

Measuring Lexical Quality: The Role of Spelling Ability

Sally Andrews, Aaron Veldre, & Indako E. Clarke

The University of Sydney

Short title: Measuring lexical quality

Contact details

Sally Andrews

School of Psychology, The University of Sydney

Sydney, NSW 2006

Australia

sally.andrews@sydney.edu.au

+61 2 9351 8297

Abstract

The construct of 'lexical quality' (Perfetti, 2007) is widely invoked in literature on word recognition and reading to refer to a systematic dimension of individual differences that predicts performance in a range of word identification and reading tasks in both developing readers and skilled adult populations. Many different approaches have been used to assess lexical quality, but few have captured the orthographic precision that is central to the construct. This paper describes, evaluates and disseminates spelling dictation and spelling recognition tests that were developed to provide sensitive measures of the precision component of lexical quality in skilled college student readers – the population that has provided most of the benchmark data for models of word recognition and reading. Analyses are reported for 785 students who completed the spelling tests in conjunction with standardized measures of reading comprehension, vocabulary and reading speed, of whom 107 also completed author recognition and phonemic decoding tests. Internal consistency analyses showed that both spelling tests were relatively unidimensional and displayed good internal consistency, although the recognition test contained too many easy items. Item-level analyses are included to provide the basis for further refinement of these instruments. The spelling tests were moderately correlated with the other measures of written language proficiency but factor analyses revealed that they consistently defined a separate component, demonstrating that they tap a dimension of variability that is partially independent of variance in reading comprehension, speed and vocabulary. These components appear to align with the precision and coherence dimensions of lexical quality.

Keywords

Reading ability, spelling ability, individual differences, lexical quality hypothesis

Word count: 8798 (excluding Abstract, References, Footnotes, Tables, Figures, and Appendices)

This paper has both practical and theoretical goals. The major practical goal is to validate and disseminate two measures of spelling ability that were developed to discriminate among samples of skilled, native English-speaking university students. Although other measures of spelling ability are available in batteries such as the Wide Range Achievement Tests (Wilkinson & Robertson, 2017), they were not specifically developed for university student populations. This may, in part, account for why spelling ability has received little attention in the literature on individual differences among skilled readers relative to measures of word and nonword reading (e.g., Kuperman & Van Dyke, 2011), reading speed (e.g., Jackson & McClelland, 1975), reading comprehension (e.g., Ashby, Rayner, & Clifton, 2005), and vocabulary (e.g., Yap, Balota, Sibley, & Ratcliff, 2012). The lack of attention to spelling ability also reflects a common assumption that, at least for skilled readers, measures of word identification and spelling tap the same dimension of individual variability, typically conceptualized as *word identification* or *decoding*. While spelling is seen as a useful index of reading development (e.g., Treiman, 2017), at higher levels of skill, reading and spelling words are often assumed to be 'two sides of the same coin' (Ehri, 2000), suggesting that spelling is unlikely to account for unique variance. More generally, the contribution of individual differences in word identification to explaining variability in reading comprehension is typically assumed to reduce across reading development relative to measures of comprehension-related factors like vocabulary and listening comprehension (e.g., Braze et al., 2016).

Challenging these assumptions, a series of studies using the tests of *spelling dictation* and *spelling recognition* evaluated in this paper has demonstrated that spelling ability accounts for unique variance in a variety of measures of performance across a range of single-word and sentence-processing tasks within samples of monolingual, English-speaking university students (see Andrews, 2008, 2012, 2015; Andrews & Veldre, 2019, for reviews). Most of this research has been conducted in Andrews' laboratory at the University of Sydney, Australia, where the tests were

developed, but they are now beginning to be used by independent psycholinguistic research groups (Adelman et al., 2014; Emmorey, Midgley, Kohen, Sehyr & Holcomb, 2017; Eskenazi, Swischuk, Folk, & Abraham, 2018; Meade, Grainger, Midgley, Emmorey & Holcomb, 2018; Slattery & Yates, 2018; Tan & Yap, 2016). It is therefore timely to report analyses evaluating the tests' internal consistency and validity and to make them publicly available along with details of the typical administration procedures and norms so that they can be consistently applied by other researchers to investigate individual differences among skilled readers.

The Lexical Quality Hypothesis

The theoretical goal is to evaluate the validity and utility of combining measures of spelling ability with other measures of written language proficiency to assess individual differences in *lexical quality* among skilled readers. Perfetti (1985) coined the term 'lexical quality' to refer to a critical determinant of the efficiency and effectiveness of the procedures involved in retrieving linguistic codes during reading comprehension. He subsequently refined the definition to refer to qualities of skilled readers' lexical knowledge (Perfetti & Hart, 2001): High-quality representations are "orthographically fully specified", represent redundant word-specific, context-sensitive phonology, and are "semantically more generalized and less context-bound" (Perfetti, 2007, p. 359). This focus on the causal role of lexical *knowledge* distinguishes the lexical quality hypothesis (LQH) from many other accounts of individual differences in reading. Rather than attributing reading difficulties to deficits in phonological, semantic, or working memory *processes*, "the LQH is about knowledge that has not been acquired or practiced to a high-enough level" (Perfetti, 2007, p. 380) to achieve the properties required for efficient, effective retrieval and subsequent higher-level processing.

The utility of the construct of lexical quality depends on how clearly it is defined. Perfetti (2007) emphasized the *precision* and *flexibility* of lexical knowledge. Precision is identified with the *content* of lexical knowledge: the specificity and completeness of the orthographic, phonological,

grammatical, and semantic constituents of the lexical representation. *Orthographic precision* is particularly critical because written forms become the gateway to lexical knowledge (Dehaene, 2009). Flexibility arises from the interconnectedness or binding between the different constituents of lexical representations. The precision and redundancy of high-quality representations strengthens the binding between the orthographic, phonological, and semantic constituents that define a word's identity. These strong connections allow printed word forms to trigger synchronous, coherent activation of all components of the word's identity required for the higher order meaning integration processes underpinning effective comprehension (Perfetti, 2007). This property of greater *coherence* between the constituents of a lexical complex may be at least partially independent of the *precision* of the orthographic representation of a word (Andrews, 2015).

According to Perfetti's definition, lexical quality is a graded, word-specific attribute. The quality of lexical representations varies within individuals, as they gradually increase the size of their vocabulary and refine the specificity, redundancy, and inter-connectedness of the knowledge stored for existing words through reading experience. There are also differences between individuals in the extent to which they have established high-quality representations for most of the words in their written vocabulary. Such differences could arise from genetic differences in foundational skills such as phonological awareness or orthographic learning (Byrne et al., 2008) or from environmental factors such as reading experience and methods of instruction. They may also interact with differences in reading strategy: Readers who rely heavily on context to identify words may devote little attention to the details of words' internal structure and therefore be less likely to develop fully specified orthographic codes for all words (Frith, 1986).

Assessing lexical quality

Perfetti's (2007) construct of lexical quality is increasingly widely invoked in the literature on both developing and skilled readers to refer to a systematic dimension of individual differences that predicts performance in a variety of word identification and reading tasks (e.g., Breadmore & Deacon, 2019; Rossi, Martin-Chang, & Oulette, 2019; Slattery & Yates, 2018) but a range of different measures have been used to assess it. The most widely used indices of lexical quality are measures of vocabulary and decoding skill. Although word-level measures of decoding continue to predict significant variance in reading comprehension, even in skilled readers (e.g., Landi & Perfetti, 2007), vocabulary has been found to account for more variance than decoding at later stages of reading development (e.g., Protopapas, Sideris, Mouzaki, & Simos, 2007). In Verhoeven's systematic studies of the contribution of lexical quality to the development of reading comprehension in late primary age Dutch children, tests of both the breadth and depth of vocabulary accounted for a substantial proportion of variance in reading comprehension in a large cross-sectional sample, after controlling for decoding measures of word and nonword reading, short-term memory, and non-verbal intelligence (Swart et al., 2017). Vocabulary and category fluency also predicted growth in reading comprehension in a longitudinal study over grades 4 to 6 (Nouwens, Groen, Kleemans, & Verhoeven, 2017). Measures of 'lexical richness' obtained from tests of synonym knowledge and verbal analogies have also been found to predict eye-movement measures of lexical processing and semantic integration in English-speaking adolescents' sentence reading (Luke, Henderson, & Ferreira, 2015).

Individual differences in vocabulary, decoding, and meaning retrieval have also been used to assess lexical quality in adult readers. Many of the investigations conducted by Perfetti's group have used scores derived from factor analyses of a broad battery of tests of reading-related skills administered to large samples of college-student readers to define "functionally distinct dimensions of variability" (Taylor & Perfetti, 2016). Early studies (e.g., Perfetti & Hart, 2002; Landi & Perfetti,

2007) identified separate 'meaning knowledge' (assessed by vocabulary and reading comprehension) and 'form knowledge' factors (defined by spelling, phonology, and decoding) and suggested that form knowledge split into separate orthographic and phonological factors among less-skilled readers (Perfetti & Hart, 2002), supporting the LQH's assumption that lexical constituents become more integrated at higher levels of reading skill. Taylor and Perfetti's (2016) more recent research using an expanded battery including subjective reports of reading history (Lefly & Pennington, 2000) and a measure of print exposure based on author recognition (Stanovich & West, 1989), found that the strongest source of shared variance was a 'reading experience' factor defined by measures of reading speed, print exposure, book reading, and reading attitude, which was independent of a robust 'lexical knowledge' factor defined by decoding, word recognition, and spelling skill. Both factors predicted more efficient eye movements – more word skipping and fewer regressions – during reading of short passages, but the lexical factor was a stronger predictor of these reading behaviors for texts containing unfamiliar, recently acquired words. Convergent evidence that word identification predicts individual differences in eye movements in a broader community sample of adults derives from Kuperman and Van Dyke's (2011) finding that word identification and rapid naming were the only measures from a large battery of tests assessing decoding, working memory, and listening and reading comprehension that significantly predicted unique variance in eye movements during sentence reading. Better performance on these measures was associated with efficient oculomotor control that Kuperman and Van Dyke suggested to be "an affordance of the overall quality of ... [the reader's] lexical representations" (p. 56).

The test-battery approach adopted in the studies described above is very resource-intensive, particularly for the large samples that are desirable for individual-differences research. Many researchers have therefore attempted to capture individual differences in lexical proficiency with a single or limited set of measures. Vocabulary has been systematically used as a coarse index

of 'lexical integrity' in Yap and colleagues' thorough investigations of how individual differences modulate both behavioral performance and the underlying processes revealed by mathematical modeling methods across a range of word identification tasks (e.g., Pexman & Yap, 2018; Yap, Balota, Sibley, & Ratcliff, 2012; Yap, Tse, & Balota, 2009). Another relatively widely used single measure that aims to assess the reading experience assumed to underpin lexical quality is the Author Recognition Test (ART). Originally developed by Stanovich and West (1989), and subsequently refined and extended by Acheson, Wells, and MacDonald (2008), these quick, easily administered tests use the ability to discriminate between real and fabricated authors (or magazine titles; Acheson et al., 2008) as a surrogate measure of the extent of exposure to print (see Moore & Gordon, 2015, for a review).

The accumulating evidence that a range of measures of lexical proficiency predict systematic variance in skilled readers' performance not only in single-word identification and priming tasks but also in eye-movement measures of sentence reading confirms that individual differences in word-level processes remain significant predictors of variance among adult readers. However, to provide evidence for a *specific* contribution of lexical quality requires specification of how the predictor variables map to the precision, redundancy, and coherence of lexical representations that define their quality. As highlighted by Braze, Tabor, Shankweiler, and Mencl (2007), vocabulary knowledge is central to general linguistic comprehension processes that are shared between spoken and written language processing. Correlations between vocabulary and reading performance may, therefore, tap general comprehension processes rather than the quality of reading-specific lexical representations. Consistent with this interpretation, factor analyses typically show that vocabulary loads on the same latent factor as listening comprehension (Braze et al., 2016). However, in two independent studies of large community samples of adult readers, vocabulary also accounted for a small, but significant, component of unique variance in reading comprehension when word

decoding and listening comprehension were controlled for (Braze et al., 2007; Braze et al., 2016), suggesting a reading-specific contribution. Braze et al. (2016) suggested this may be because higher-quality representations "that incorporate subtle gradations of meaning, may integrate more flexibly into representations of discourse of narrative and ...be more readily recognized in context" (p. 447). Nevertheless, in studies that only measure vocabulary, it is not clear whether observed effects reflect lexical quality specifically, or factors related to general comprehension – or some combination of the two. Similarly, ART measures of print exposure are typically at least moderately correlated with measures of vocabulary and word identification (e.g., Moore & Gordon, 2015). Moreover, as well as contributing to the refinement of lexical representations, the reading experience that these tests are presumed to assess will influence a range of reading processes (Falkauskas & Kuperman, 2015).

