also completed the reading comprehension subtest, where they read passages of increasing length and complexity (types of text included short fictional stories, informational text, advertisements, and how-to passages) aloud or silently before answering literal and inferential comprehension questions aloud when asked by the experimenter. The internal consistency reliability reported in the WIAT-II user manual is high (α = .98). The WIAT-II is normed with UK and US populations of children from 4 to 16 years, and adults

17 to 85 years. Scores for all WIAT-II subtests were combined, standardised and age-normed according to the examiners' manual.

Vocabulary Knowledge

LexTALE task (Lemhöfer & Broersma, 2012). An English word or pseudoword was presented on screen and participants were required to indicate whether the word was a real English word or a pseudoword. The test includes 40 real English words and 20 non-words. Scores are calculated as an average of the percentage of real words correctly identified and the percentage of nonwords correctly identified, therefore scores may range from 0 to 100 % (Lemhöfer & Broersma, 2012). Lemhöfer and Broersma (2012) report split-half reliability for the LexTALE test ranging from .64 to .81. It was initially created to provide a measure of English proficiency in second language learners, however the test has been used in studies comparing native and non-native adult speakers without performance reaching ceiling levels (e.g., Ernestus, et al., 2017).

Spelling

Spelling Dictation and Spelling Recognition (Andrews & Hersch, 2010). In the spelling dictation task participants listened to recordings of 20 single key words and sentences containing them. Participants were then asked to spell the key words correctly using a keyboard. The spelling dictation task was reported by Andrews and Hersch (2010) to have a test–retest reliability coefficient of .90. Scores were the number of words correctly spelled (range 0 – 20). In the spelling recognition task participants were given a list of 44 correctly spelled and 44 incorrectly spelled words and were asked to select all the words on the screen that they identified as being spelled incorrectly. Scoring is such that each correctly spelled word selected scores 1, and each incorrectly spelled word scores -1, therefore scores may range from -44 to 44. The spelling recognition was also reported to have very high test–retest reliability (.93) (Andrews & Hersch, 2010). These tasks were created to assess skilled reader's spelling abilities for research purposes.

Print Exposure

Author Recognition Test (Acheson et al., 2008). Participants were given a list of 65 real author names and 65 foil names presented on screen. Participants were asked to select any real authors they were familiar with. Participants were informed that there were some foil names included in the list. Scores were calculated as the total number of correct authors selected minus the number of foils selected. Since there were 65 real items and 65 foils, the highest possible score was 65 and the lowest possible score was -65. The ART has been found to have high test reliability (α = .84 in Stanovich & West, 1989; and α = .75 - .89 in a review by Mol & Bus, 2011). In addition, McCarron and Kuperman (2021) found it to be an informative and accurate measure of print exposure for Native English university students.

Rapid Automatized Naming

Alphanumeric RAN (Denkla & Rudel, 1974). A series of letters were presented in random order on screen and participants were required to verbally name the characters as quickly as possible while being timed by the experimenter. This was then repeated with a series of digits. Total RAN times were calculated as the sum of the time taken to read the letters and the time taken to read the digits (recorded in seconds). Gordon et al., (2020) reported high test-retest reliability for RAN digits (.89) and RAN letters (.85).

Working Memory Capacity

Digit Span Backward (Gathercole et al., 2004). Number sequences were presented on screen and participants were asked to recall their order in two different tasks: Digit Span Forward (the participant attempts to repeat randomised digit sequences of increasing length forwards), and Digit Span Backward (the participant attempts to repeat randomised digit sequences of increasing length backwards). Mean scores calculated from the largest successful backward digit span sequences (before two consecutive errors) were used in our analyses. Gathercole et al., (2004) report the mean test–retest reliability coefficient of the backwards digit span task to be .62.

Eye Tracking

Participants were required to read sentences on screen while their eye movements were tracked. Related questions were then presented on the screen and participants were asked to respond using the keyboard to ensure they were reading for comprehension. Sentences featured one word that was either low or high frequency (e.g., “gourd” versus “apple”). There were 78 high and 78 low frequency target words matched on word length. All were nouns with a word length between four and nine characters ($M = 5.40$, $SD = 0.87$). The high frequency words had significantly higher Zipf values (Sublex-UK Zipf word frequencies: Van Heuven et al., 2014), ($M = 5.28$, $SD = 0.36$) than the low frequency words ($M = 3.22$, $SD = 0.43$), t (147) = -32.73, $p < .001$. Each pair of target words was embedded in the same neutral sentence frame. Eight participants who did not take part in the main experiment completed a cloze task: they were provided with the experimental sentences up to the target word and were asked to provide single words that could follow this partial sentence. Only 1.28 % of guesses were correct, demonstrating that target words were not predictable from the initial sentence context. Participants read all 78 experimental sentences, with either the high or low frequency target words in each. Stimuli were organised across two lists using a Latin square design. Testing began with 12 practice trials with practice questions. Comprehension questions followed 26 % of the sentences.

Design and Procedure

Participants were given an information sheet and consent form. Participants were first asked to answer some reading background questions. Task order was then randomised apart from eye tracking which always presented in the middle of testing. Participants were allowed breaks where needed. Participants were first asked to complete the NDRT, WIAT-II, LexTALE, Spelling, ART, alphanumeric RAN and backwards digit span tasks.

During the eye tracking task, participants were asked to sit comfortably at the computer, resting their chin on a chinrest and were guided through the set up and calibration of the eye tracker by the researcher. Participants were then required to direct their gaze to a fixation cross presented on the left of the screen. When ready, sentences were presented following the fixation cross. Participants were asked to read and answer questions presented on the screen using the keyboard to respond to ensure they were reading for comprehension.

Results

Analytic Approach

Our analyses were organised into three stages: First, separate models predicting gaze durations were examined for each individual difference test to assess whether findings replicate previous research which often included only one of these tests. Second, a principal components analysis was performed to determine which tasks load together. This stage identified two components which were extracted and used as composite scores in subsequent models. Third, GLMM models were tested to examine eye movement measures (word skipping probabilities, first fixation durations (FFD), single fixation durations (SFD), gaze durations (GD), go past times and sentence reading times) and whether they may be predicted by the following fixed factors: the two identified components (PC1 and PC2), two reading comprehension subtests (from the NDRT and WIAT-II), and a word reading subtest (from the WIAT-II), word frequency, saccade launch site and trial number. Two-way interactions between a single individual differences score (PC1, PC2, NDRT comprehension, WIAT-II comprehension or WIAT-II word reading) and target word frequency were also considered.

Individual Differences Tests

To correct for the unequal proportion of real words and non-words in the LexTALE test, scores were calculated by taking the mean of the percent of correct responses (yes) for real words and the percent of correct responses (no) for nonwords ((number of correct words/40 * 100) + (number of correct nonwords/20 * 100))/2) (Lemhöfer & Broersma, 2012). An overall spelling score was calculated as an average of the two subtests, spelling recognition and spelling dictation[1] (see Andrews & Hersch, 2010). For the ART, total scores were collated whereby a correct selection of a real author scored 1 and an incorrect selection of foil name scored -1. The time taken to read aloud both RAN digits and RAN letters was calculated as an overall time in seconds. The backward version of the digit span was taken as an overall score (see Gathercole et al., 2004). Overall scores for the NDRT were calculated as an average of comprehension and vocabulary subtest scores. Overall WIAT-II scores were calculated using the age adjusted scoring materials provided which resulted in numerical scores as well as categorical assessments of reading proficiency (Borderline/ Low Average/ Average/ High Average/ Superior). For the purpose of examining these tests as predictors of eye movements, in light of their use in previous research, overall composite test scores were used for models which included only one test as a predictor. However, for the PCA, it is more appropriate to consider the cognitive skills that are included within reading ability tests separately as the skills assessed differ between the NDRT and WIAT-II and may align more closely with other tasks (e.g., it is reasonable to suggest that NDRT vocabulary scores will be more closely associated with vocabulary knowledge than with reading comprehension). Therefore, for our PCA and subsequent GLMMs, subtests from the WIAT-II and NDRT were included instead. These consisted of reading comprehension (WIAT-II and NDRT), word reading (WIAT-II), pseudoword decoding (WIAT-II), and vocabulary (NDRT). Outliers were examined using the interquartile range (IQR) criterion for each test and as mentioned 12 participants were removed due to very low scores on one or more measures[2]. Of this 12, two were also flagged due to technical difficulties during eye tracking. Descriptive statistics for scores on each test are summarised in Table 1.

---

[1] Spelling subtests (Recognition and Dictation) were highly correlated, $r = 0.71$. As a result, we followed procedures by Andrews and Hersch (2010) to calculate an overall spelling score.

[2] WIAT-II comprehension (2 outliers), NDRT comprehension (1 outlier), NDRT Vocabulary (2 outliers), WIAT-II word reading (2 outliers), WIAT-II pseudoword reading (2 outliers), LexTALE (4 outliers), Spelling (2 outliers), RAN (2 outliers), ART (1 outlier).

Table 1

Descriptive Statistics for Tests and Subtests

	Test Range	Min	Max	Mean	*SD*
NDRT Total	0 - 78	33.50	68.00	50.72	8.02
NDRT Comprehension	0 - 76	10.00	60.00	33.77	12.32
NDRT Vocabulary	0 - 80	51.00	79.00	67.65	6.79
WIAT-II Total	40 - 160	86.00	127.00	108.94	9.61
WIAT-II Comprehension	40 - 160	71.00	125.00	106.25	13.57
WIAT-II Pseudoword Decoding	40 - 160	82.00	118.00	103.89	8.12
WIAT-II Word Reading	40 - 160	100.00	121.00	111.48	4.62
LexTALE	0 - 100	75.00	97.50	89.94	5.86
Spelling	0 - 54	9.5	31	21.11	5.14
Spelling Dictation	0 - 20	6	20	12.82	3.34
Spelling Recognition	-44 - 44	9	44	29.42	7.65
Author Recognition Test	-65 - 65	-2.00	37.00	11.66	7.64
Rapid Automatized Naming	(seconds)	29.98	53.85	40.74	5.29
Digit Span Test	2 - 15	3.83	10.83	6.13	1.32

Eye Tracking Analyses

Comprehension accuracy during the eye tracking task was very high (*mean accuracy* = 95.6%). Trials with blinks on the target word or featuring tracking loss were removed prior to analysis. Fixations shorter than 80 *ms* that were made within one character of a previous or subsequent fixation were merged. Afterwards, remaining fixations shorter than 80 *ms* and longer than 800 *ms* were removed. Additionally, trials that consisted of fewer than 3 fixations across the sentence were also removed. This stage of data cleaning removed 0.18% of data. Trials where the target word was skipped were removed from analyses based on fixation times. This stage removed 6.26 % of data for the fixation time analyses. Dependent variables were then calculated; first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), go past times (GOPAST), sentence reading times and word skipping probabilities[3]. Word skipping probability, single fixation duration, first fixation

---

[3] Split half reliabilities of the word frequency effect in the eye movement DVs were calculated (SFD, r = 0.33; FFD, r = 0.47, GD, r = 0.30; GOPAST, r = 0.32; Skipping Probability, r = 0.07; Total Reading Time, r = 0.18).

duration and gaze duration were selected as appropriate 'early' eye movement measures (measures that relate to the first time a word is read, before rereading; Rayner, Sereno, et al., 1989). Go past times and total sentence reading times were selected as 'late' eye movement measures (measures that relate to eye movements made after first pass reading of a target word, including rereading). When calculating these eye movement measures, data falling outside of 3 standard deviations from the mean for each participant within a condition (high and low frequency) were removed (SFD; 2.18 %, FFD; 2.12 % GD; 2.83 %, go past times; 2.74 %, Sentence reading times; 1.96 %). Descriptive statistics based on participant means for these measures are displayed in Table 2 below.

Table 2

Descriptive Statistics for Eye Movement Measures

	Condition	Min	Max	Mean	*SD*
SFD	HF	132.81	277.71	204.65	27.84
	LF	140.21	312.29	232.86	36.76
FFD	HF	131.80	280.28	204.50	27.46
	LF	143.68	324.41	229.66	34.66
GD	HF	135.53	313.12	217.83	33.22
	LF	156.97	420.68	258.01	48.31
GOPAST	HF	135.53	313.12	272.18	36.20
	LF	156.97	442.29	282.34	53.29
Sentence Reading Times	HF	1324.27	4321.74	2587.02	612.67
	LF	1224.90	4779.45	2693.63	657.91
Skipping Probability	HF	0.00	0.49	0.22	0.12
	LF	0.00	0.42	0.14	0.11

Note. Means and *SDs* are calculated based on participant means per condition.

The reliabilities of the word frequency effect in these eye movement measures were comparable to those noted by Staub (2021).

Generalized Linear Mixed Models

A gamma distribution was used for all Generalised linear mixed models (GLMM), following guidance for analysing skewed reaction time data without transformation (see Lo & Andrews, 2015) with participants and items as random factors.

Models with Single Test Predictors

Separate GLMMs were conducted to model each test focusing exclusively on gaze durations. Fixed factors included the test score measuring the cognitive skill in question, word frequency and the interaction of the test score with word frequency, as well as trial number and launch site. Trial number was never found to be significant in any model and did not contribute to the model fit and was therefore never included in final models. Launch site indicates the distance from which a saccade was launched prior to the target word. Previous research has documented that when a saccade is launched close to a target word, fixation times are often reduced and skipping rates increase (Pollatsek et al., 1986). The following procedure was followed to obtain the final model (see also Dirix & Duyck, 2017). The random effects structure started with intercepts only for subjects and items. The fixed structure was trimmed backwards by removing an interaction or fixed factor and using pairwise chi-square model fit comparisons to determine whether this removal negatively influenced the model fit. However, we always maintained the test itself as fixed factor given that this was the focus of the analysis. Following this, the random effects structure was forwards fitted to find the largest possible structure that converged again using pairwise chi-square model fit comparisons to see if an extra parameter added to the fit of the model. Once the maximal random effects structure possible had been established, we finally checked whether parameters in the fixed structure could be removed.

Model outputs are presented in Table 3 below. To correct for multiple comparisons, Benjamini and Hochberg's (1995) false-discovery rate correction was used to calculate adjusted *p* values. Gaze Duration models consistently revealed that high frequency words received significantly shorter gaze durations than low frequency words and that saccades launched from a position close to the target word were significantly more likely to result in shorter gaze durations on the target word. Analyses revealed that - when considered in separate models - high scores in the NDRT, WIAT-II, LexTALE and Spelling tests predicted shorter gaze durations than low scores. After correcting for multiple comparisons, the main effect of LexTALE scores on gaze durations was marginal. NDRT, WIAT-II and LexTALE- models revealed that scores on these tests significantly moderated the impact of word frequency on gaze durations. High scorers on these tests were not slowed down as much as low scorers when encountering a low frequency word within a sentence (see Figure 1). Note that scores on both reading ability tests are composites of two or more subtests, the NDRT is comprised of comprehension and vocabulary and the WIAT-II includes comprehension, word reading and pseudoword decoding. It was therefore unclear whether comprehension or other skills included in these composites were key in driving the effects found in these models.

Table 3

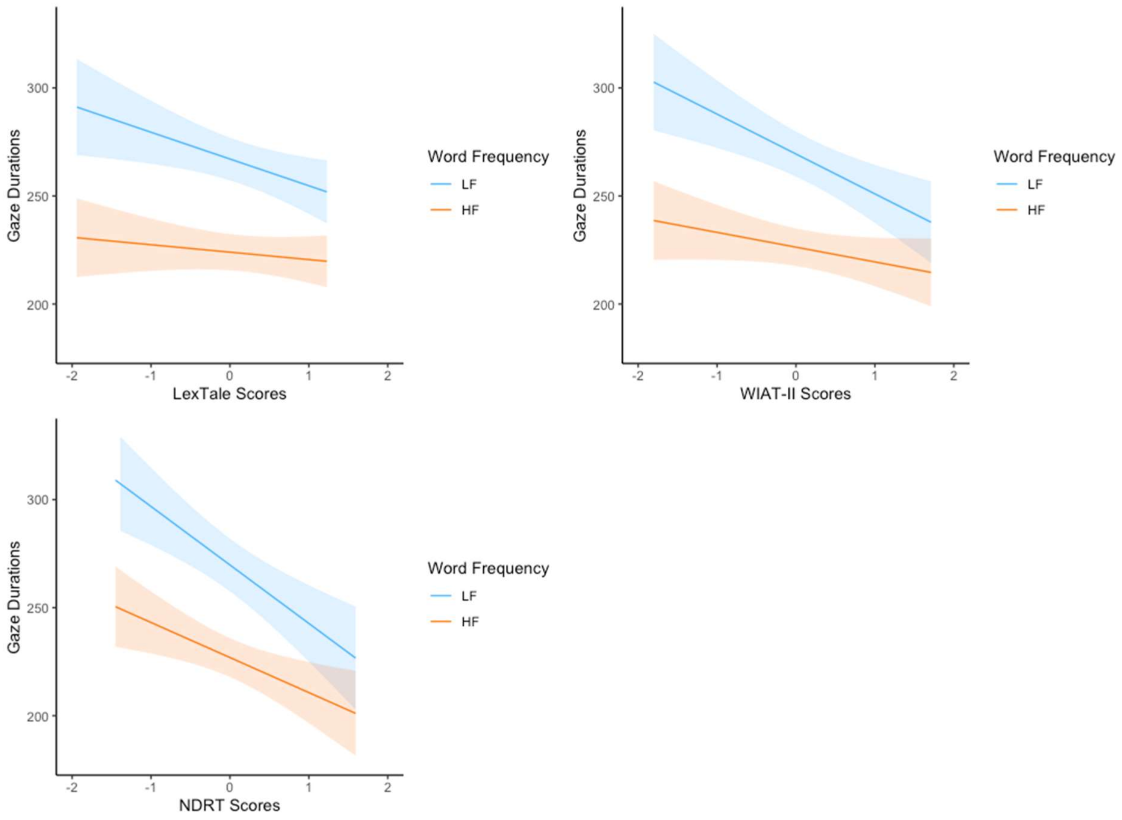
GLMMs to predict Gaze Durations where Models included a Single Test Predictor

		Est	*SE*	*t*	*p*
WIAT-II	Intercept	248.00	4.53	54.81	<.001 ***
	Word Frequency (LF-HF)	-43.11	3.81	-11.30	<.001 ***
	Launch Site	4.76	0.76	6.30	<.001 ***
	WIAT-II Overall Score	-12.64	4.29	-2.95	.004 **
	WIAT-II Overall Score * Word Frequency	11.61	3.99	2.91	.004 **
NDRT	Intercept	248.47	5.01	49.64	<.001 ***
	Frequency	-42.84	4.46	-9.62	<.001 ***
	Launch Site	4.84	0.76	6.36	<.001 ***
	NDRT Overall Score	-21.64	5.77	-3.75	<.001 ***
	NDRT Overall Score * Word Frequency	10.84	4.17	2.60	.011 *
LexTALE	(Intercept)	245.67	4.25	57.78	<.001 ***
	Word Frequency (LF-HF)	-43.12	3.89	-11.07	<.001 ***
	Launch Site	4.24	0.57	7.42	<.001 ***
	LexTALE	-7.91	3.99	-1.98	.055.
	LexTALE * Word Frequency	8.95	4.15	2.16	.037 *
Spelling	Intercept	247.79	4.53	54.75	<.001 ***
	Word Frequency (LF-HF)	-42.23	4.05	-10.44	<.001 ***
	Launch Site	4.82	0.75	6.43	<.001 ***
	Spelling	-13.06	4.20	-3.11	.003 **
	Spelling * Word Frequency	4.98	3.94	1.27	0.212
ART	Intercept	245.34	4.71	52.06	<.001 ***
	Word Frequency (LF-HF)	-41.64	4.21	-9.88	<.001 ***
	Launch Site	4.17	0.59	7.13	<.001 ***
	ART	-7.77	3.98	-1.95	.058.
	ART * Word Frequency	-	-	-	
RAN	Intercept	244.98	4.48	54.73	<.001 ***
	Word Frequency (LF-HF)	-41.34	4.98	-8.30	<.001 ***
	Launch Site	4.18	0.58	7.26	<.001 ***
	RAN	3.45	3.98	0.87	0.386
	RAN * Word Frequency	-	-	-	-
Backwards Digit Span	Intercept	245.02	4.64	52.77	<.001 ***
	Word Frequency (LF-HF)	-42.12	4.04	-10.42	<.001 ***
	Launch Site	4.11	0.58	7.14	<.001 ***
	Backwards Digit Span	-6.15	3.81	-1.61	.118
	Backwards Digit Span * Word Frequency	5.15	3.32	1.55FSIMR	.128

Note. Significance is denoted by * < .05, ** < .01, *** < .001. *p* values are adjusted for multiple comparisons via Benjamini and Hochberg's (1995) false-discovery rate correction.

Figure 1

The Effect of Word Frequency on Gaze Durations (ms) Moderated by LexTALE, WIAT-II and NDRT Scores



Note. Shaded areas represent 95% confidence intervals.