Orthographic precision and lexical quality

Precision is central to Perfetti's (2007) definition of lexical quality, but little attention has been paid to establishing the extent to which the measures of individual differences used in studies of skilled reading capture this attribute. *Orthographic precision* is particularly critical to lexical quality: successful word identification requires readers to extract the relevant features from the perceptual input and map them to existing lexical representations. Phonology shapes the orthographic units that need to be extracted, but neuroimaging evidence suggests that skilled readers develop a specialized visual system for mapping visual input to word-specific knowledge (Dehaene et al., 2010). Spelling ability provides a direct index of the orthographic precision of readers' lexical representations of known words that the LQH suggests may play a specific role in predicting effective reading, independently of other measures of lexical proficiency.

Consistent with this view, recent studies of skilled readers that have indexed lexical quality by combining the spelling dictation and recognition tests described in this paper with standardized

measures of vocabulary and reading comprehension have demonstrated that spelling ability predicts unique variance in both masked priming studies of single word identification tasks (Andrews & Hersch, 2010; Andrews & Lo, 2012, 2013; Andrews, Lo, & Xia, 2017) and in behavioral (Andrews & Bond, 2008; Hersch & Andrews, 2012) and eye-movement indices of sentence reading (Drieghe, Veldre, Fitzsimmons, Ashby, & Andrews, 2019; Veldre & Andrews, 2014, 2015a, 2015b, 2016; Veldre, Drieghe, & Andrews, 2017). Independent studies using one or both of the same spelling tests have provided converging evidence for the unique contribution of spelling ability to predicting adult readers' lexical decision performance (Adelman et al., 2014), eye movements during sentence reading (Slattery & Yates, 2018), and electrophysiological indices of word processing (Meade et al., 2018). A second electrophysiological study found that the selective effect of spelling was more marked in pre-lingually deaf adults than hearing readers (Emmorey et al., 2017), suggesting that orthographic precision may play a particularly important role in reading development when phonological processes are compromised.

This accumulated evidence confirms that tests of spelling ability account for individual differences among skilled adult readers that are not captured by other measures of written language proficiency. However, further evidence about the internal consistency and validity of these tests is required to confirm and elaborate the relationship between spelling and lexical quality. To address these issues, the present paper reports analyses of data collated from over 800 individuals tested across nine independent samples of between 46 and 110 University of Sydney students who completed individual-differences tests in conjunction with their participation in eye-tracking studies of sentence reading.

The present research

The effects of spelling ability summarized above were derived from tests of *spelling dictation* and *spelling recognition* developed to discriminate among samples of skilled, university

student readers that have been briefly described in a number of previous papers (e.g., Andrews, 2008, 2012, 2015), and in most detail by Andrews and Hersch (2010). Spelling-production tasks, like dictation, directly test the precision of orthographic knowledge, but some theories of spelling assume that spelling production relies on different information than that required for recognition of correct spellings. Readers may rely on partial orthographic information to identify words (e.g., Frith, 1980) and therefore be able to correctly recognize words for which they cannot produce a correct spelling. Performance on spelling recognition tasks is also influenced by the type of misspellings included. Katz and Frost (2001) found that participants were more likely to accept a repeated misspelling as being correct when it was phonologically plausible, even if it had been correctly rejected on its first presentation, and adopted laxer criteria for judging spelling acceptability when phonologically implausible spellings were included in the recognition list than when all misspellings were phonologically plausible. Such findings demonstrate the role of decision processes in spelling recognition tasks and suggest that participants can be induced to vary the relative weighting of orthographic and phonological information. More extreme differences between production and recognition tasks are suggested by cognitive neuropsychological evidence of dissociations between reading and spelling performance in brain-injured patients that has been interpreted as indicating separate input and output representations that can be independently accessed and damaged (e.g., Ellis, 1993). Such views predict that readers may be able to correctly recognize the spelling of words that they cannot accurately produce by relying on their more accurate input representation (Holmes & Babauta, 2005).

Thus, spelling-recognition tasks may tap factors not captured by spelling dictation. To comprehensively assess spelling ability it is therefore important to include both measures. As reported by Andrews and Hersch (2010), the spelling dictation and recognition tests both

demonstrate high test-retest reliability ($r = 0.90$ and 0.93 , respectively) but they have not previously been analyzed for internal consistency. This is the goal of the first set of analyses reported here.

The further goal of the present research is to extend the evidence for the convergent and divergent validity of the spelling tests by assessing their relationships with other measures of written language proficiency across a large sample of participants. Before tackling this question, we addressed an issue relevant to assessing reading comprehension and vocabulary, two of the other major indices of lexical quality. Specifically, we investigated the implications of varying the time limits allowed for the reading-comprehension and vocabulary subtests of the Nelson-Denny Reading Test (NDT; Brown, Fishco, & Hanna, 1993), a widely used standardized measure of adults' reading proficiency. This test has been developed for use with students from grade 9 of high school through to the 4th year of college/university. Our extensive experience administering the test to samples of university students has revealed that a substantial number of participants complete one or both subtests before the recommended time limit expires. Presumably because of the potential reduction in the tests' discriminative power, combined with pragmatic constraints on testing time, applications of these tests in Perfetti's studies of individual differences in lexical quality among skilled readers have typically reduced the time limits for each subtest to half the length required for standard administration (e.g., Perfetti & Hart, 2001; Taylor & Perfetti, 2016). The same reduced time limits have been used in approximately half of the studies collated for the present research while the remainder used the standard time limits. This provided an opportunity to evaluate the impact of administration time on the distribution of NDT scores and their relationship to scores on other tests.

Method

Participants

Data were collated for 813 students who participated in exchange for credit in introductory Psychology courses. The mean age of the participants was 19.59 years¹ (SD=3.65 years) and approximately 71.3% of the sample was female. The recruitment criteria specified that participants spoke English, and began learning to read and write English by no later than age 6. Most participants (87.1%) reported that English was the first language they learned to speak, and English was the first language that virtually all (96.6%) the sample learned to read and write.

Procedure

The individual-differences data were collected over a period of approximately 5.5 years (late 2011-early 2017) as part of a series of nine eye-movement experiments investigating the role of written language proficiency in sentence reading. Almost all participants completed the two spelling tests (798 participants). In addition, most participants also completed the vocabulary, reading-comprehension, and reading-rate subsections of the NDT (785 participants). The Author Recognition Test (ART; Moore & Gordon, 2015) and the phonemic-decoding subtest of the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1998) were completed by 107 participants. The individual-differences battery was administered either before or after the eye-movement experiment, individually or in small groups.

Spelling dictation test. This test consists of 20 words selected from a larger set administered to samples of Australian university students by Burt and Tate (2002) to cover a broad range of spelling accuracy. In their sample, the words were correctly spelled by between 35% and 92% of participants, and discriminated between above-average and below-average spellers. The

¹ Demographic data were not collected from all participants. The summary statistics reported here are based on between 466 and 609 participants, but are representative of all samples tested.

experimenter read aloud each word and included it in a short sentence to resolve any ambiguity. Most participants handwrote the word on a response sheet, but one sub-sample ($n=62$) completed a computerized version of the test (implemented in Qualtrics but administered in the laboratory) in which they typed the word into a response box. Handwritten responses were scored manually by the experimenter and typed responses were automatically scored by Qualtrics. Administration of the spelling dictation test typically took between 5 and 7 minutes.

Spelling recognition test. This test consists of 88 items, half correctly spelled and half incorrect. Incorrect spellings were constructed to be phonologically plausible to increase the difficulty of the test and encourage reliance on orthographic knowledge (Katz & Frost, 2001). Participants were given unlimited time to select all incorrectly spelled items. They viewed all the items together in four columns and either recorded their responses by circling items on paper, or by selecting items in a Qualtrics survey. Responses were manually scored by the experimenter or automatically scored in Qualtrics. Scores on the spelling recognition test are calculated out of a total possible score of 88, subtracting the number of correct spellings selected (i.e., false alarms) and the number of incorrectly spelled items that were not selected (i.e., misses). Participants typically took between 5 and 10 minutes to complete the spelling recognition test.

Nelson-Denny Test (NDT). All participants completed Form H of this instrument, which includes two subtests. In the 80-item *vocabulary* test participants are given a word and asked to select the best-matching word or phrase from five options. The separately timed *comprehension* subtest includes 38 items relating to 7 short passages on a range of topics. It also provides an assessment of *reading rate* by instructing participants to mark their progress through the first passage after one minute has elapsed. A total of 361 of the participants were administered the vocabulary and comprehension sections with standard timing procedures (i.e., 15 minutes for vocabulary and 20 minutes for comprehension); the remaining 345 participants were allowed only

half the usual time limit for each subtest. To allow direct comparison of scores under the two administration procedures, a further sample of 107 participants (referred to as the *Full+Half* sample) completed the full-timed version of the NDT but marked where they were up to at the half-way point for each section so that both full and half-timed scores could be computed. This sample also completed the two additional tests described below.

Author Recognition Test (ART). The Full+Half sample of 107 participants completed this test which required them to identify which of a list of 100 names they recognized to be authors. The test items were taken from Moore and Gordon (2015) who removed 15 poorly discriminating author names from the ART scale developed by Acheson et al. (2008), leaving 50 author names and 50 foils. All names were listed in alphabetical order by surname on a response sheet and participants circled the names they recognized as authors. Participants were instructed not to guess because they would be penalized for incorrect responses. Administration of the ART took between 3 and 5 minutes. Scores were computed by subtracting the number of false alarms to foil names from the number of correctly selected authors, the standard method of scoring the ART (Acheson et al., 2008). Moore and Gordon (2015) found that this scoring method yielded slightly higher correlations with measures of reading behavior than a measure of hit rate alone.

Phonemic decoding. The Full+Half subsample also completed the phonemic-decoding subtest of the *TOWRE* (Torgeson et al., 1999) which consists of a list of 63 nonwords ranging from 1 to 4 syllables. After completing a practice list of 8 items, participants were given 45 seconds to read aloud as many nonwords as they could. Responses were recorded by a microphone and the number of correctly pronounced items was checked offline by a research assistant. Pronunciations were deemed correct if they applied plausible grapheme-phoneme correspondences for the complete nonword string. Partial scores were not applied.

Results and Discussion

Descriptive Statistics

Table 1 presents the descriptive statistics for each measure – spelling dictation, spelling recognition, vocabulary, reading comprehension, reading rate, author recognition, and phonemic decoding efficiency – separately for the three sub-samples administered full-timed and half-timed versions of the NDT.

As expected for a selected sample of predominantly native English-speaking university students, the average level of reading performance was relatively high: the means for the NDT vocabulary and comprehension tests administered with standard timing corresponded to the 74th and 75th percentiles, respectively, of the norms for US students in the first year of a 4-year college program; while the average reading rate corresponded to approximately the 60th percentile of that cohort. The mean and standard deviation of the ART scores was very similar to those obtained by Moore and Gordon (2015) for the 50-item test ($M=13.75$; $SD = 6.81$), while the average phonemic decoding score was above the 90th percentile for 6th grade children – the oldest members of the Australian normative sample (Marinus, Kohnen, & McArthur, 2013). Despite the high average performance, the wide score ranges indicate that the samples showed considerable variability on all measures in the battery.

Unsurprisingly, the NDT vocabulary and comprehension scores were lower for the half-timed version, but vocabulary scores were relatively less affected by the reduced time limit than comprehension. The reasons that the two sub-tests differ in sensitivity to speed pressure and the implications for their discriminative power are explored below. The manipulation of administration time did not affect the NDT reading-rate measure, because it is based on only the first passage of the comprehension test. The similarity between the mean reading speeds of the three sub-samples provides reassurance that they are of relatively equivalent reading proficiency.