Principal Components Analysis

As mentioned previously, WIAT-II and NDRT subtests were included separately when assessing grouping variables via principal components analysis. Scores on all tasks within the test battery were centred to allow comparisons to be made. Correlations between all tests within the test battery are presented in Table 4. Tests were moderately positively correlated, except for RAN and backwards digit span tests where smaller interactions were observed. A PCA was conducted on all tests and subtests to identify which tests loaded together using the in-built function ‘prcomp’ in R (version 4.1.1 R Core Team, 2022). This procedure was used to decompose the correlation matrix and reduce the dimensions for further analysis. Two principal components were identified using parallel analysis which calculates adjusted eigenvalues for the data based on random noise expected via a simulated parallel dataset (Horn, 1965; see Figure 2). This method provides guidance about which components should be accepted for interpretation and further analyses. Principal components which fall above the mean of the random eigenvalues generated by the simulated data should be retained. The two components that meet this criterion in our data (PC1 and PC2) collectively contributed 49.23 % of the variance within the data. An individual test loading on a component was considered important if its contribution exceeded 10 % (expected average contribution calculated from 1/number of variables = 1/10).

PC1 was explained by the vocabulary subtest from the NDRT, Spelling, ART, LexTALE and the pseudoword decoding subtest of the WIAT-II (see Figure 3). This component uniquely contributed 35.40 % of the variance. We suggest that these tests are related to lexical quality (Perfetti, 2007) though we will refer to them as 'lexical proficiency' to avoid confusion with the theoretical construct. PC2 was explained predominantly by the RAN and to a lesser extent the backwards digit span task, and uniquely contributed 13.83 % of the variance (see Figure 4). RAN tasks have been interpreted as an index of general processing speed in cognitive tasks (e.g., Cutting & Denkla, 2001; Powell et al., 2007; Kail & Hall, 1994; Savage et al., 2007; Wolf & Bowers, 1999). Working memory is often suggested to be a "common cause" variable in the relationship between RAN and reading (Papadopoulos et al., 2016). Therefore, the assumption that working memory is related to the RAN when predicting reading ability is supported by the finding that they load together on this component. The remaining tests (NDRT comprehension, WIAT-II comprehension and WIAT-II word reading) fell outside of these two components and were considered separately in subsequent analyses. Surprisingly, the two comprehension subtests (NDRT and WIAT-II) were not found to load on the same component. We will return to this issue in the general discussion.

Table 4

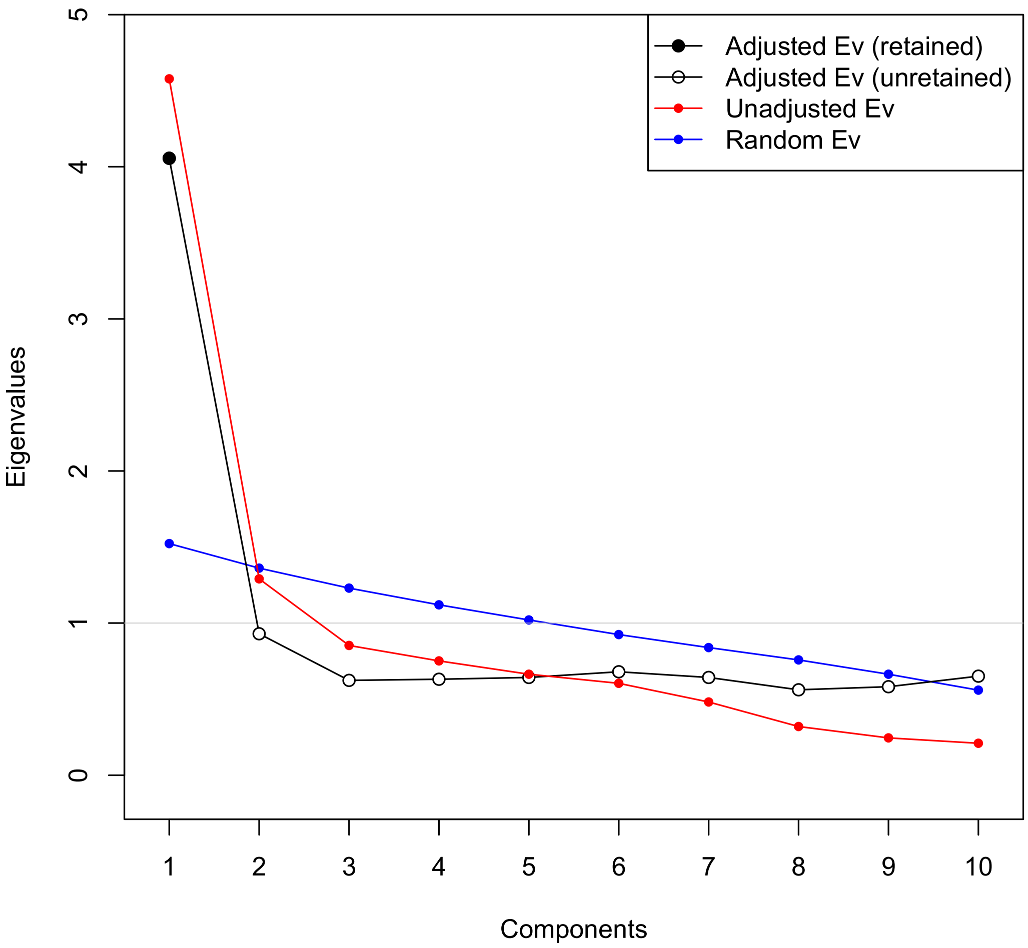
Correlations between Subtests

	NDRT Comp	NDRT Vocab	WIAT-II Comp	WIAT-II Pseudo	WIAT-II Word Read	Spelling	ART	LexTALE	Digit Span
NDRT Vocabulary	0.36***								
WIAT-II Comprehension	0.21	0.51**							
WIAT-II Pseudoword Decoding	0.13	0.30**	0.22**						
WIAT-II Word Reading	0.25*	0.32	0.12	0.44***					
Spelling	0.38***	0.48***	0.27*	0.47***	0.33**				
ART	0.22*	0.58***	0.36***	0.35***	0.24*	0.51***			
LexTALE	0.19	0.48***	0.29**	0.33**	0.43***	0.41***	0.36***		
Digit Span	0.08	0.22*	0.08	0.15	0.14	0.30**	0.06	0.16	
RAN	0.11	0.07	-0.17	-0.27*	0.00	-0.15	0.10	-0.20	-0.24*

Note. Significance is denoted by * < .05, ** < .01, *** < .001.

Figure 2

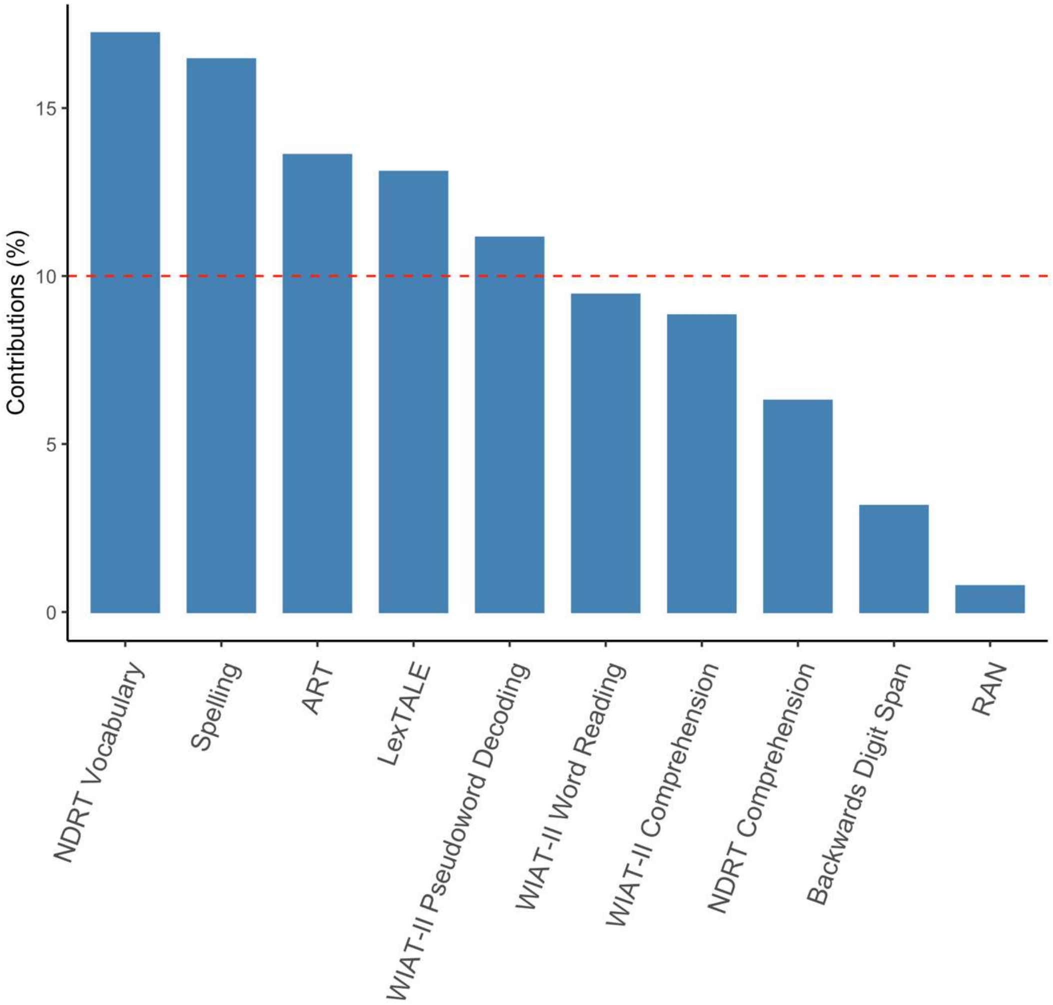
A Plot of Principal Components Identified within the Full Test Battery using Parallel Analysis



Note. Random Ev (blue) refer to randomly generated eigenvalues from a simulated parallel dataset. Unadjusted Ev (red) are the eigenvalues given by the real data. Adjusted Ev (black) show these eigenvalues adjusted to account for expected noise in the data, filled points on this line represent principal components which fall above the mean of the random eigenvalues (1), these components are retained for further analysis.

Figure 3

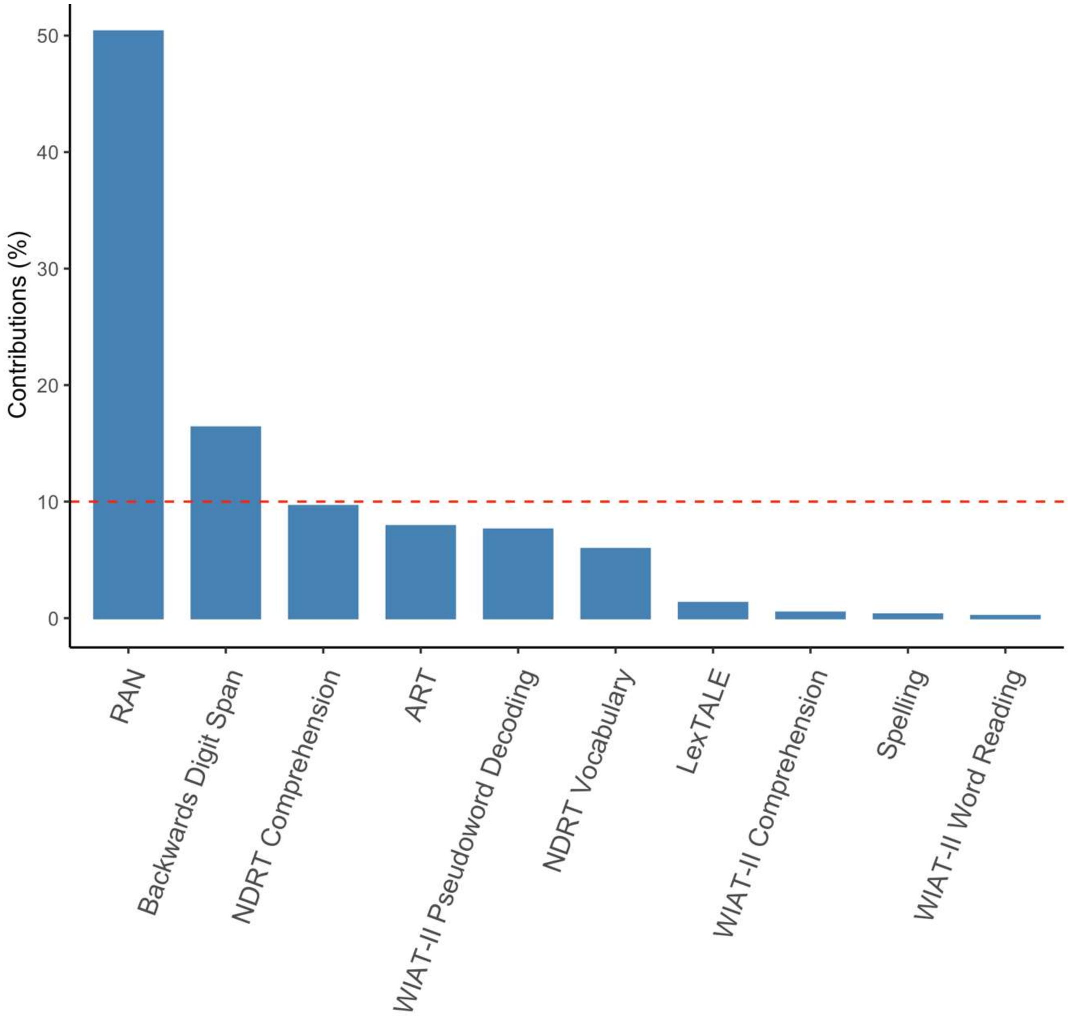
Individual Contributions of Each Individual Differences Measure on PC1



Note. The dotted line represents the expected average contribution (10 %). A contribution above this line is considered important in explaining the component.

Figure 4

Individual Contributions of Each Individual Differences Measure on PC2



Note. The dotted line represents the expected average contribution (10 %). A contribution above this line is considered important in explaining the component.

Analysis of Grouping variables

GLMMs were conducted for FFD, SFD, GD and go past times using a gamma distribution (as used in the single predictor models as well) with participants and items as random factors. Skipping probabilities were modelled via GLMMs with a binomial distribution. Model trimming followed the same procedure as for the single test models. First, models included all fixed effects: word frequency, trial number, launch site, PC1, PC2, the reading comprehension subtests NDRT and WIAT-II, the word reading subtest of the WIAT-II and two-way interactions between each individual differences test and word frequency (three-way interactions never contributed to the fit). Model comparison Chi-square tests were conducted to investigate whether non-significant fixed effects or interactions could be trimmed without reducing the model fit.

Next, we began to build up the random effects structure to find a converging model closest to the maximal model. Random effects were forward fitted, adding slopes in order of theoretical importance, starting with word frequency, the individual difference test, launch site and trial number. These effects were retained if they contributed to the model fit. Finally, any non-significant fixed effects were examined whether they could be trimmed again.

Focusing first on the main effects, models revealed that low frequency words received longer first and single fixations, gaze durations and go past times and a lower likelihood of being skipped than high frequency words (Tables Table 5 and Table 6). In line with previous research (Pollatsek et al., 1986), when saccades were launched nearby the target word, it was subsequently fixated for shorter durations (FFD, SFD, GD and go past times) and skipped more often. Participants who scored highly in tests associated with PC1 had shorter gaze durations and go past times, and more word skipping than participants with lower scores[4]. High scores on the WIAT-II word reading subtest were associated with shorter go past times than low scores.

Turning to the interactions, we found that scores associated with PC1, and WIAT-II comprehension influenced the relationship between word frequency on go past times. Participants who scored highly in the tests associated with PC1 were less negatively impacted by a low frequency word embedded within a sentence than low scorers (see Figure 5). High scorers on the WIAT-II

---

[4] For clarity, PC1 was negatively associated with the vocabulary subtest from the Nelson Denny Reading Test, Spelling, ART, LexTALE and the pseudoword decoding subtest of the WIAT-II. As such, results in Table 5 and 6 demonstrate that higher PC1 scores are associated with longer gaze durations, go past times, sentence reading times and increased skipping probabilities.

comprehension test also exhibited smaller differences in go past times between high and low frequency words than those with low scores. However, we observe that go past times for low frequency words were fairly stable across WIAT-II comprehension scores, and instead, we see shorter go past times for high frequency words read by low scorers. The difference here is unexpected and does not fit the pattern we find for WIAT-II overall scores when modelled separately on gaze durations (Figure 1) or for WIAT-II comprehension in sentence reading times, as we will discuss next.

Sentence reading times were included in analyses since previous studies have indicated that some individual differences are more closely related to late eye movement measures. Better spellers make longer saccades and better comprehenders read sentences more quickly and make fewer fixations (Veldre et al., 2017), those with a faster RAN score make fewer regressions and refixations (Kuperman & Van Dyke, 2011), and those with a large WMC display shorter sentence reading times (Long & Freed 2021). Sentence reading time models included these fixed effects: word frequency, trial number, PC1, PC2, the reading comprehension subtests NDRT and WIAT-II, the word reading subtest of the WIAT-II and two-way interactions between each individual differences test and word frequency and were built using the procedure described for previous eye movement models. Results revealed that a low frequency word embedded within a sentence significantly slowed the total reading time of that sentence (Table 7). Trials occurring later during testing featured shorter sentence reading times than earlier trials. Participants who scored highly in tests associated with PC1 and PC2, and also in NDRT comprehension, WIAT-II comprehension and word reading subtests read sentences more slowly than low scorers on these tests. Participants who scored highly in the tests associated with PC1, PC2, WIAT-II comprehension and WIAT-II word reading tasks were also less slowed down by a low frequency word in sentence reading times (see Figure 6).

Table 5

GLMMs with Multiple Test Predictors to predict Skipping Probability, First Fixation Durations and Single Fixation Durations

	Skipping Probability				FFD				SFD			
	Est	*SE*	*t*	*p*	*Est*	*SE*	*t*	*P*	*Est*	*SE*	*t*	*p*
Intercept	-2.67	0.17	-15.41	<.001***	220.57	3.54	62.31	<.001***	224.69	4.52	49.67	<.001***
Frequency	0.68	0.08	8.70	<.001***	-24.81	3.76	-6.60	<.001***	-29.19	4.17	-7.00	<.001***
Trial Number	0.00	0.00	1.73	.083	-	-	-	-	-	-	-	-
Launch Site	-0.54	0.04	-13.81	<.001***	1.48	0.62	2.38	.017*	1.86	0.42	4.45	<.001***
PC1	-0.17	0.06	-2.85	.004**	1.75	2.04	0.86	.392	-	-	-	-
PC2	-	-	-	-	-	-	-	-	-	-	-	-
NDRT Comprehension	-	-	-	-	-	-	-	-	-	-	-	-
WIAT Comprehension	-	-	-	-	-	-	-	-	-	-	-	-
WIAT Word Reading	-	-	-	-	-	-	-	-	-	-	-	-
Frequency * PC1	-	-	-	-	-	-	-	-	-	-	-	-
Frequency * PC2	-	-	-	-	-	-	-	-	-	-	-	-
Frequency * NDRT Comprehension	-	-	-	-	-	-	-	-	-	-	-	-
Frequency * WIAT Comprehension	-	-	-	-	-	-	-	-	-	-	-	-
Frequency * WIAT Word Reading	-	-	-	-	-	-	-	-	-	-	-	-

Note. Significance is denoted by * < .05, ** < .01, *** < .001.