TABLE 1 ABOUT HERE

Internal consistency of spelling tests

To address our first aim of assessing the psychometric properties of the two novel tests of spelling ability, analyses of the internal consistency of each test were conducted.

Spelling dictation

Factor analysis of the 20 items from the spelling dictation test using maximum likelihood estimation to test for unidimensionality showed a clear break in the scree plot after the first factor (from an eigenvalue of 4.56 to 1.21), which accounted for 18.88% of the variance.² Factor loadings (λ) and extracted communalities (h^2), which represent the proportion of the item's variance accounted for by this factor, are shown in Table A1 with descriptive statistics for each item. All but three items (*warranty*, *asymmetry*, *diligent*) had factor loadings greater than .30, indicating adequate prediction of the item from the latent construct assessed by the test as a whole. Internal consistency indexed by Cronbach's alpha of .814 is in the range classified as 'good', and the mean inter-item correlation of .18 falls at the lower end of the range taken to indicate test homogeneity (Clark & Watson, 1995). Further evidence of homogeneity is provided by the point biserial correlations (r_{pb} ; see Table A1) which index whether the item discriminates between individuals in the same way as the total test. All items had positive point biserial correlation coefficients $>.30$ (65% above .40), indicating that the items consistently predicted the total test score.

To provide further insight into the psychometric properties of the test, Rasch analysis of the 20 items was conducted using RUMM2030, yielding the Rasch item-difficulty indices (in logits) shown in Table A1 for each item, and summarized in the person-item map in Figure 1. Rasch item-difficulty indices estimate the level of person ability at which an item has a 50% chance of being

² Confirming that a single-factor solution captured the majority of predictable item variance, a second factor analysis that extracted two factors revealed that they accounted for 19.04% and 2.42% of variance, respectively, and were highly correlated ($r=.678$).

correctly/incorrectly endorsed. Higher Rasch difficulty values indicate more difficult items (i.e., a greater level of ability required). The most difficult item (*conciliatory*), which yielded a Rasch index of 2.45, was correctly spelled by less than 20% of the sample while over 88% correctly spelled the easiest item (*euphoric*, Rasch index of -1.94). Rasch analysis also allows items and individuals to be measured on the common scale of Rasch logit units depicted in the person-item map (Figure 1) which plots the participants by ability against the test items by difficulty. In general, the distribution of item difficulty aligns with the distribution of person ability, however the range of the ability distribution is wider than that of the item difficulty distribution. This indicates that the test items may not effectively capture individual differences in spelling ability at the lowest and highest levels.³

FIGURE 1 ABOUT HERE

Spelling recognition

Four of the 88 items (*appreciate, distinguish, exhibition, annual*) of the spelling recognition test were identified as correctly spelled by all participants and so were excluded from item-level analyses. The remaining 84 items were factor analyzed using maximum-likelihood estimation to check for unidimensionality. Examination of the scree plot showed a clear break after the first factor, which accounted for 8.83% of the variance. Factor loadings (λ) and extracted communalities (h^2) are shown in Table A3 with descriptive statistics for each item. Overall internal consistency was

³ To evaluate whether these limitations in discrimination were due to categorical scoring of spellings as correct/incorrect, the dictation responses for a subset of 125 participants were rescored by calculating the Levenshtein distance between the correct spelling and the participant's response using the *vwr* package (Keuleers, 2013) in R (R Core Team, 2019). On average, participants' spellings were very similar to the correct spelling – the mean edit distance was less than one letter ($M=0.66$, $SD=0.46$). Levenshtein-based scores of spelling accuracy were also highly correlated with the categorical dictation scores by both subjects ($r=-0.78$) and items ($r=-0.95$), and with the Rasch difficulty index based on categorical scores ($r=-0.97$). The Levenshtein scores also showed similar, albeit slightly weaker, relationships as the categorical scores with variability in NDT vocabulary (-0.42 vs 0.51), comprehension (-0.36 vs 0.46) and reading rate (-0.29 vs 0.37) and spelling recognition (-0.58 vs 0.80). Thus, at least for this set of items, the greater precision offered by Levenshtein distance measures does not appear to improve discrimination among skilled readers, presumably because their spelling errors are generally limited to specific ambiguous phoneme-grapheme correspondences.

'good' (Cronbach's alpha = .859). However, the factor loadings for some items were marginally negative (e.g., *sufficient*, *elementary*, *inhibition*) and many items had low communalities (e.g., *attitude*, *consequence*, *parallel*) indicating inadequate prediction of the item from the latent construct of spelling recognition ability. The mean inter-item correlation was .06, indicating that the test is likely not homogeneous. The point biserial correlations (r_{pb}) for many items were also negative or near-zero (e.g. *senior*, *fulcrum*, *guitar*), indicating poor discrimination of spelling ability. These limitations arise, at least in part, because the test contains too many easy items. Although the percentage of participants correctly identifying each item as either a correct or incorrect spelling ranged from 26.6% to 100%, more than half of the items were correctly classified by over 90% of participants, limiting its sensitivity.

This problem is evident in the person-item map for the spelling recognition test presented in Figure 2. The Rasch item-difficulty indices (in logits) reveal considerable variability among items: *consequence* was the easiest, with a Rasch score of -3.55, and *vigilant*, at 3.71, was the most difficult. Participants were also relatively normally distributed on the Rasch difficulty scale, but the distributions of ability and item difficulty were not aligned. Many items fell below the lowest level of spelling ability and therefore made little contribution to discrimination between people. It is noteworthy that the majority of the items in the easier half of the difficulty range are correct spellings (e.g., 37 the 41 items with Rasch values below 0), suggesting a bias to classify items as correctly spelled— i.e., participants were much more likely to fail to identify an incorrect spelling than to falsely classify a correct spelling as incorrect.

Thus, the results of item-level analyses for the Spelling Recognition test clearly indicate that there is room for psychometric improvement. This might be achieved by removing items with low communalities, negative factor loadings, and/or negative and near-zero point biserial correlations, but, as elaborated in the Discussion, attention will need to be paid to the distribution of correct and

incorrect spelling and response requirements of the recognition task. The present set of items may also be more effective in discriminating between individuals in samples with lower levels of written-language proficiency than those included in our studies. Even though a large proportion of the items yielded very high average accuracy, the Rasch item distribution for the easy items is graded, and only four items were correctly classified by all participants. Importantly, even in our restricted sample, scores on the recognition test were highly correlated with spelling dictation, and the factor analyses reported below show that they tapped the same dimension of variability between people.

FIGURE 2 ABOUT HERE

Effects of administration time on vocabulary and comprehension performance

To investigate our second research question of how differences in administration time influence the estimates of vocabulary and reading comprehension obtained from the NDT, scores were collated separately for participants using the standard 'Full-time' procedures ($n=361$), those tested using 'Half-time' procedures ($n=345$), and the Full+Half NDT sample ($n=107$) for whom both Full-time and Half-time measures of vocabulary and reading comprehension were obtained. As illustrated in Figure 3, the distribution of scores obtained in both tests was highly negatively skewed under the standard Full-time conditions (vocabulary: Skew=-0.97; comprehension: Skew=-1.54), reducing discrimination at the upper end of the score distribution, but substantially more normal with the Half-time limit conditions (vocabulary: Skew=-0.11; comprehension: Skew=0.46). The Half-time procedure therefore increased discrimination among more-proficient readers. Similar differences in the distribution of Full-time and Half-time scores were evident in the Full+Half NDT sample. To determine whether the differential discrimination of the Half-time procedure influenced the relationship of vocabulary and comprehension to other measures, the analyses of convergent and divergent validity reported below were conducted separately on the Full-time and Half-time samples.

FIGURE 3 ABOUT HERE

Correlations

Table 2 shows the correlations among spelling and reading measures for the samples that completed the NDT under full- and half-timed administrations. All measures were moderately to strongly positively correlated, although the relationship of reading rate with the other measures tended to be weaker than those between spelling, vocabulary, and comprehension. The two spelling tests were very highly correlated in both the full-timed and half-timed NDT subsamples, and they showed almost identical strong correlations between vocabulary and comprehension, and similar moderate relationships between these measures and the two spelling tests. The most substantial difference between the samples was a significantly higher correlation between NDT comprehension and reading rate in the Half-time ($r=.51$) than the Full-time sample ($r=.30$), $z = 3.018$, $p = .003$, suggesting that the relationship between reading speed and comprehension is stronger under time constraints. However, in general, the simple correlations indicate that the half-timed version of the NDT captures similar dimensions of individual differences among skilled readers as the standard full-timed version.

TABLE 2 ABOUT HERE

Further support for this conclusion derives from the similar patterns of correlations observed for the full and half-timed measures in the Full+Half subsample (see Table 3). The data for this sample also showed that the additional ART and phonemic decoding scores were only weakly related to each other, but both were at least moderately correlated with all of the measures of spelling and reading proficiency.

TABLE 3 ABOUT HERE

Principal Component Analysis

To evaluate whether these positively correlated measures of linguistic processing can be reduced to a smaller number of independent dimensions, principal component analyses with promax rotation were conducted.⁴ Each analysis was conducted for both the full-timed and half-timed NDT subsamples, excluding the Full+Half subsample to determine whether administration time changed the dimensional structure. Two components were extracted in each analysis.

TABLE 4 ABOUT HERE

The first set of analyses was conducted on the spelling dictation, spelling recognition, vocabulary, and comprehension scores that have been used in our previous published investigations of individual differences in masked priming and sentence reading. As summarized in Table 4, for both full-time and half-time subsamples, the two moderately correlated components accounted for around 85% of variance, with the spelling tests forming Component 1, and NDT vocabulary and comprehension forming Component 2. A second set of analyses added reading rate (rightmost columns of Table 4) to evaluate whether it moderated the impact of administration time. Again, both full-time and half-time subsamples yielded two similar, moderately correlated components on which the two spelling tests formed Component 1. In the full-timed subsample, vocabulary, comprehension, and reading rate again formed Component 2. However, for the half-timed subsample, comprehension and reading rate loaded selectively on Component 2 but vocabulary loaded equally on both components. There was no evidence of a similar cross-component contribution of vocabulary when reading rate was not included as a predictor, suggesting that reading rate captures variance shared with vocabulary that particularly affects performance in half-timed conditions.

⁴ All principal component analyses were also conducted with no rotation. Component structure prior to rotation also yielded two components in which the first centroid had high loadings from all variables and the second differentiated the spelling tests from the other measures.

TABLE 5 ABOUT HERE

A third set of principal component analyses was conducted on the Full+Half subsample. Separate analyses were conducted using full-timed and half-timed measures of vocabulary and comprehension to evaluate the similarities and differences to the structure observed for samples administered the independent full- and half-timed measures. These analyses also included ART and phonemic decoding scores to confirm and refine understanding of the dimensions of individual differences. Two components were extracted using a promax rotation. Component loadings are shown in Table 5.

Paralleling the results for the independent samples, analyses including both the full- and half-timed scores yielded moderately correlated components that separated the spelling tests from the other measures: the two spelling tests fell on Component 2, and the NDT measures of vocabulary, comprehension, and reading rate on Component 1.⁵ The two additional individual-differences measures collected for this sample provided useful evidence of convergent and divergent validity because they loaded on different components. The ART index of print exposure loaded on Component 1 with the three NDT measures, while the phonemic decoding score loaded on Component 2 with the two spelling tests. As well as providing converging evidence for the two components of reading proficiency identified in the principal component analyses, the differential loadings of these two measures also provide useful additional information about the nature of the dimensions they assess. The relationship between phonemic decoding and spelling ability is consistent with the view that orthographic precision depends on the amalgamation of orthography and phonology (Ehri, 2015; Perfetti, 2007), and that this dimension is partially independent of the

⁵ The same two-factor structure was observed in analyses including the smaller subset of predictors available for the full and half-timed samples (see Appendix). NDT vocabulary loaded selectively on the same component as NDT comprehension both when NDT reading rate was, and was not, included as a predictor.

higher-order linguistic knowledge and skills captured by tests of vocabulary, comprehension, and reading speed, and the broad reading experience tapped by the ART.