Table 6

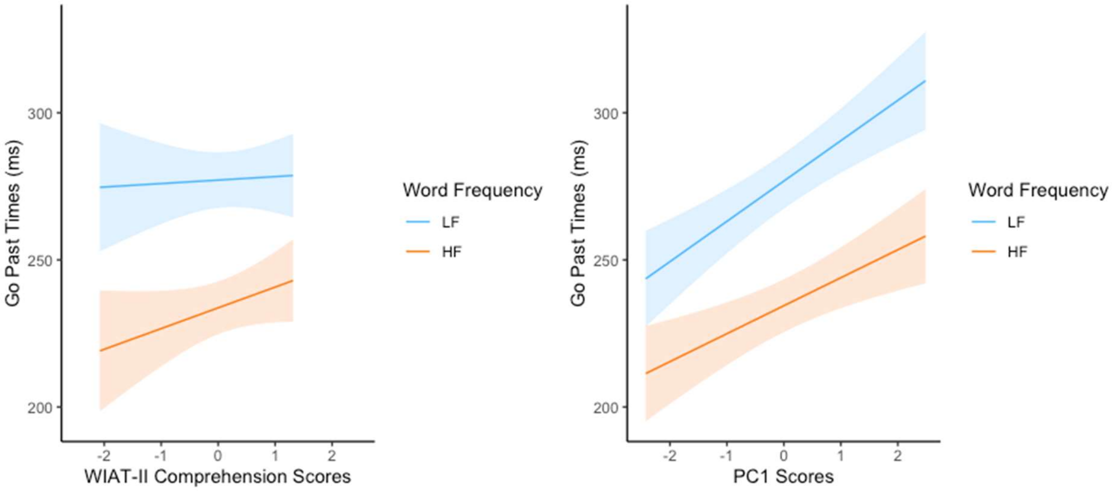
GLMMs with Multiple Test Predictors to predict Gaze Durations and Go Past Times

	GD				GOPAST			
	Est	*SE*	*t*	*p*	*Est*	*SE*	*t*	*p*
Intercept	245.87	3.85	63.94	<.001***	253.26	4.51	56.11	<.001***
Frequency	-41.50	3.78	-10.97	<.001***	-43.27	3.35	-12.91	<.001***
Trial Number	-	-	-	-	-	-	-	-
Launch Site	4.82	0.76	6.38	<.001***	5.73	0.73	7.83	<.001***
PC1	5.15	2.23	2.31	.021*	11.61	2.73	4.26	<.001***
PC2	-	-	-	-	-	-	-	-
NDRT Comprehension	-	-	-	-	-	-	-	-
WIAT Comprehension	-	-	-	-	4.11	4.28	0.96	.337
WIAT Word Reading	-	-	-	-	9.81	4.72	2.08	.038*
Frequency * PC1	-	-	-	-	-4.19	1.24	-3.39	<.001***
Frequency * PC2	-	-	-	-	-	-	-	-
Frequency * NDRT Comprehension	-	-	-	-	-	-	-	-
Frequency * WIAT Comprehension	-	-	-	-	5.86	2.63	2.23	.026*
Frequency * WIAT Word Reading	-	-	-	-	-	-	-	-

Note. Significance is denoted by * < .05, ** < .01, *** < .001.

Figure 5

The Effect of Word Frequency on Go Past Times (ms) as a Function of PC1 and WIAT Comprehension Scores



Note. Shaded areas represent 95% confidence intervals.

Table 7

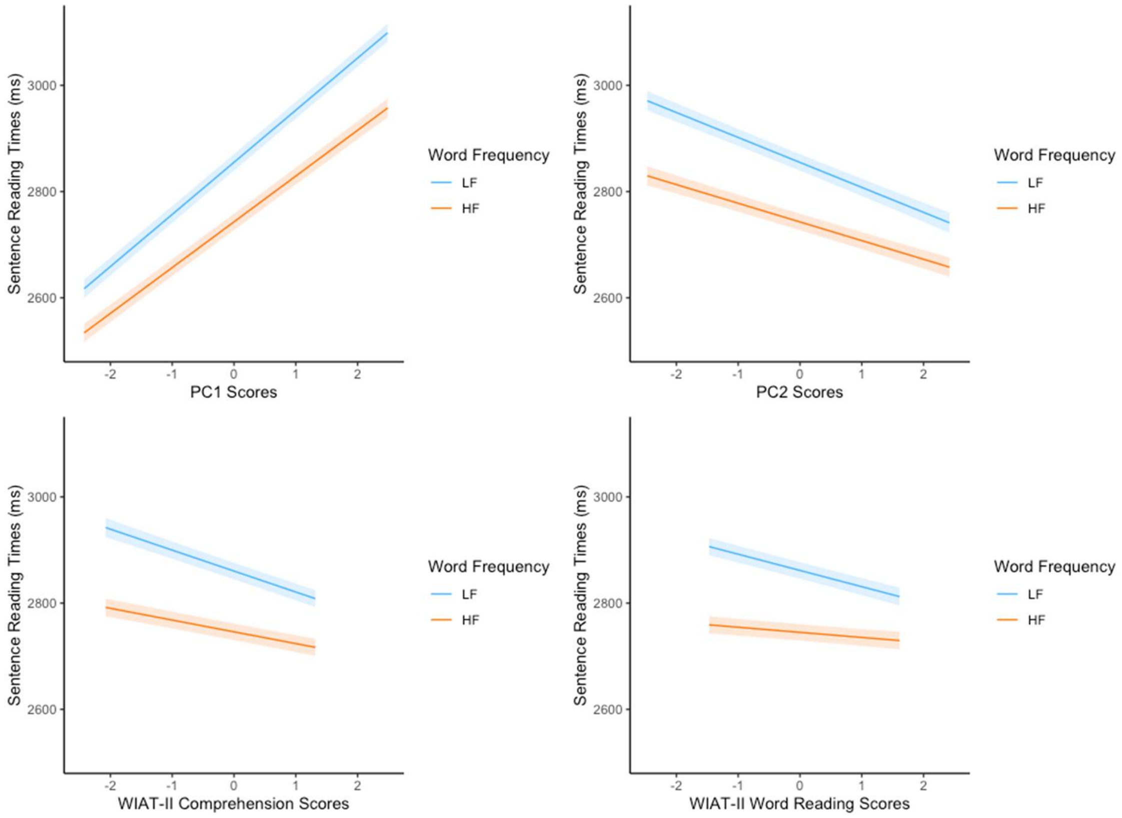
GLMM with Multiple Test Predictors to predict Sentence Reading Times

	Est	*SE*	*t*	*p*
Intercept	3040.72	1.37	2226.43	<.001***
Frequency	-118.93	1.36	-87.27	<.001***
Trial Number	-4.39	0.15	-29.19	<.001***
PC1	92.14	1.31	70.12	<.001***
PC2	-41.07	1.88	-21.81	<.001***
WIAT Comprehension	-30.78	1.23	-24.99	<.001***
WIAT Word Reading	-20.07	1.62	-12.38	<.001***
NDRT Comprehension	-113.07	1.42	-79.68	<.001***
Frequency * PC1	-11.94	1.39	-8.58	<.001***
Frequency * PC2	11.83	1.46	8.11	<.001***
Frequency * WIAT Comprehension	17.30	2.45	7.07	<.001***
Frequency * WIAT Word Reading	20.96	1.47	14.29	<.001***
Frequency * NDRT Comprehension	-	-	-	-

Note. Significance is denoted by * < .05, ** < .01, *** < .001.

Figure 6

The Effect of Word Frequency on Sentence Reading Times (ms) as a Function of PC1, PC2, WIAT-II Comprehension and WIAT-II Word Reading Scores



Note. Shaded areas represent 95% confidence intervals.

General Discussion

Our approach to assessing the influence of individual differences on the effect of word frequency consisted of first examining separate models in which each test was a single predictor – besides word frequency and the interaction - to allow for a more straightforward comparison with the existing literature. Tests were then grouped based on overlapping variance using a PCA. Finally, eye movement patterns were modelled using these factors alongside individual tests that were not grouped as predictors. This approach followed a more appropriate method to examine the influence of multiple skills alongside one another whilst reducing multicollinearity. Before delving into the three stages of our analytic approach, we mention that throughout our analyses all models showed the well-documented effect of word frequency on eye movement measures (Rayner & Fischer, 1996; Rayner et al., 1996, Rayner et al., 2004), whereby high frequency words received shorter fixation times and were more likely to be skipped than low frequency words. In addition, all models for word-level measurements (first and single fixation durations, gaze durations, go past times and skipping probabilities) demonstrated that when saccades were launched close to the target word, it was subsequently fixated for shorter durations and skipped more often, again in line with previous research (e.g., Pollatsek et al., 1986).

Models with Single Test Predictors

Shorter gaze durations were associated with high scores in both overall reading ability tests (NDRT and WIAT-II), vocabulary knowledge and spelling. These findings are consistent with the idea that shorter gaze durations are associated with better readers compared with average readers (Ashby et al, 2005). However, shorter gaze durations were not predicted by faster RAN or higher scores on the backwards digit span task, suggesting that efficient word identification is facilitated more by top-down processes related to comprehension and word knowledge than bottom-up processes related to WMC and faster processing of single characters.

In addition, we replicated previous research that linked high reading abilities (Ashby et al., 2005) to a reduced word frequency effect and partially replicated research based on large vocabularies (Cop et al., 2015)[5]. Both overall reading ability measures, the NDRT and WIAT-II, and the LexTALE

---

[5] Trends in our data were fairly consistent with Cop et al. (2015) who analysed eye movements of 14 English-speaking monolinguals when reading an entire book (this was a baseline to compare with bilinguals, who read half of the book in their first language and half in their second language). Previous research has indicated that reading single sentences can produce quite different patterns in reading behaviour compared to reading paragraphs. Radach et al. (2008) demonstrated that frequency effects are more pronounced when reading sentences rather than passages.

Cop et al. (2015) observed the interaction between LexTALE and word frequency in SFDs rather than GDs. SFD analyses in the current study did not reveal a significant main effect of the LexTALE (t = - 0.51, p = .612) or any effect of the LexTALE on the relationship between word frequency and SFDs (t = 1.45, p = .146). There were clear effects of word frequency (t = -7.68, p < .001) and saccade launch site (t = 3.92, p < .001).

task moderated the impact of the low frequency word on reader's gaze durations. High scorers on these tests were not slowed down by a low frequency word as much as low scorers were.

Next, we consider spelling. As part of a study looking at spaced vs unspaced text, Veldre et al. (2017) observed that in normally spaced text, higher spelling ability predicted smaller fixation durations and a reduced frequency effect on all duration measures. In our sample, we found that highly skilled spellers had significantly shorter gaze durations than less skilled spellers, but spelling ability was not found to predict differences in the size of the frequency effect (though trends were in the same direction as Veldre et al., 2017). This discrepancy may be due to a difference in modelling as Veldre and colleagues included NDRT scores within the same models as spelling, whereas we (initially) report models with single predictors, in this case spelling. The influence of spelling ability on gaze durations supports the Lexical Quality Hypothesis (Perfetti, 1992; 2007; Perfetti & Hart, 2002) in that higher quality lexical representations appear to result in more efficient word processing (measured by gaze durations, Rayner, 1998). When considering word frequency, it is not a prerequisite of a good speller to be more familiar with low frequency words, though spelling ability and vocabulary size are often correlated. Therefore, we demonstrate the distinction between high spelling abilities and large vocabularies in this research. Though a good speller may process all words more efficiently than a poor speller, a larger vocabulary may be a better predictor of more efficient processing of less common words in particular.

As predicted by previous findings from Moore and Gordon (2015), individuals with higher ART scores had shorter gaze durations in our eye tracking task, which is consistent with the idea that greater experience leads to more efficient word processing. Print exposure measured by this test has also been previously linked to higher reading ability and a reduced effect of word frequency on gaze durations (Moore & Gordon, 2015). However, an interaction between ART scores and word frequency was not replicated in our data. Kuperman and Van Dyke (2013) offer an explanation for a null effect of print exposure on the size of the word frequency effect. They observed that when print exposure is matched, differences in eye movements in response to high and low frequency words still occurred (Kuperman & Van Dyke, 2013). The authors suggested that it may not be as simple as measuring the amount of exposure to words that determines greater quality lexical representations – but rather that highly skilled readers are better at utilising experience with words to create high-quality lexical representations. This distinction is likely to be subtle in the literature since this ability often correlates with print exposure.

In addition, we note that there are some differences in our sample of readers compared to the sample that the ART was based on for the UK (Acheson et al., 2008). One discrepancy is that ART scores in the current study (M = 11.66, SD = 7.64) were much lower than scores in the sample from Acheson et al. (2008), when this test was adapted for UK use (M = 22.7, SD = 10.8). The ART is based on the idea that people who read a lot are likely to encounter the names of a greater number of authors than people who do not read as much. However, Moore and Gordon (2015) note that the ART is less effective at discriminating differences for lower scores. They also suggested authors appearing on the ART created in 2008 may be somewhat out of date, and therefore may be less useful for determining print exposure for the college-aged samples in 2023[6]. We agree, and also propose that in all likelihood it is now increasingly possible to be a voracious reader without reading printed books, or any books for that matter, due to the accessibility of online text for other reasons (e.g., news articles, emails, social media). In this instance we may not expect everyone who reads often to be more familiar with the names of authors. Researchers should consider improvements suggested by Moore and Gordon (2015) when considering the ART for future experiments.

---

[6] Moore and Gordon (2015) conducted a factor analysis on the ART and concluded that two factors were present rather than one clear 'print exposure' factor. They theorized that one of these factors was related to academic or literary reading rather than reading for leisure as the test originally intended to measure (West et al., 1993). They also identified a positive correlation between selecting author names and foil names, suggesting that some participants used a lower criterion for deciding that they know an author, this correlation was moderate in our sample ($r = 0.38$, $p < .001$). Moore and Gordon suggested that stricter penalties for selecting a foil name or including a measure of self-rated confidence could provide more information about a participant's criteria.

Principal Components Analysis

Unsurprisingly, we found that scores across most of the test battery were correlated, since individual performances on cognitive tasks often are (Deary, 2000), which make it difficult to interpret individual skills as unique predictors of reading. A principal components analysis revealed two components, allowing related skills to be grouped where variances overlap. The first, lexical proficiency (PC1), was negatively associated with scores on the NDRT vocabulary test, spelling, ART, LexTALE, and WIAT-II pseudoword decoding (higher scores on these tests were associated with lower scores on PC1). We suggest that this variable is related to lexical quality based on the associated skills in line with our second prediction that skills associated with lexical quality would feature overlapping variance. We refer to PC1 as lexical proficiency to avoid confusion with the theoretical construct of lexical quality.

The second factor (PC2) was associated with RAN (faster scores on the RAN were associated with higher scores on PC2) and to a lesser extent the backwards digit span test (more items recalled in the backwards digit span test were associated with higher scores on PC2). We suggest that this variable describes participants' speed of processing. Three tests did not load on these components and were treated as distinct skills in subsequent analyses: NDRT comprehension, WIAT-II comprehension and WIAT-II word reading. It should be noted that word reading was close to the threshold for inclusion in PC1, and had it been included it would not have been surprising since it is related to word knowledge which appears to be a commonality across PC1 measures. Neither of the two comprehension measures included in the test battery (NDRT and WIAT-II) loaded on the extracted components PC1 and PC2, though we note that the NDRT comprehension was close to the threshold for inclusion in PC2. Perhaps more surprisingly, the comprehension subtests did not load together despite supposedly measuring the same construct, this important finding will also be discussed later.

Analyses with Identified Components

Models that included the two identified components and other distinct tests revealed main effects that are consistent with previous research that found more skilled readers identify words more quickly, read sentences faster and skip over words more often than less skilled readers (Rayner, 1998; Ashby et al., 2005). PC1 was predictive of shorter gaze durations, go past times, sentence reading times and higher skipping probabilities when individuals scored highly on the associated tests. Higher skipping probabilities may suggest that participants who score highly on tests associated with PC1 may be advancing faster in the recognition process of the parafoveal word and subsequently make a decision to skip it (Inhoff & Rayner 1986; White, 2008), alternatively they may employ riskier saccade targeting strategies (O'Regan 1990, 1992; Rayner & Fischer, 1996; Rayner, 1998). A risky reading strategy is where some readers (often seen in older adults; Rayner et al., 2006) compensate for less advanced lexical processing by frequently guessing upcoming words, skipping them and often returning when wrong guesses are realised. We suggest some directions for experiments in this area. First, a follow-up experiment should feature more difficult questions to encourage reading for comprehension, since a risky strategy is likely to have been quite successful in the present study. In addition, researchers should utilise a boundary paradigm to investigate how multiple reading-related skills influence the efficient extraction of information in the parafovea. Some previous studies conducted by Veldre and Andrews (2014; 2015b) suggest that an increased capacity to extract information from the parafovea is related to reading and spelling abilities.

We predicted that a factor reflecting the shared variance of skills related to lexical quality would be a strong predictor of a reduced word frequency effect as suggested by previous research about each contributing skill[7] (spelling; Veldre et al., 2017; vocabulary; Cop et al., 2015; print exposure; Chateau & Jared, 2000; Gordon et al., 2020). This prediction was only partially upheld as

---

[7] It should be noted that factors obtained via PCA only reflect shared variance amongst the skills measured, filtering out residual variance that is unique to individual measures. We can therefore only draw conclusions about proxies of latent variables such as lexical quality that ignore the nuance of each cognitive task.

our lexical proficiency factor PC1 influenced the relationship between word frequency only in some eye movement measures, but not others. In comparison to low scorers, high scorers exhibited shorter go past times and shorter sentence reading times for low frequency words. These findings are in late eye movement measures suggesting that for skilled readers, very precise lexical representations are associated more with faster embedding of meaning into sentence context, rather than faster orthographic decoding of low frequency words.

An important comment needs to be made regarding not finding an interaction between PC1 and word frequency in earlier measures. Our reported Linear Mixed Models were created using a pruning strategy that aimed to achieve the largest random effects structure possible (see Barr et al., 2013). We suspect pruning techniques used in previous years would often be comparable to intercept only models. If we run intercept only models for the current analyses, consistent interactions between PC1 and word frequency can be found in SFD, FFD and GD, and an WIAT-II comprehension by word frequency interaction in GD can be found. We mention this as it could explain parts of the discrepancies with previous related research. In addition, earlier sections of our own analyses (on single test predictors) found differences in the magnitude of the word frequency effect in gaze durations, which were not found when looking at models based on multiple test predictors. One reason for this may be that PCA groups factors via variance shared amongst the tests included. There may be unique features of the LexTALE and reading ability tests (NDRT & WIAT-II) that are key in predicting this pattern, or perhaps more likely, there is a reduction in statistical power to estimate an effect where more parameters are estimated within the model, as in the current analysis.

PC2 was highly associated with the RAN and to a lesser extent the working memory measure (backwards digit span test). High scores were associated with faster processing of information within whole sentences, but not of single words. In addition, a small reduction in the word frequency effect was found associated with high PC2 scores in sentence reading times which may reflect a small decrease in rereading time following an uncommon word for readers with faster processing speeds. We first consider the RAN, as it influences PC2 more than any other measure, and again mention that previous research suggests that the RAN is more predictive of later measures in the eye tracking record such as refixations, regressions, foveal-on-parafoveal effects and second pass reading times (e.g., Gordon et al., 2020). The observation that PC2 is predictive of shorter sentence reading times in our data therefore supports the idea that the RAN reflects “efficient coordination of perceptual-motor and attentional processing during reading” (Gordon et al., 2020, p. 553).

PC2 was also somewhat reflective of higher scores on a backwards digit span task (WMC), which supports the idea that working memory is a “common cause” in the relationship between the RAN and reading (Papadopoulos et al., 2016). Overall, results supported the prediction that lexical quality was a stronger predictor of differences in eye movements for average-to-very-skilled readers than RAN and WMC. However, future research should focus on longer passages for reading comprehension and should continue to assess later measures of eye movements in relation to variables associated with WMC and RAN scores to fully assess the literature.

Our test battery featured two comprehension subtests, first we discuss the WIAT-II. This comprehension subtest was not found to influence early eye movement measures in models with multiple test predictors, but was found to predict sentence reading times, a later eye movement measure, and also influenced the relationship between word frequency and eye movements in both sentence reading times and go past times. High scores were associated with shorter sentence reading times when a low frequency word was embedded compared to low scores. However, in go past times, the pattern observed was slightly different, low scorers were found to spend less time reading a high frequency word before moving on than high scorers. No difference was observed for low frequency words. We interpret this pattern with caution as we note that scores on this test are clustered towards high scores, with very few low scores (note that the low scores did not fall outside 2.5 SDs). We suspect this unusual pattern might be due to strategic differences in whether a processing difficulty is resolved by making a regression or by first finishing the entire sentence. Sentence reading times were in line with the more established pattern that the low frequency word impacts better readers less.

Earlier, we discussed the finding that the overall WIAT-II score (which includes comprehension, word reading and pseudoword decoding) was associated with differences in gaze durations in the model when it was the only predictor entered and interacted with the word frequency effect on gaze durations. It may be that effects seen in early measures were driven by a subtest related to lexical proficiency: pseudoword decoding (which was associated with PC1, and generated similar effects in the models with multiple test predictors) whereas comprehension in the WIAT-II influenced later eye movement measures in the bigger models.

Turning to the NDRT, we found no main effect associated with the NDRT comprehension subtest on the word-based measures (but do on sentence reading times), or interactions between this test and word frequency in any of the models with multiple test predictors. Note that in the earlier reported analyses, when the entire NDRT (including the comprehension subtest) was the only predictor, it led to a main effect and interaction with word frequency on gaze durations. A likely hypothesis is that the reduced word frequency effect seen in associated with NDRT total was driven by the vocabulary subtest (associated with PC1).

We observe here some discrepancies in findings related to comprehension based on which test is selected. High WIAT-II comprehension scores were associated with a reduced word frequency effect in later eye movement measures whereas high NDRT comprehension scores were not in any measures. Since these two comprehension measures were included separately, it could be argued that some of the variance due to NDRT comprehension may have overlapped with variance attributed to WIAT-II comprehension. However, these tests did not load together in the PCA and were not very highly correlated ($r = 0.21$). We suspect that this demonstrates a case of the jingle fallacy (Thorndike, 1904), – the false assumption that two instruments measure the same construct because they share a name, namely that both tests are measures of a specific "reading comprehension skill" and we think this finding has important repercussions for comparing research that used one of these tests.