GENERAL DISCUSSION

The central aims of the present research were to assess the internal consistency and validity of two recently developed tests of spelling ability that have been used in several investigations of individual differences among skilled readers and to evaluate whether they assessed a dimension of variability that is not effectively captured by other widely used tests of written language proficiency.

The results confirmed previous evidence that the tests of spelling dictation and spelling recognition were highly correlated across the entire sample of close to 800 skilled readers, and in the three independent sub-samples of 107-361 participants ($r=.75-.82$) defined by the different administration timing conditions of the NDT. There is therefore little evidence that dictation and recognition tests tap independent input and output representations. Rather, spelling appears to be a relatively unitary ability in skilled adult readers but, as in other domains, production tests like dictation are more difficult than recognition tests. However, the failure to demonstrate independent dimensions of spelling ability may reflect psychometric limitations of the test of spelling recognition discussed below.

Factor analyses showed that the items in each test had good internal consistency (Cronbach's $\alpha > 0.8$) and assessed a relatively homogeneous dimension of individual variance. Rasch analysis revealed that the item and ability distributions for the spelling dictation test were relatively well aligned, although the items failed to tap either the upper or lower extreme of the ability distribution. The spelling recognition test was more poorly calibrated. The distributions of ability scores for both participants and items was relatively normal, but the test contained too many easy items. The analyses also revealed a strong bias to withhold 'incorrect' responses –

yielding high accuracy for correct spellings at the expense of a high miss rate for incorrect spellings. This conservative strategy may reflect the response requirement of the present recognition task – i.e., to check incorrect spellings. The opposite instruction – to endorse correct spellings – may yield the reverse bias: participants may be cautious about endorsing spellings as correct and therefore show a high miss rate to avoid making false alarms to incorrect spellings. Such biases might be reduced by telling participants that 50% of the items are correctly spelled. However, a more effective strategy for developing a psychometrically sound spelling recognition test may be to use a forced-choice format in which two alternative spellings of a single item are simultaneously presented to avoid the inherent problems of yes/no procedures.

Both spelling tests could undoubtedly be psychometrically improved to yield better ability-item alignment by adding more difficult items, particularly to the recognition test. However, care would be required to ensure that increasing discrimination did not reduce the specificity of the tests. Difficulty in producing or recognizing a correctly spelled word may reflect low vocabulary or reading experience rather than a specific limitation in the precision of orthographic knowledge for known words. Consistent with this possibility, there were small, significant correlations between Rasch difficulty and log frequency estimates from both the SUBTLEX and HAL corpora for both correctly ($r = -.24$ and $-.30$, respectively) and incorrectly spelled words ($r = -.40$; -0.37). Adding further items that are sufficiently difficult to be discriminating may therefore risk confounding spelling ability with vocabulary. As discussed in more detail in the next section, the importance of distinguishing the effects of spelling and vocabulary depends on the research goals.

The relatively poor discriminative power of the spelling recognition test may also, in part, reflect the relatively elite population of participants our samples were drawn from: monolingual tertiary students from a metropolitan, research-intensive, high-entry university. The test may yield more effective discrimination between individuals in more diverse samples of adult readers.

Nevertheless, even within our restricted samples it demonstrated strong convergent validity with the spelling dictation test. The two measures were highly correlated, and they defined a separate principal component in the three independent samples of participants, regardless of differences in the administration time of the NDT vocabulary and comprehension tests. Thus, despite its poorer discrimination, specific variance in performance on the recognition test was clearly shared with the dictation test rather than the other measures of written-language proficiency.

The inclusion of the additional measures of ART and phonemic decoding in the Full+Half sample contributed to establishing the convergent and divergent validity of the two components. Both measures showed at least moderate simple correlations with the tests defining both factors, despite their low correlation with each other. Confirming that they tapped different dimensions of individual difference, they loaded on different components in analyses using both full and half-timed NDT measures. Their differential loadings are conceptually consistent with the lexical quality hypothesis. The broad index of reading experience tapped by the ART loaded with other measures of higher-level reading processes, while phonemic decoding ability selectively loaded with the two tests of spelling ability. This evidence that phonemic decoding is specifically related to spelling ability, is consistent with the view that knowledge of orthographic-phonological correspondences is a critical foundation for the development of precise orthographic knowledge (Ehri, 2015); Nation, 2017). The foundational role of the alphabetic principle and phonological decoding in learning to read English has been well established in developmental populations (e.g., Byrne, 1998; Share, 1995) and argued to provide “the means for orthographic learning – the gradual accumulation of orthographic knowledge, via reading experience” (Nation, 2017, p.3). The present evidence that phonemic decoding is specifically related to spelling ability even in skilled adult readers provides strong support for such claims.

The consistency of the component structure regardless of whether vocabulary and comprehension were assessed in time-pressured conditions also suggests that the separable dimensions of individual differences exert a stronger influence than the method variance associated with strategic adjustments of speed-accuracy in response to different timing constraints. As summarized in Figure 3, reducing the time limits for these two tests had a dramatic influence on the distribution of scores for both subtests of the NDT. Such differences would be expected to influence the discriminative capacity of the tests and compromise their sensitivity to different sources of variability. However, the overall proportion of variance accounted for, and the component structure, was very consistent across variations in administration time. The only observed change in component structure was due to the inclusion of NDT reading rate in the Half-time sample. In this analysis, NDT vocabulary showed similar moderate loadings on both principal components, rather than loading selectively with the higher-level processing measures, as it did in all other analyses. No such shift was evident in the analysis using half-timed scores for the Full+Half sample, which was tested under the same longer time limits as the independent Full-time sample but instructed participants to mark where they were up to at the half-time limits. This sample was therefore not subject to the same time pressure as the Half-time sample. The contribution of vocabulary to the 'precision' component therefore appears to be associated with strategies that are enhanced under conditions of speed pressure.

Thus, with respect to the practical goals of this research, the analyses demonstrate that the two tests of spelling ability provide relatively internally consistent, converging measures of a source of individual differences among skilled readers that is partially independent of the more typically used measures of vocabulary, reading comprehension, and speed. Despite some psychometric limitations, these instruments have already yielded new empirical insights, briefly elaborated in the next section, that demonstrate their utility in research on individual differences in reading.

One practical issue that may warrant further investigation concerns the value of assessing both spelling dictation and spelling recognition. The high correlation between the two tests provides little support for the view that spelling production and recognition tap different representations or processes (e.g., Ellis, 1993). Nevertheless, the use of two converging measures that involve different encoding and response requirements may contribute to more effectively isolating variance specifically associated with the spelling ability. If a single test was selected, spelling dictation would certainly be favored on psychometric grounds. Assessing production of the complete orthographic form also provides a more conceptually valid index of precise orthographic knowledge and yields errors that can potentially be analyzed to diagnose the source of spelling problems. However, the informal experience we have gained by making the tests available to other research groups is that the spelling recognition test is preferred, presumably because of both its ease of administration and greater acceptability to participants: spelling dictation tests often evoke performance anxiety even in skilled readers. It may be possible to develop a more psychometrically sound spelling recognition test by using a forced-choice rather than yes-no format to reduce the biases of recognition tasks, and selecting distractor items that systematically manipulate different dimensions of similarity to the target (e.g., orthographic, phonological, morphological). Such an approach may allow construction of a recognition test that achieves the same level of discrimination, and diagnostic capacity as spelling dictation. The item analyses reported here provide the foundation for such developments.

The role of spelling in assessing lexical quality

The consistent evidence that tests of spelling ability tap a partially independent dimension of variability in reading skill aligns with the precision component of Perfetti's (2007) construct of lexical quality. It also confirms Andrews' (2008, 2012, 2015) proposal that measures of spelling ability can be used to tap lexical precision and account for unique variance that is not captured by

the measures of reading comprehension and vocabulary more typically used to assess individual differences among skilled readers.

Lexical quality is defined by both the precision of orthographic representations and the coherent, synchronous activation of their associated phonological and semantic codes (Andrews, 2015; Perfetti, 2007). The second component consistently identified in the principal component analyses was defined by measures of higher-level knowledge and processes: the semantic knowledge indexed by vocabulary, text-level measures of reading comprehension and speed, and the ART index of reading experience. It therefore appears to tap the processes associated with efficient, synchronous retrieval and use of the complex of lexical codes associated with a specific word form – processes related to the ‘coherence’ dimension of lexical quality (Perfetti, 2007). This is consistent with evidence that individuals with higher scores on the coherence than the precision component show stronger semantic influences of masked primes in both lexical decision (Andrews & Lo, 2013) and semantic categorization tasks (Andrews et al., 2017). The present evidence that NDT vocabulary loaded equally on both the precision and coherence factors for the half-time sample only when reading rate was included as a predictor suggests that knowledge of word meanings combines with high orthographic precision to facilitate rapid, coherent retrieval of the complex of lexical codes that define a word, as well as contributing to higher-level word integration processes. It also implies that the contribution of the coherence between orthographic and semantic codes to early lexical retrieval may be more effectively tapped when vocabulary is assessed under speed pressure.

The multidimensional nature of lexical quality revealed by these analyses highlights the value of including multiple measures of lexical proficiency in studies designed to investigate individual differences among highly skilled readers. As confirmed by the present data, measures of written language proficiency are generally at least moderately inter-correlated so a single reliable

measure can potentially tap variability in the broad dimension of lexical proficiency. For example, Yap's systematic investigations of the role of vocabulary in predicting individual differences across a range of tasks including single-word lexical decision (Yap et al., 2012) and semantic-categorization tasks (Pexman & Yap, 2018) and studies of unmasked (Yap, Tse, & Balota, 2009) and masked priming (Tan & Yap, 2016) have been interpreted as demonstrating the role of the "integrity" of readers' lexical representations in facilitating automatic lexical retrieval. However, vocabulary is also correlated with measures of higher-order linguistic processing such as listening comprehension (Braze et al., 2016). Similarly, in the present analyses, vocabulary most consistently loaded with the text-based measures of reading comprehension and reading rate rather than with spelling ability. The consistency of the modulating effects of vocabulary across a range of word-identification tasks may reflect the fact that it contributes to both the precision and coherence components of lexical quality, at least when it is assessed under speeded testing conditions. Vocabulary tests may therefore be a useful index of the broad construct of lexical quality when researchers are limited to a single measure.

However, studies using a broader test battery that allows different facets of lexical quality to be separated provide additional insights into *how* lexical quality contributes to effective reading that are obscured by more global measures. Investigations of masked orthographic priming of lexical-decision performance have shown that spelling selectively predicts competition from similar words on both behavioral performance (Andrews & Hersch, 2010; Andrews & Lo, 2012) and the N400 component of the event-related potential waveform (Meade et al., 2018). An 'orthographic profile' of relatively higher spelling than vocabulary is also associated with form-based morphological decomposition indexed by equivalently strong masked priming for genuine e.g., (*hunter-HUNT*) and pseudomorphemically related (*corner-CORN*) prime-target pairs (Andrews & Lo, 2013). In contrast, the opposite 'semantic profile' of higher vocabulary than spelling ability predicts

strong masked priming for genuinely morphological pairs (Andrews & Lo, 2013) and stronger semantic congruence priming in a semantic categorization task (Andrews et al., 2017).

Isolating the different dimensions of lexical quality has also yielded new insights into the processes underlying skilled readers' eye movements during sentence reading. Reading comprehension is typically the best predictor of the overall speed and efficiency of sentence reading, consistently predicting shorter fixation durations and sentence/passage reading times, but spelling ability selectively predicts measures related to eye-movement control such as saccade length (Veldre & Andrews, 2014) and word skipping (Slattery & Yates, 2018; Veldre & Andrews, 2016b; Veldre, Drieghe, & Andrews, 2017). The combination of spelling and reading comprehension is also a stronger predictor of the extent and depth of parafoveal processing than either measure alone (Veldre & Andrews, 2015a, 2015b). However, the two measures have counteracting effects on *semantic* processing of parafoveal previews: higher reading comprehension was associated with stronger preview benefit from a semantically related or contextually plausible word, whereas better spellers showed reduced semantic/plausibility preview benefit relative to poorer spellers (Veldre & Andrews, 2016a, 2016b; see Andrews & Veldre, 2019, for a review). These data converge with the masked morphological and semantic priming results in suggesting that high spelling ability supports rapid retrieval of a word's orthographic form. In the masked orthographic priming paradigm, this yields competition from word primes that are orthographically similar to the target. This competition eliminates the benefits of sublexical overlap between the prime and target which is responsible for priming in poor spellers. In studies of parafoveal preview, the conflict between the orthographic form of the preview and the target eliminates the benefits of the preview's contextual acceptability – which is the source of semantic preview benefit in poorer spellers. In both cases, the detrimental effects of high orthographic precision are due to the conflicting perceptual input introduced by the masked prime, or parafoveal preview. In normal reading, where

such misleading information is not presented, rapid retrieval of either orthographic or semantic features of a briefly presented, or parafoveal, word will facilitate lexical retrieval and enhance the efficiency of reading.

This evidence of systematic individual differences among skilled readers reviewed above contributes to resolving some of the contradictory findings obtained in typical analyses of the averaged data for samples of skilled readers that have sustained ongoing debates about visual word recognition and reading (Andrews, 2012). The evidence that spelling ability selectively modulates lexical competition suggests that efforts to ‘crack *the* orthographic code’ (emphasis added) for reading (e.g., Grainger, 2008) may be misguided. Similarly, the differential effects of vocabulary and spelling on morphological and semantic priming suggest that there may not be a single answer to the debate between form-first versus cascaded activation accounts of semantic retrieval (e.g., Forster, 2013; Rodd, 2004). Individual differences also influence the interactions between foveal and parafoveal processing that provide a critical source of evidence for distinguishing between serial and parallel models of eye movement control (e.g., Engbert, Nuthman, Richter, & Kliegl, 2005; Reichle, Pollatsek, Fisher, & Rayner, 1998; Snell, van Leipsig, Grainger, & Meeter, 2018).

It is important to emphasize that the different components of lexical quality are not entirely independent. Across each of our samples, the correlation between the two rotated components identified in our factor analyses was consistently around 0.5. In unrotated factor solutions, all the NDT and spelling tests loaded on the first common factor, which accounted for 54% and 56% of variance in the full and half-time samples, respectively; while the discrepancies between spelling and NDT measures emerged on the second component, which captured an additional 19-20% of variance. Thus, the primary dimension of individual differences tapped by our test battery is common to spelling, vocabulary, and reading comprehension – an index we have referred to as ‘overall proficiency’ (e.g., Andrews, 2015; Andrews & Veldre, 2019). However, when this shared

variance is partialled out, a second, weaker dimension of variance is revealed that is defined by discrepancies between individual's level of spelling and vocabulary/comprehension ability.

The fact that spelling taps a common dimension of variance that is shared with vocabulary and comprehension is not surprising. Knowledge of the meanings and the spellings of words both depend critically on experience reading (and writing). Although spoken language experience clearly makes an important contribution to vocabulary acquisition, the majority of children's vocabulary learning occurs through reading (Nagy, Herman, & Anderson, 1985), and print exposure is clearly essential for spelling in both children and adults (e.g., Mol & Bus, 2011) – particularly in the quasi-regular writing system of English.⁶ Nation (2017) argues that repeated exposure to words across a diversity of contexts, episodes, and experiences supports development of a rich interconnected database and leads to “local variation at the word level: a ‘lexical legacy’ that is measurable during word reading behavior” (p. 1). Vocabulary and spelling knowledge both depend on the capacity to extract invariant patterns from text and establish connections both between the different components of a lexical form (orthography, phonology, and semantics), and between related linguistic units. Individual differences in these capacities, and their consequences for the quality of the lexical database that is acquired through reading experience, may account for the shared

⁶ An issue that is beyond the scope of this paper, but warrants further research, is whether the independent effects of spelling ability on reading behavior generalize to languages other than English. The construction of precise word-specific orthographic representations may be a specific response to the notoriously inconsistent and idiosyncratic spelling-sound correspondences of English which have been demonstrated to be more complex than any other alphabetic orthography (Share, 2008). This complexity may drive extraction of multi-letter units at a range of different ‘grain-sizes’ (Ziegler & Goswami, 2005) to capture systematic, higher-order consistencies in the mapping between orthography and phonology (e.g., Kessler & Treiman, 2001). If so, the independent contributions of spelling ability to predicting reading behavior may be limited to ‘orthographically deep’ scripts (Frost, Katz, & Bentin, 1987) but absent in transparent alphabetic script like Finnish and Spanish, for which word-specific orthographic representations may be unnecessary or redundant. Alternatively, the contribution of orthographic knowledge to reading behavior may reflect general cognitive principles of skill acquisition (Anderson, 1981). Across a range of domains of skilled human behavior, the transition from novice to expert performance is characterized by a shift from slow, deliberate, effortful, algorithmic processing to rapid, automatic performance mediated by direct-retrieval mechanisms (e.g., Logan, 1998). From this perspective, unitised orthographic representations may be a signature of expert word recognition and reading even in highly transparent scripts. However, spelling tests like those reported here are unlikely to provide a sensitive measure of orthographic knowledge in such languages so other methods will be required to investigate this possibility.

variance in spelling, vocabulary, and reading. However, the evidence of a second dimension of individual variability that is tapped by the discrepancy between orthographic and semantic knowledge implicates an additional source of variation in either the knowledge extracted from reading experience, or the manner in which it is applied to word identification and reading. Further systematic investigations of orthographic learning in developing readers (e.g., Nation, 2017; Joseph, Wonnacott, Forbes, & Nation, 2014; Tamura, Castles, & Nation, 2017) and computational modeling of this development (e.g., Ziegler, Perry, & Zorzi, 2014) will contribute to determining the individual, instructional, and contextual factors responsible for these differences.

Conclusions

The spelling dictation test and spelling recognition test described here each only take approximately 5 minutes to administer but capture unique variance among the population of skilled, college-aged readers that has been shown to modulate effects in masked priming and sentence reading tasks. Both spelling tests were unidimensional, displayed good internal consistency, correlated with other measures of reading ability, and formed a distinct component across a large college-aged sample. The item-level analyses reported in the present paper provide a basis for further refinement of these instruments, particularly the spelling recognition test, to better discriminate among participants at the highest levels of proficiency, but the present versions may discriminate more effectively in less-skilled adult samples. The results also demonstrate that using shorter time limits for standard tests of vocabulary and comprehension enhances discrimination among skilled readers without substantially changing the relationships between different components of written-language proficiency. These practical contributions to measuring lexical quality will support the development of a richer body of empirical evidence about how individual differences modulate skilled word recognition and reading. Incorporating such variation into models of these processes is widely acknowledged to be a critical step for future theoretical

development in understanding of these essential educational and vocational skills (e.g., Andrews & Reichle, 2019; Radach & Kennedy, 2013; Rayner, Abbott, & Plummer, 2015).

Author Note

The data, test instruments, and norms for the Spelling Dictation Test and the Spelling Recognition Test are available at <https://osf.io/t4x7r/>. This research was supported under Australian Research Council's *Discovery Projects* funding scheme (project numbers DP16203224, DP190100719).

References

- Acheson, D.J., Wells, J.B. & MacDonald, M.C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*, 278-289.
- Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., . . . Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*, 1052-1067.
- Anderson, J.R. (1981). *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.
- Andrews, S. (2008). Lexical expertise and reading skill. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 49, pp. 247-281). San Diego, CA: Elsevier.
- Andrews, S. (2012). Individual differences in skilled visual word recognition and reading: The role of lexical quality. In J. S. Adelman (Ed.), *Visual Word Recognition* (Vol. 2, pp. 151-172). London: 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. & Bond, R. (2009). Lexical expertise and reading skill: Bottom-up and top-down processing of lexical ambiguity. *Reading & Writing: An Interdisciplinary Journal*, *22*, 687-711.
- Andrews, S. & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, *139*, 299-318.
- Andrews, S. & Lo, S. (2012). Not all skilled readers have cracked the code: Individual differences in masked form priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 152-163.

- Andrews, S. & Lo, S. (2013). Is morphological priming stronger for transparent than opaque words? It depends on individual differences in spelling and vocabulary. *Journal of Memory and Language, 68*, 279-296.
- Andrews, S., & Reichle, E.D. (2019). The cognitive architecture of reading: The organization of an acquired skill. In P. Hagoort (Ed.), *Human Language: From genes and brains to behaviour*. MIT Press.
- Andrews, S., & Veldre, A. (2019). What is the most plausible account of the role of parafoveal processing in reading? *Language and Linguistics Compass*.
- Andrews, S., Lo, S., & Xia, V. (2017). Individual differences in automatic semantic priming. *Journal of Experimental Psychology: Human Perception and Performance, 43*, 1025-1039.
- 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, 58*, 1065-1086.
- Braze, D., Katz, L., Magnuson, J.S., Mencl, W.E., Tabor, W., Van Dyke, J.A. et al., (2016). Vocabulary does not complicate the simple view of reading. *Reading and Writing, 29*, 435-451.
- Braze, D., Tabor, W., Shankweiler, D., & Mencl, W.E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities, 40*, 226-243.
- Breadmore, H.L., & Deacon, S.H. (2019). Morphological processing before and during children's spelling. *Scientific Studies of Reading, 23*, 178-191.
- Brown, J. I., Fishco, V. V., & Hanna, G. (1993). *Nelson-Denny Reading Test*. Rolling Meadows, IL: Riverside Publishing Company.
- Burt, J. S., & Tate, H. (2002). Does a reading lexicon provide orthographic representations for spelling? *Journal of Memory and Language, 46*, 518-543.

- Byrne, B., Coventry, W. L., Olson, R. K., Hulstender, J., Wadsworth, S., DeFries, J. C., . . . Samuelsson, S. (2008). A behaviour-genetic analysis of orthographic learning, spelling and decoding. *Journal of Research in Reading, 31*, 8-21.
- Chace, K.H., Rayner, K., Well, A.D. (2005). Eye movements and phonological parafoveal preview: Effects of reading skill. *Canadian Journal of Experimental Psychology, 59*, 209-217.
- Choi, W., Lowder, M. W., Ferreira, F., & Henderson, J. M. (2015). Individual differences in the perceptual span during reading: Evidence from the moving window technique. *Attention, Perception, & Psychophysics, 77*, 2463-2475.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.
- Dehaene, S. (2009). *Reading in the brain*. New York: Penguin Group.
- Dehaene, S., Pegado, F., Braga, L., Ventura, P., Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, M. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 3*, 1359-1364.
- Drieghe, D., Veldre, A., Fitzsimmons, G., Ashby, J., & Andrews, S. (2019). The influence of number of syllables on word skipping during reading revisited. *Psychonomic Bulletin & Review, 26*, 616-621.
- Ehri, L. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders, 20*, 19-37.
- Ehri, L. (2015). How children learn to read words. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 293-310), Oxford, England: Oxford University Press
- Emmorey, K., Midgley, K.J., Kohen, C.B., Sevcikova Sehyr, Z. & Holcomb, P.J. (2017). The N170 ERP component differs in laterality, distribution and association with continuous reading measures for deaf and hearing readers. *Neuropsychologia, 106*, 298-309.

- Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777-813.
- Falkauskas, K. & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*, 1607-1627.
- Forster, K.I. (2013). How many words can we read at once? More intervenor effects in masked priming. *Journal of Memory and Language*, *69*, 563-573.
- Frith, U. (1980). Unexpected spelling problems. In U. Frith (Ed.), *Cognitive Processes in Spelling* (pp. 495-515). London: Academic Press.
- Frith, U. (1986). A developmental framework for developmental dyslexia. *Annals of Dyslexia*, *36*, 69-81.
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 104-115.
- Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Processes*, *23*, 1-35.
- Hersch, J. & Andrews (2012). Lexical expertise and reading skill: Bottom-up and top-down contributions to sentence processing. *Scientific Studies of Reading*, *16*, 240-262.
- Holmes, V. M., & Babauta, M. L. (2005). Single or dual representations for reading and spelling? *Reading and Writing*, *18*, 257-280.
- Jackson, M. D., & McClelland, J. L. (1975). Sensory and cognitive determinants of reading speed. *Journal of Verbal Learning and Verbal Behavior*, *14*, 565-574.
- Joesph, H.S., Wonnacot, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, *133*, 238-248.
- Katz, L. & Frost, S.J. (2001). Phonology constrains the internal orthographic representation. *Reading and Writing*, *14*, 297-332.
- Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of Memory and Language*, *44*, 592-617.