Indeed, previous research has indicated that different cognitive abilities are addressed by different tests of reading ability (Keenan et al., 2008; Kendeou et al., 2012). We note some qualitative differences in the way that these tests are conducted that might indicate such differences in the constructs measured. The WIAT-II comprehension test includes a variety of texts including single sentences, fiction and non-fiction passages with the option to read aloud or silently, whereas the NDRT consists of non-fiction passages that participants read silently and the passages remains accessible while individuals answer self-paced questions on screen. Previous research has suggested that differences in the format of reading materials (sentences vs paragraphs; Radach et al., 2008; fiction vs non-fiction; e.g., Graesser et al., 1998; Zwaan, 1994) and in reading strategy (reading aloud or silently; e.g., Hale et al., 2007) can alter the behaviour measured. Additionally, the use of non-fiction passages in the NDRT may mean this test is more related to general knowledge than reading comprehension, as suggested by Coleman et al. (2010) who found greater than chance levels of correct answers after administering a "passageless" version of the test to college students (see also Ready et al., 2013).

Although both tasks were timed, during the NDRT participants were unaware of the time limit until they were asked to stop. We therefore have no reason to expect that participants adopted an unnatural reading pace. The time given to complete the NDRT comprehension subtest in this study was half of the time usually allocated for this test which may produce increased discrimination of scores between more proficient readers (Andrews et al., 2020) but participants were not aware of this. In contrast, participants were aware of the timed element in the WIAT-II comprehension test since the test is administered face-to-face and the timing is explicit in the instructions. However, measures of WIAT-II reading speed (m = 360.84 ms, SD = 89.61 ms) and NDRT words per minute (m = 272.18, SD = 74.36) were highly correlated (r = -0.65), therefore, we do not expect time pressure differences to be the cause of any differential predictive value between the tests. Instead, we propose that the face-to-face aspect of testing in the WIAT-II comprehension task might have influenced participants' performance. This procedural difference could lead to performance anxiety especially where some sections of the test required participants to read aloud and where questions had to be answered directly to the experimenter, thereby creating qualitative differences in the experience of the participant in the NDRT versus the WIAT-II.

Importantly, the jingle fallacy found here may explain some inconsistent results across studies that model comprehension on different tests (e.g., Mézière et al., 2023; Cutting & Scarborough, 2006). However, it is unclear whether such findings are actually in disagreement (failure to conceptually replicate) or whether they are simply measuring separate latent constructs as suggested here when two comprehension tests were not grouped by a single latent construct in our PCA. An analysis of all the conflicting findings in the literature that may be attributed to weak convergent validity is beyond the scope of this paper but should be considered in future research. We advise that these comprehension tests should not be used interchangeably since they appear to tap into independent aspects of reading comprehension. Researchers should proceed with caution when selecting a reading comprehension test in future research.

Summary

This study investigated the patterns of individual differences in skilled reading that have been mentioned in the literature, that is, generally faster reading is associated with higher reading ability (e.g., Ashby et al., 2005; Veldre et al, 2017; Moore & Gordon, 2015;). Our focus was also on the size of the frequency effect given that a comparatively smaller frequency effect during reading has been suggested as indicative of better reading skills (Ashby et al., 2005; Haenggi & Perfetti, 1994).

We first examined separate models in which each test was a single predictor – besides word frequency and the interaction - to allow for a more straightforward comparison with the existing literature. Broadly in agreement with the literature, higher skill was associated with shorter gaze durations in two reading ability tests (WIAT-II and NDRT), and in spelling. This pattern was marginally significant for higher vocabulary knowledge scores (LexTALE). A reduced word frequency effect was associated with reading ability tests and vocabulary knowledge but was not found to be associated with other tests included in our test battery.

We grouped shared variance across our test battery into two grouping variables, lexical proficiency and speed of processing. The two comprehension tests (NDRT and WIAT-II) were not grouped into one of the two grouping factors and surprisingly were not grouped in a single factor. We urge researchers to be cautious when selecting one of these two comprehension tests for future research as these tests did not load together and demonstrated an example of the jingle fallacy. We discussed reasons why these two tests might have qualitatively different ways in measuring comprehension.

When all measures were included in models with multiple test predictors, eye movements often associated with skilled reading (shorter fixation times and higher skipping rates) were most consistently related to a factor we identified as lexical proficiency. Our findings support the Lexical Quality Hypothesis (Perfetti, 1992; 2007; Perfetti & Hart, 2002) in that precise lexical representations support faster word recognition processes as a reduced word frequency effect was associated with higher lexical proficiency (PC1).

Constraints on Generality

This study sampled average-to-very-skilled readers, and it should be noted that differences based on skill may be more clearly observed between less skilled and average readers. Though the sample included a wide age range and included some participants from the local community, a large proportion were undergraduate psychology students.

References

Acheson, D. J., Wells, J. B., & Macdonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students, *Behavioural Research, 40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Andrews, S. (2008). Lexical expertise and reading skill. *Psychology of Learning and Motivation, 49*, 247-281. https://doi.org/10.1016/S0079-7421(08)00007-8

Andrews, S. (2012). Individual differences in skilled visual word recognition and reading: The role of lexical quality. In J. Adelman (Ed.), *Visual Word Recognition* (pp. 151-172). Hove: Psychology Press.

Andrews, S. (2015). Individual differences among skilled readers: The role of lexical quality. In A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading* (pp. 129–148): Oxford University Press.

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General, 139*(2), 299–318. https://doi.org/10.1037/a0018366

Andrews, S., Veldre, A. & Clarke, I. E. (2020). Measuring Lexical Quality: The Role of Spelling Ability. *Behavior Research Methods, 52,* 2257-2282. https://doi.org/10.3758/s13428-020-01387-3

Angele, B., Laishley, A. E., Rayner, K., & Liversedge, S. P. (2014). The effect of high- and low-frequency previews and sentential fit on word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1181–1203. https://doi.org/10.1037/a0036396

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 58*(6), 1065–1086. https://doi.org/10.1080/02724980443000476

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

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289-300. http://www.jstor.org/stable/2346101.

Bowey, J. A. (2005). Predicting individual differences in learning to read. In M. J. Snowling & C. Hulme (Eds.), *Blackwell handbooks of developmental psychology. The science of reading: A handbook* (pp. 155-172). Malden, Blackwell Publishing. http://dx.doi.org/10.1002/9780470757642.ch9

Brown, J. A., Fishco, V. V., & Hanna, G. (1993). *Nelson-denny reading test: Manual for scoring and interpretation, forms G & H*. Rolling Meadows, IL: Riverside Publishing

Chateau, D., & Jared, D. (2000) Exposure to print and word recognition processes. *Memory and Cognition 28,* 143–153. https://doi.org/10.3758/BF03211582

Clarke, P., Hulme, C., & Snowling, M. (2005). Individual differences in RAN and reading: a response timing analysis*. Journal of Research in Reading, 28*, 73–86. https://doi.org/10.1111/j.1467-9817.2005.00255.x

Clifton C.., Traxler, M.J., Mohamed, M.T., Williams, R.S., Morris, R.K., & Rayner, K. (2003) The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language. 49*, 317–334. https://doi.org/10.1016/S0749-596X%2803%2900070-6

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg K. N. (2010) Passageless comprehension on the Nelson-denny reading test: well above chance for university students. *Journal of Learning Disabilities 43*(3):244-9. https://doi.org/10.1177/0022219409345017.

Cop, U., Dirix, N., Drieghe, D. & Duyck, W (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods 49*, 602–615 https://doi.org/10.3758/s13428-016-0734-0

Cop, U., Keuleers, E., Drieghe, D. & Duyck, W. (2015) Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review 22*, 1216–1234. https://doi.org/10.3758/s13423-015-0819-2

Cutting, L. E., & Denckla, M. B. (2001). The relationship of rapid serial naming and word reading in normally developing readers: An exploratory model. *Reading and Writing, 14,* 673-705. https://doi.org/10.1023/A:1012047622541

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading, 10*(3), 277-299.

Daneman, M., & Carpenter, P. A. (1980) Individual differences in working memory and reading, *Journal of Verbal Learning and Verbal Behavior, 19(*4), 450-466. https://doi.org/10.1016/S0022-5371(80)90312-6.

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain.* Oxford, England: Oxford University Press.

Denckla, M. B. (1988). *"What does it take to learn (after you've learned to read)?*" in Paper Presented at the Research Symposium Conducted at the Meeting of the New York Orton Society, (New York, NY, USA: Literacy Assistance Center).

Denckla, M. B., & Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex, 10*, 186-202. http://dx.doi.org/10.1016/S0010-9452(74)80009-2

Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175–197. https://doi.org/10.1017/S1366728902003012

Dirix, N., & Duyck, W. (2017). An eye movement corpus study of the age-of-acquisition effect. *Psychomomic Bulletin & Review, 24*(6), 1915–1921. https://doi.org/10.3758/s13423-017-1233-8

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*, 777-813. https://doi.org/10.1037/0033-295X.112.4.777

Ernestus, M., Kouwenhoven, H., & Van Mulken, M. (2017). The direct and indirect effects of the phonotactic constraints in the listener's native language on the comprehension of reduced and unreduced word pronunciation variants in a foreign language. *Journal of Phonetics, 62,* 50-64. https://doi.org/10.1016/j.wocn.2017.02.003

Freed, E. M., Hamilton, S. T., & Long, D. L. (2017). Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language, 97*, 135–153. doi:10.1016/j.jml.2017.07.008

Flake, J.K., & Fried, E.I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science. 3*(4):456-465. https://doi.org/10.1177/2515245920952393

Furnes, B., & Samuelsson, S. (2010) Predicting reading and spelling difficulties in transparent and opaque orthographies: a comparison between Scandinavian and US/Australian children. *Dyslexia. 16*(2):119-42. http://dx.doi.org/10.1002/dys.401.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Gordon, P.C., Moore, M., Choi, W., Hoedemaker, R.S., & Lowder, M.W. (2020) Individual differences in reading: Separable effects of reading experience and processing skill. *Memory & Cognition, 48*, 553–565 https://doi.org/10.3758/s13421-019-00989-3

Graesser, A. C., Kassler, M. A., Kreuz, R. J., & McLain-Allen, B. (1998). Verification of statements about story worlds that deviate from normal conceptions of time: What is true about Einstein's Dreams? *Cognitive Psychology, 35*(3), 246–301. https://doi.org/10.1006/cogp.1998.0680

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

Haenggi, D., & Perfetti, C. A. (1994). Processing components of college-level reading comprehension. *Discourse Processes, 17*(1), 83-104. http://dx.doi.org/10.1080/01638539409544860

Hale, A. D., Skinner, C. H., Williams, J., & Hawkins, R., Neddenriep, C. E., & Dizer, J. (2007). Comparing comprehension following silent and aloud reading across elementary and secondary students: Implication for curriculum-based measurement. *The Behavior Analyst Today. 8*(1), 9. https://doi.org/10.1037/h0100101.