- Keuleers, E. (2013). vwr: Useful functions for visual word recognition research (R package version 0.3.0). <https://CRAN.R-project.org/package=vwr>
- 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*, 42-73.
- Landi, N., & Perfetti, C. A. (2007). An electrophysiological investigation of semantic and phonological processing in skilled and less-skilled comprehenders. *Brain and Language, 102*, 30-45.
- Lefly, D. L., & Pennington, B. F. (2000). Reliability and validity of the Adult Reading History Questionnaire. *Journal of Learning Disabilities, 33*, 286-296.
- Logan, G.D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.
- Luke, S. G., Henderson, J. M., & Ferreira, F. (2015). Children's eye-movements during reading reflect the quality of lexical representations: An individual differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1675-1683.
- Marinus, E., Kohnen, S., & McArthur, G. (2013). Australian comparison data for the Test of Word Reading Efficiency (TOWRE). *Australian Journal of Learning Difficulties, 18*, 199-212.
- Meade, G., Grainger, J., Midgley, K.J., Emmorey, K. & Holcomb, P.J. (2018). From sublexical facilitation to lexical competition: ERP effects of masked neighbor priming. *Brain Research, 1685*, 29-41.
- 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.
- Moore, M. & Gordon, P. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods, 47*, 1095-1109.
- Nagy, W.E., Herman P.A., & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly, 20*, 439-458.

- Nation, K. (2017). Nurturing a lexical legacy: reading experience is critical for the development of reading skill. *npj Science of Learning*, 2:3.
- Nouwens, S., Groen, M. A., Kleeman, T., & Verhoeven, L. (2017). The role of semantic retrieval in children's reading comprehension development in the upper primary grades. *Journal of Research in Reading*, 41, 597-614.
- Perfetti, C. A. (1985). *Reading Ability*. New York: Oxford University Press.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357-383.
- Perfetti, C. & Hart, L. (2001). The lexical basis of comprehension skill. In D. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67-86). Washington DC: American Psychological Association.
- Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189-214). Amsterdam, The Netherlands: John Benjamins Publishing.
- Pexman, P.M. & Yap, M.J. (2018). Individual differences in semantic processing: Insights from the Calgary Semantic Decision Project. *Journal of Experimental Psychology: Learning Memory and Cognition*, 44, 1091-1112.
- R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *Quarterly Journal of Experimental Psychology*, 66, 429-452.
- Rayner, K., Abbott, M. J., & Plummer, P. (2015). Individual differences in perceptual processing and eye movements in reading. In P. Afflerbach (Ed.), *Handbook of individual differences in reading: Reader, text, and context* (pp. 348-363). New York, NY: Routledge.

- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125-157.
- Rodd, J.M. (2004). When do leotards get their spots? Semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 434-439.
- Rossi, M., Martin-Chang, S., & Ouellette, G. (2019). Exploring the space between good and poor spelling: Orthographic quality and reading speed. *Scientific Studies of Reading*, *23*, 192-201.
- Share, D.L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*, 151-218.
- Share, D.L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an “outlier” orthography. *Psychological Bulletin*, *134*, 584-615.
- Slattery, T.J., & Yates, M. (2018). Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology*, *71*, 250-259.
- Snell, J., van Liesig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*, 969-984.
- Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*, 402-433.
- Swart, N. M., Muijselaar, M. M. L., Steenbeek-Planting, E. G., Droop, M., de Jong, P. F., & Verhoeven, L. (2017). Differential lexical predictors of reading comprehension in fourth graders. *Reading and Writing*, *30*, 489-507.
- Tan, L. C., & Yap, M. J. (2016). Are individual differences in masked repetition and semantic priming reliable? *Visual Cognition*, *24*, 182-200.
- Tamura, N., Castles, A., & Nation, K. (2017). Orthographic learning, fast and slow: Lexical competition effects reveal the time course of word learning in developing readers. *Cognition*, *163*, 93-102.

- Taylor, J. N., & Perfetti, C. A. (2016). Eye movements reveal readers' lexical quality and reading experience. *Reading and Writing, 29*, 1069-1103.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Test of word reading efficiency. Austin, TX: ProEd.
- Treiman, R. (2017). Learning to spell: Phonology and beyond. *Cognitive Neuropsychology, 34*, 83-93.
- 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*, 703-727.
- Veldre, A. & Andrews, S. (2015a). Parafoveal lexical activation depends on skilled reading proficiency. *Journal of Experimental Psychology: Learning Memory and Cognition, 41*, 586-595.
- Veldre, A. & Andrews, S. (2015b). Parafoveal lexical activation is modulated by the precision of skilled readers' lexical representations. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 219-232.
- Veldre, A., & Andrews, S. (2016a). Semantic preview benefit in English: Individual differences in the extraction and use of parafoveal semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 837-854.
- Veldre, A., & Andrews, S. (2016b). Is semantic preview benefit due to relatedness or plausibility? *Journal of Experimental Psychology: Human Perception and Performance, 42*, 939-952.
- 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*, 1612-1628.
- Wilkinson, G.S., & Robertson, G.J. (2017). *Wide Range Achievement Test* (5th edition). Pearson Education Inc.

- Yap, M.J., Balota, D.A., Sibley, D.E. & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 53-79.
- Yap, M. J., Tse, C. S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language*, 61, 303-325.
- Ziegler, J.C., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Philosophical Transactions of the Royal Society B*, 369, 20120397.

Table 1*Descriptive statistics for spelling and reading ability measures*

	Full time NDT (n = 361)				Half time NDT (n = 345)				Full+Half (n = 107)			
	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range	N
Age	19.90	4.78	18-56	130	19.67	3.35	17-38	336	18.95	2.83	17-34	107
Spelling Dictation	12.21	4.03	2-20	345	12.16	4.39	1-20	345	13.22	3.94	2-20	107
Spelling Recognition	74.65	7.44	48-88	346	75.19	7.27	53-88	345	76.70	7.64	51-87	107
NDT Vocabulary	64.09	9.83	25-80	344	53.63	13.47	19-80	334	64.97 ^a	9.72 ^a	40-78 ^a	107 ^a
									45.72 ^b	11.12 ^b	18-76 ^b	103 ^b
NDT Comprehension	63.81	8.54	20-76	344	32.57	12.64	8-68	334	62.56 ^a	9.82 ^a	34-74 ^a	107 ^a
									31.22 ^b	11.71 ^b	10-58 ^b	103 ^b
NDT Reading Rate	289.04	82.05	164-610	341	276.05	84.14	116-610	329	274.94	72.72	164-512	107
Author Recognition Test									12.59	6.46	1-35	107
Phonemic Decoding									52.09	9.20	16-63	106

Notes. NDT = Nelson-Denny Reading Test; The standard scoring procedure for the NDT doubles the comprehension score, to yield a maximum of 76, so that Vocabulary and Comprehension make relatively equal contributions to the Total score calculated by summing them.

^a Full time administration. ^b Half time administration

Table 2

Correlations (n in parentheses) among measures for full-timed (below diagonal) and half-timed (above diagonal) samples.

	1.	2.	3.	4.	5.
1. Spelling Dictation		.82** (345)	.52** (334)	.42** (334)	.32** (329)
2. Spelling Recognition	.75** (345)		.47** (334)	.37** (334)	.27** (329)
3. NDT Vocabulary	.41** (343)	.43** (344)		.62** (334)	.32** (329)
4. NDT Comprehension	.39** (343)	.44** (344)	.60** (344)		.51** (329)
5. NDT Reading Rate	.25** (341)	.24** (341)	.32** (341)	.30** (341)	

Note. NDT = Nelson-Denny Reading Test.

** $p < .01$ (two-tailed).

Table 3

Correlations (n in parentheses) among measures for the Full+Half sample

	1.	2.	3.	4.	5.	6.	7.	8.
1. Spelling Dictation								
2. Spelling Recognition	.78** (107)							
3. ND Vocabulary (full-timed)	.38** (107)	.38** (107)						
4. NDT Comprehension (full-timed)	.44** (107)	.41** (107)	.69** (107)					
5. NDT Vocabulary (half-timed)	.36** (103)	.33** (103)	.78** (103)	.59** (103)				
6. NDT Comprehension (half-timed)	.41** (103)	.32** (103)	.61** (103)	.77** (103)	.65** (103)			
7. NDT Reading Rate	.33** (107)	.28** (107)	.34** (107)	.43** (107)	.43** (103)	.58** (103)		
8. Author Recognition	.41** (107)	.37** (107)	.60** (107)	.46** (107)	.45** (103)	.46** (103)	.38** (107)	
9. Phonemic Decoding	.52** (106)	.46** (106)	.37** (106)	.40** (106)	.33** (103)	.35** (103)	.33** (106)	.18 (106)

Note. NDT = Nelson-Denny Reading Test.

** $p < .01$ (two-tailed).

Table 4

Component loadings for full-timed and half-timed Nelson-Denny subsamples for spelling dictation, spelling recognition, vocabulary, and comprehension. Analyses including Nelson-Denny reading rate are listed in the rightmost columns.

	Full-timed ND (n = 343)		Half-timed ND (n = 334)		Full-timed ND (n = 341)		Half-timed ND (n = 329)	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Spelling Dictation	.95	-.03	.92	.05	.94	-.05	.94	-.01
Spelling Recognition	.91	.04	.98	-.05	.93	-.02	.98	-.10
NDT Vocabulary	.02	.88	.14	.81	.22	.68	.45	.48
NDT Comprehension	-.01	.90	-.10	.97	.22	.67	.09	.83
NDT Reading Rate	-	-	-	-	-.25	.88	-.17	.91
Total variance explained	62.87%	21.09%	65.43%	20.92%	53.89%	18.33%	58.02%	19.36%
Inter-component correlation	.49		.50		.49		.49	

Note. NDT = Nelson-Denny Reading Test. Component loadings > |.40| shown in **bold**.

Table 5

Component loadings for Full+Half subsample for spelling dictation, spelling recognition, vocabulary, comprehension, reading rate, author recognition, and phonemic decoding efficiency

	Full-timed ND (<i>n</i> = 107)		Half-timed ND (<i>n</i> = 103)	
	PC1	PC2	PC1	PC2
Spelling Dictation	.03	.89	.03	.91
Spelling Recognition	-.01	.89	-.05	.93
NDT Vocabulary	.88	-.02	.80	.01
NDT Comprehension	.78	.11	.88	-.01
NDT Reading Rate	.59	.08	.80	-.06
Author Recognition	.85	-.09	.64	.11
Phonemic Decoding	-.00	.76	.05	.71
Total variance explained	51.28%	15.52%	50.07%	16.99%
Inter-component correlation	.52		.48	

Note. NDT = Nelson-Denny Reading Test. Component loadings > |.40| shown in **bold**.

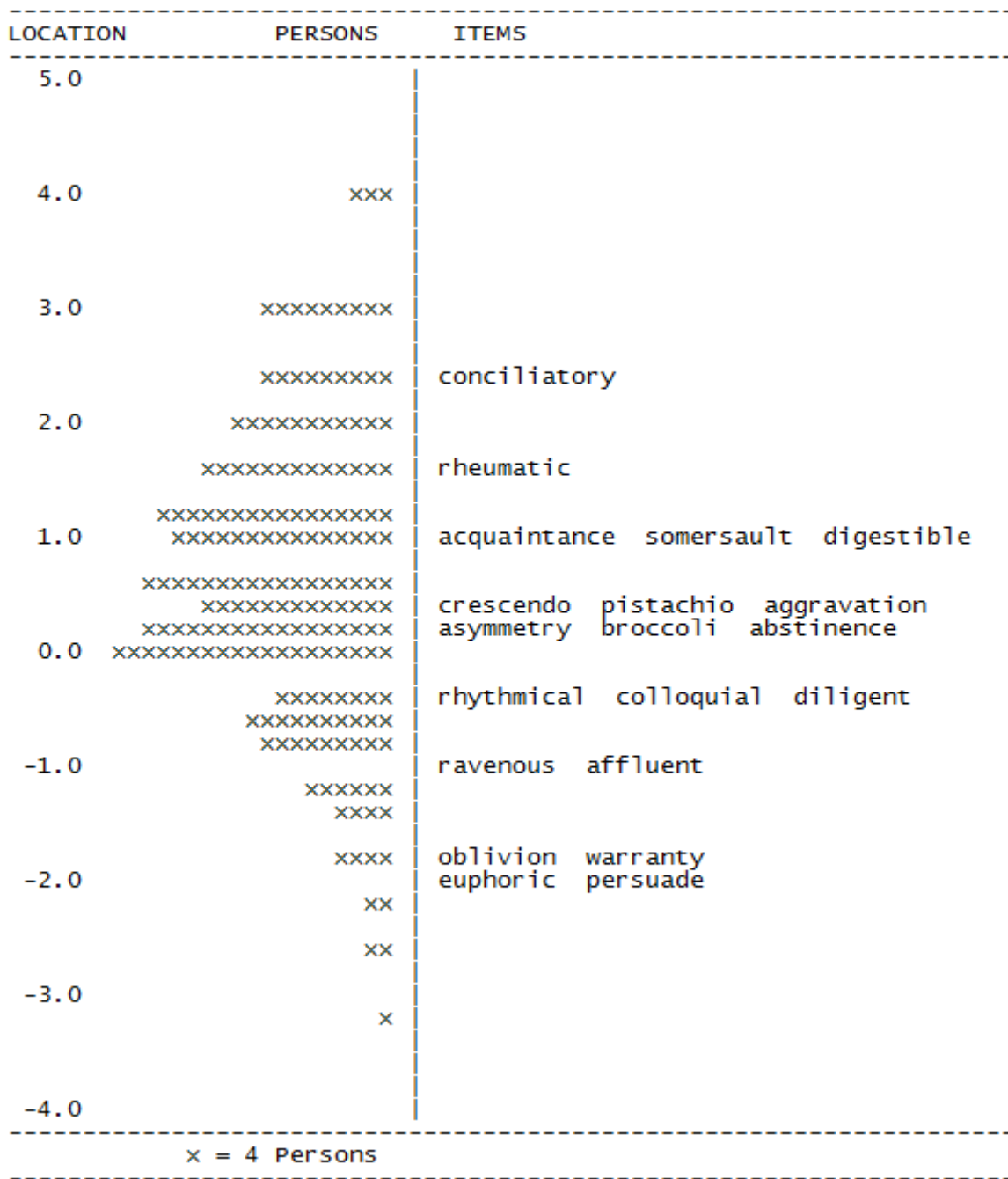


Figure 1. Plot of participants ('PERSONS') by performance on the Spelling Dictation Test against the 20 test items ('ITEMS') by Rasch difficulty (in logit units). Better test performance and more difficult items appear at the top of the figure.

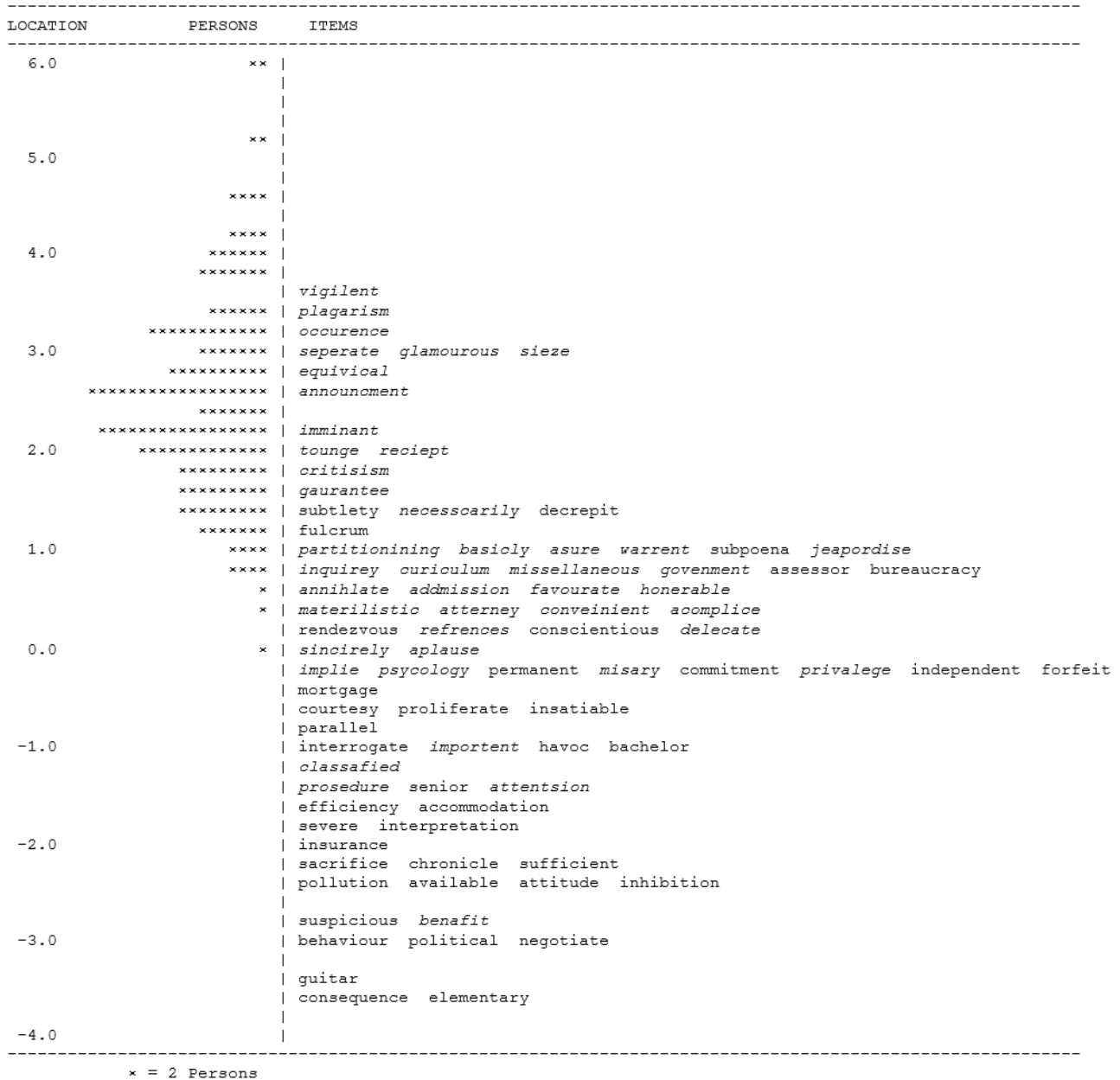


Figure 2. Plot of participants ('PERSONS') by test performance on the Spelling Recognition Test against 84 test items ('ITEMS') by Rasch difficulty (in logit units). Better test performance and more difficult items appear at the top of the figure. Items in italicized font are incorrect spellings.

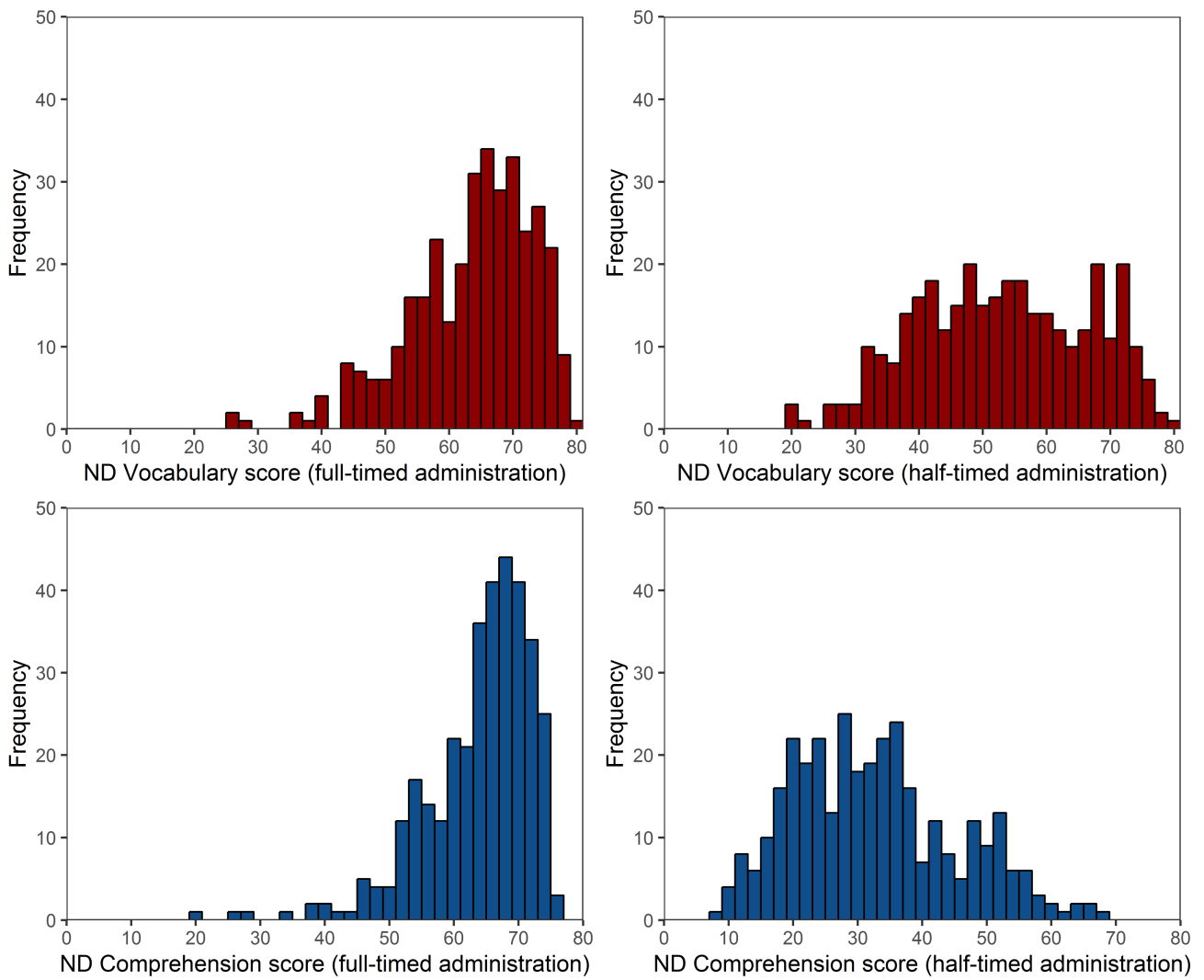


Figure 3: Frequency distributions of scores on the Vocabulary (upper panels) and Comprehension (lower panels) subtests of the Nelson-Denny Reading Test for participants tested under full-timed versus half-timed administration conditions.

APPENDIX

Spelling Dictation Test

Administration instructions: The experimenter reads the word aloud to the participant, followed by the sentence containing the word.

1. ABSTINENCE

The ex-alcoholic found it very difficult to maintain complete abstinence from drinking.

2. ACQUAINTANCE

She knew the woman as an acquaintance, but she was not a close friend.

3. DIGESTIBLE

The nurses had to blend the food to make it digestible for the patient.

4. CONCILIATORY

She tried to adopt a conciliatory approach to avoid further conflict.

5. PISTACHIO

The biscuit with pistachio nuts was delicious.

6. WARRANTY

The TV seemed a good buy because it had a 3 year warranty.

7. RHEUMATIC

The old woman suffered rheumatic pain in all of her joints.

8. CRESCENDO

The music reached a crescendo towards the end of the symphony.

9. ASYMMETRY

She found the asymmetry of the design very appealing.

10. AFFLUENT

By comparison with most Asian countries, Australia is very affluent.

11. DILIGENT

Most of the students are lazy but this boy is very diligent in completing his work.

12. AGGRAVATION

The noise from the next classroom was a constant aggravation to the teacher.

13. COLLOQUIAL

It is usually not appropriate to use colloquial language in an essay.