Hamilton, S.T., Freed, E.M., & Long, D.L. (2013). Modelling reader and text interactions during narrative comprehension: A test of the lexical quality hypothesis, *Discourse Processes, 50*:2, 139-163, https://doi.org/10.1080/0163853X.2012.742001

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika 30*, 179–185. https://doi.org/10.1007/BF02289447

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics, 40*(6), 431–439. https://doi.org/10.3758/BF03208203

Jones, M. W., Snowling, M. J., & Moll, K. (2016). What automaticity deficit? Activation of lexical information by readers with dyslexia in a rapid automatized naming Stroop-switch task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(3), 465. https://doi.org/10.1037/xlm0000186

Joseph, H. S. S. L., Bremner, G., Liversedge, S. P., & Nation, K. (2015) Working memory, reading ability and the effects of distance and typicality on anaphor resolution in children, *Journal of Cognitive Psychology,* 27:5, 622-639, https://doi.10.1080/20445911.2015.1005095

Just M.A., & Carpenter P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kail, R., & Hall, L. K. (1994). Processing speed, naming speed, and reading. *Developmental Psychology, 30*(6), 949. https://doi.org/10.1037/0012-1649.30.6.949.

Kazak, A.E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1–2. https://doi.org/10.1037/amp0000263

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281–300. https://doi.org/10.1080/10888430802132279

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*(5), 354–367. https://doi.org/10.1016/j.learninstruc.2012.02.001

Kirby, J. R., Georgiou, G. K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45*, 341–362. https://doi.org/10.1598/RRQ.45.3.4

Kuperman, V., Matsuki, K., & Van Dyke, J.A. (2018). Contributions of reader- and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 1687–1713. https://doi.org/10.1037/xlm0000547

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language, 65*(1), 42–73. https://doi.org/10.1016/j.jml.2011.03.002

Kuperman, V., & 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), 802–823. https://doi.org/10.1037/a0030859

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods, 44*, 325-343. https://doi.org/10.3758/s13428-011-0146-0

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences. 4*(1):6-14. https://doi.org/10.1016/s1364-6613(99)01418-7.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology. 6:*1171. https://doi.org/10.3389/fpsyg.2015.01171

Long, D. L., & Freed, E. M. (2021). An individual differences examination of the relation between reading processes and comprehension. *Scientific Studies of Reading, 25*:2, 104-122, https://doi.org/10.1080/10888438.2020.1748633

McCarron, S.P., & Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavioural Research Methods, 53*, 2226–2237. https://doi.org/10.3758/s13428-021-01556-y

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception. An account of basic findings. *Psychological Review, 88,* 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Mézière, D.C., Yu, L., Reichle, E.D., von der Malsburg, T. and McArthur, G. (2023), Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly.* https://doi.org/10.1002/rrq.498

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*, 267–296. https://doi.org/10.1037/a0021890.

Monsell, S. (1991). The nature and locus of word frequency effects inreading. In D. Besner & G. W. Humphreys (Eds.), Basic Process in reading: Visual word recognition (pp. 148–197). Hillsdale, NJ: Erlbaum.

Moore, M., & Gordon, P.C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods. 47*(4), 1095-1109. https://doi.org/10.3758/s13428-014-0534-3.

Morton, J. (1970). A functional model of human memory. In D. A. Norman (Ed.), Models of human memory (pp. 203–254). New York: Academic Press

O'Regan J. K. (1990). Eye movements and reading. In E. Kowler (Ed.), Eye movements and their role in visual and cognitive processes (pp. 395-453). Amsterdam: Elsevier, North-Holland.

O'Regan, J. K. (1992). Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 333-354). New York: Springer-Verlag https://doi.org/10.3758/BF03213829

Papadopoulos, T.C., Spanoudis, G.C., & Georgiou, G.K. (2016). How is RAN related to reading fluency? A comprehensive examination of the prominent theoretical accounts. *Frontiers in Psychology. 7*:1217. https://doi.org/10.3389/fpsyg.2016.01217

Perfetti, C. A. (1985). *Reading ability*. New York, NY, US: Oxford University Press.

Perfetti, C. A. (1992) The representation problem in reading acquisition. In Gough PB, Ehri LC and Treiman R (eds), *Reading acquisition*. Hillsdale, NJ: Erlbaum, pp. 145–174.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of Functional Literacy, 11*, 67-86. https://doi.org/10.1075/swll.11.14per

Pollatsek, A., Rayner, K., & Balota, D. A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception and Psychophysics, 40*, 123–130. https://doi.org/10.3758/BF03208192

Powell, D., Stainthorp, R., Stuart, M., Garwood, H., & Quinlan, P. (2007). An experimental comparison between rival theories of rapid automatized naming performance and its relationship to reading. *Journal of Experimental Child Psychology, 98*(1), 46-68. https://doi.org/10.1016/j.jecp.2007.04.003

R Core Team (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/.

Radach, R.R., Huestegge, L., & Reilly, R.G. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research, 72*, 675-688. https://doi.org/10.1007/s00426-008-0173-3

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., Ashby, J., Pollatsek, A., & Reichte, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 720–732. https://doi.org/10.1037/0096-1523.30.4.720

Rayner, K., & Fischer, M. H. (1996). Mindless reading revisited: Eye movements during reading and scanning are different. *Perception and Psychophysics, 58*(5), 734–747. https://doi.org/10.3758/BF03213106

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21*, 448– 465. https://doi.org/10.1037/0882-7974.21.3.448

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. Journal of Experimental Psychology: *Human Perception and Performance, 22*(5), 1188–1200. https://doi.org/10.1037/0096-1523.22.5.1188

Rayner, K., Schotter, E.R., Masson, M.E.J., Potter, M.C. & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest, Supplement. 17*(1), 4 - 34. https://doi.org/10.1177/1529100615623267

Ready, R. E, Chaudhry, M. F., Schatz, K.C., & Strazzullo, S. (2013). "Passageless" administration of the nelson-denny reading comprehension test: Associations with IQ and reading skills. *Journal of Learning Disabilities 46*(4):377-84. http://dx.doi.org/10.1177/0022219412468160.

Reichle, E. D. (2020). *Computational models of reading: A handbook.* Oxford University Press.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*(1), 125-157. http://dx.doi.org/10.1037/0033-295X.105.1.125

Reichle, E.D., Rayner, K., & Pollatsek, A. (1999) Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research, 39*(26), 4403-4411.

Savage, R., Pillay, V., & Melidona, S. (2007). Deconstructing rapid automatized naming: Component processes and the prediction of reading difficulties. *Learning and Individual Differences, 17*(2), 129-146. https://doi.org/10.1016/j.lindif.2007.04.001

Slattery, T. J., & Yates, M. (2017). Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology, 71*(1), 250–259. https://doi.org/10.1080/17470218.2017.1310264

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review, 125*(6), 969- 984. http://dx.doi.org/10.1037/rev0000119

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading and Research Quarterly, 24,* 402–429. https://doi.org/10.2307/747605

Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language, 116,* 104190. https://doi.org/10.1016/j.jml.2020.104190

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997) Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second-to fifth-grade children, *Scientific Studies of Reading, 1*(2), 161-185, https://doi.org/10.1207/s1532799xssr0102_4

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. Teacher's College, Columbia University, Science Press, New York. https://doi.org/10.1037/13283-000

Traxler, M.J., Long, D.L., Tooley, K.M., Johns, C.L., Zirnstein, M., & Jonathan, E. (2012). Individual Differences in Eye-Movements During Reading: Working Memory and Speed-of-Processing Effects. *Journal of Eye Movement Research, 5*(1), 5. https://doi.org/10.16910/jemr.5.1.5

Underwood, G., Hubbard, A., & Wilkinson, H. (1990). Eye fixations predict reading comprehension: The relationship between reading skill, reading speed, and visual inspection. *Language and Speech, 33,* 69–81. https://doi.org/10.1177/002383099003300105

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition, 131*(3), 373-403. https://doi.org/10.1016/j.cognition.2014.01.007

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67,* 1176-1190.

Veldre, A., & Andrews, S. (2014). Lexical quality and eye movements: Individual differences in the perceptual span of skilled adult readers. *Quarterly Journal of Experimental Psychology, 67*(4), 703–727. https://doi.org/10.1080/17470218.2013.826258

Veldre, A., & Andrews, S. (2015a). Parafoveal lexical activation depends on skilled reading proficiency. *Journal of Experimental Psychology: Learning Memory and Cognition, 41*(2), 586–595. https://doi.org/10.1037/xlm0000039

Veldre, A., & Andrews, S. (2015b). Parafoveal preview benefit is modulated by the precision of skilled readers' lexical representations. *Journal of Experimental Psychology: Human Perception and Performance, 41*(1), 219–232. https://doi.org/10.1037/xhp0000017

Veldre, A., Drieghe, D., & Andrews, S. (2017). Spelling ability selectively predicts the magnitude of disruption in unspaced text reading. *Journal of Experimental Psychology: Human Perception and Performance, 43*(9), 1612–1628. https://doi.org/10.1037/xhp0000425

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a grouping variable longitudinal study. *Developmental Psychology, 30*(1), 73–87. https://doi.org/10.1037/0012-1649.30.1.73

Waters G.S., & Caplan D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103*(4), 761–772. https://doi.org/10.1037/0033-295X.103.4.761

Wechsler, D. (2005). *Wechsler Individual Achievement Test 2nd Edition (WIAT-II II).* London: The Psychological Corp.

West, R. F., Stanovich, K. E., & Mitchell, H. R. (1993). Reading in the real world and its correlates. *Reading Research Quarterly,* 35-50. https://doi.org/10.2307/747815

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. Journal of Experimental Psychology: *Human Perception and Performance, 34*(1), 205. https://doi.org/10.1037/0096-1523.34.1.205

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*(3), 415. https://doi.org/10.1037/0022-0663.91.3.415

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., ... & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science, 21*(4), 551-559. https://doi.org/10.1177/0956797610363406

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 920–933. https://doi.org/10.1037/0278-7393.20.4.920