14. EUPHORIC

The student felt euphoric when she completed her last exam.

15. BROCCOLI

Children often dislike broccoli and other green vegetables.

16. SOMERSAULT

The gymnast turned a somersault before landing back on the bar.

17. OBLIVION

Many sportsmen achieve fame when they are young but then sink into oblivion.

18. RHYTHMICAL

The rhythmical beat of the drums was mesmerizing.

19. RAVENOUS

She had missed lunch so by dinner time she felt ravenous.

20. PERSUADE

She tried to persuade him to her point of view.

Spelling Recognition Test

Administration instructions: The participant is given unlimited time to select all the incorrectly spelled items.

Note: The item 'behaviour' should be replaced by 'behavior' if administering the test to speakers of US English.

PLEASE CIRCLE ALL ITEMS BELOW THAT YOU THINK ARE SPELLED INCORRECTLY

attitude	criticism	benafit	refrences
misary	psycology	political	glamorous
reciept	available	admission	tounge
appreciate	materilistic	independent	chronicle
seperate	senior	behaviour	attorney
sufficient	efficiency	Implie	courtesy
mortgage	govenment	basicly	privalege
consequence	sieze	suspicious	prosedure
conveinient	insurance	imminant	guitar
elementary	sacrifice	commitment	decrepit
jeapordise	forfeit	fulcrum	annihlate
distinguish	inquirey	sincirely	equivical
gaurantee	delecate	bachelor	annual
necesscarily	favourate	announcment	severe
occurence	insatiable	partitionining	asure
exhibition	warrent	interrogate	havoc
conscientious	parallel	interpretation	bureaucracy
importent	negotiate	proliferate	vigilent
missellaneous	curriculum	plagarism	acomplce
pollution	permanent	aplause	subpoena

accommodation

attentions

rendezvous

subtlety

honorable

inhibition

classified

assessor

Table A1

Descriptive statistics, communalities (h^2), factor loadings (λ), Rasch difficulty indices, and point biserial correlations (r_{pb}) for 20 Spelling Dictation Test items ordered from easiest to most difficult according to the Rasch index.

Serial position	Item	Mean	SD	h^2	λ	r_{pb}	Rasch Difficulty	% Correct (N = 742)
14	euphoric	0.88	0.32	0.22	0.47	0.48	-1.94	88.4
20	persuade	0.89	0.31	0.12	0.34	0.37	-1.87	88.9
17	oblivion	0.87	0.34	0.20	0.45	0.47	-1.68	86.5
6	warranty	0.87	0.34	0.07	0.26	0.31	-1.60	87.1
19	ravenous	0.78	0.41	0.24	0.49	0.51	-0.97	78.4
10	affluent	0.77	0.42	0.20	0.44	0.48	-0.83	76.7
18	rhythmical	0.69	0.46	0.20	0.45	0.49	-0.33	69.3
13	colloquial	0.69	0.46	0.16	0.40	0.44	-0.31	69.3
11	diligent	0.68	0.47	0.08	0.28	0.36	-0.21	67.5
9	asymmetry	0.59	0.49	0.08	0.28	0.36	0.21	59.2
15	broccoli	0.59	0.49	0.16	0.40	0.45	0.23	59.2
1	abstinence	0.59	0.49	0.27	0.52	0.55	0.25	58.8
8	crescendo	0.54	0.5	0.29	0.54	0.56	0.50	54
5	pistachio	0.53	0.5	0.11	0.33	0.39	0.52	53.1
12	aggravation	0.52	0.5	0.35	0.59	0.61	0.59	51.8
2	acquaintance	0.44	0.5	0.26	0.51	0.54	1.03	43.8
16	somersault	0.43	0.49	0.23	0.48	0.52	1.08	42.6
3	digestible	0.4	0.49	0.15	0.39	0.44	1.19	39.6
7	rheumatic	0.33	0.47	0.28	0.53	0.54	1.68	32.6
4	conciliatory	0.19	0.39	0.11	0.33	0.34	2.45	18.9

Table A2

Norms for the Spelling Dictation Test. Percentile ranks indicate the percentage of the sample with the same or lower score.

Spelling Dictation score	Percentile (N=798)
1	0.63
2	1.38
3	2.26
4	4.39
5	6.14
6	9.15
7	13.66
8	19.55
9	23.93
10	33.71
11	42.11
12	49.12
13	57.89
14	66.29
15	74.81
16	82.33
17	88.60
18	93.61
19	98.37
20	100.00

Table A3

Descriptive statistics, communalities (h^2), factor loadings (λ), Rasch difficulty indices, and point biserial

correlations (r_{pb}) for 88 Spelling Recognition Test items ordered from easiest to most difficult according to the

Rasch index. Items in bold are incorrect spellings.

Serial position	Item	Mean	SD	h^2	λ	r_{pb}	Rasch Difficulty	% Correct ($N = 308$)
4	appreciate	0.00	0.00					100.0
12	distinguish	0.00	0.00					100.0
16	exhibition	0.00	0.00					100.0
79	annual	0.00	0.00					100.0
8	consequence	0.00	0.06	0.01	0.09	0.08	-3.55	99.7
10	elementary	0.00	0.06	0.00	-0.02	0.02	-3.43	99.7
75	guitar	0.00	0.06	0.00	-0.04	-0.03	-3.33	99.7
40	negotiate	0.01	0.08	0.00	0.05	0.08	-2.81	99.4
46	political	0.01	0.08	0.00	0.05	0.11	-2.82	99.4
49	behaviour	0.01	0.08	0.01	0.10	0.12	-2.87	99.4
52	suspicious	0.01	0.08	0.00	-0.05	0.01	-2.67	99.4
1	attitude	0.01	0.10	0.00	0.03	0.06	-2.34	99.0
20	pollution	0.01	0.10	0.01	0.08	0.10	-2.39	99.0
25	available	0.01	0.10	0.01	0.08	0.10	-2.38	99.0
44	inhibition	0.01	0.10	0.01	-0.08	-0.01	-2.24	99.0
45	benafit	0.99	0.10	0.07	0.27	0.24	-2.60	99.0
6	sufficient	0.01	0.11	0.00	-0.01	0.05	-2.01	98.7
32	sacrifice	0.01	0.11	0.00	0.05	0.09	-2.06	98.7
70	chronicle	0.01	0.11	0.00	0.03	0.06	-2.02	98.7
31	insurance	0.02	0.13	0.01	0.07	0.08	-1.83	98.4
61	interpretation	0.02	0.14	0.00	0.07	0.12	-1.65	98.1
80	severe	0.02	0.14	0.02	0.15	0.16	-1.68	98.1
21	accommodation	0.02	0.15	0.00	-0.01	0.05	-1.41	97.7
28	efficiency	0.02	0.15	0.01	0.09	0.15	-1.50	97.7
27	senior	0.03	0.16	0.00	-0.04	-0.01	-1.25	97.4
43	attentsion	0.97	0.18	0.12	0.35	0.31	-1.25	96.8
74	prosedure	0.97	0.18	0.17	0.42	0.36	-1.30	96.8
60	interrogate	0.04	0.19	0.00	-0.04	0.05	-0.91	96.4
66	classafied	0.96	0.19	0.11	0.34	0.33	-1.15	96.4
82	havoc	0.04	0.19	0.00	0.05	0.11	-0.89	96.1
18	importent	0.96	0.20	0.06	0.24	0.25	-0.89	95.8
57	bachelor	0.04	0.20	0.01	0.11	0.18	-0.84	95.8
39	parallel	0.05	0.21	0.00	0.01	0.06	-0.70	95.5

72	courtesy	0.05	0.22	0.00	0.02	0.08	-0.54	94.8
62	proliferate	0.06	0.23	0.00	0.04	0.11	-0.50	94.5
37	insatiable	0.06	0.23	0.01	0.12	0.16	-0.48	94.2
7	mortgage	0.07	0.26	0.06	0.25	0.26	-0.25	92.5
54	commitment	0.07	0.26	0.01	-0.08	0.00	-0.11	92.5
24	psychology	0.92	0.27	0.02	0.16	0.16	-0.15	92.2
42	permanent	0.08	0.27	0.02	0.14	0.19	-0.15	92.2
2	misary	0.92	0.27	0.05	0.23	0.25	-0.13	91.9
48	independent	0.08	0.28	0.00	0.04	0.10	-0.01	91.6
50	imlie	0.92	0.28	0.19	0.44	0.41	-0.15	91.6
33	forfeit	0.09	0.28	0.03	0.16	0.21	-0.01	91.2
73	privalege	0.91	0.29	0.16	0.40	0.39	-0.04	90.9
56	sincirely	0.90	0.30	0.22	0.47	0.44	0.02	90.3
64	aplause	0.90	0.31	0.12	0.35	0.37	0.15	89.6
65	rendezvous	0.12	0.32	0.03	0.18	0.27	0.34	88.3
17	conscientious	0.12	0.33	0.06	0.24	0.29	0.36	87.7
67	refrences	0.88	0.33	0.17	0.41	0.41	0.34	87.7
35	delecate	0.87	0.34	0.20	0.45	0.42	0.40	87.0
26	materilistic	0.87	0.34	0.27	0.52	0.49	0.41	86.7
71	attorney	0.86	0.35	0.26	0.51	0.48	0.49	86.0
9	conveinient	0.85	0.36	0.15	0.39	0.37	0.56	85.1
85	acomplce	0.85	0.36	0.16	0.40	0.37	0.60	85.1
77	annihlate	0.85	0.36	0.09	0.31	0.30	0.62	84.7
47	admission	0.84	0.36	0.26	0.51	0.51	0.65	84.4
22	honerable	0.84	0.37	0.08	0.29	0.31	0.69	84.1
36	favourate	0.84	0.37	0.17	0.41	0.40	0.67	84.1
34	inquirey	0.82	0.38	0.31	0.55	0.53	0.80	82.5
41	curriculum	0.82	0.38	0.20	0.45	0.43	0.82	82.5
88	assessor	0.18	0.38	0.00	0.03	0.11	0.88	82.1
19	missellaneous	0.82	0.39	0.31	0.55	0.52	0.83	81.8
29	govenment	0.81	0.39	0.10	0.32	0.35	0.88	81.2
83	bureaucracy	0.19	0.39	0.01	0.07	0.17	0.93	81.2
59	partitionining	0.80	0.40	0.11	0.33	0.35	1.02	79.5
51	basicly	0.79	0.41	0.13	0.36	0.36	1.05	79.2
81	asure	0.79	0.41	0.18	0.43	0.44	1.08	78.9
38	warrent	0.78	0.42	0.25	0.50	0.47	1.12	77.9
86	subpoena	0.22	0.42	0.01	0.09	0.19	1.15	77.9
11	jeopardise	0.77	0.42	0.03	0.18	0.19	1.19	76.9
55	fulcrum	0.23	0.42	0.01	-0.08	0.01	1.24	76.6
87	subtlety	0.27	0.44	0.01	0.07	0.19	1.40	73.4
14	necessarily	0.73	0.45	0.15	0.39	0.40	1.42	72.7
76	decrepit	0.30	0.46	0.00	-0.03	0.09	1.58	69.8
13	gaurantee	0.68	0.47	0.19	0.43	0.43	1.74	67.9
23	criticism	0.64	0.48	0.28	0.53	0.54	1.98	63.6

3	reciept	0.60	0.49	0.20	0.45	0.45	2.12	60.4
69	tounge	0.60	0.49	0.17	0.41	0.42	2.11	60.4
53	imminant	0.58	0.49	0.27	0.52	0.54	2.24	58.1
58	announcment	0.49	0.50	0.23	0.48	0.48	2.69	48.7
78	equivical	0.44	0.50	0.15	0.39	0.43	2.88	44.2
5	seperate	0.42	0.49	0.23	0.48	0.52	3.07	41.9
30	sieze	0.38	0.49	0.13	0.36	0.38	3.15	38.0
68	glamorous	0.37	0.48	0.04	0.21	0.22	3.08	37.0
15	occurence	0.34	0.47	0.03	0.17	0.19	3.22	34.1
63	plagarism	0.33	0.47	0.18	0.42	0.48	3.57	32.8
84	vigilent	0.27	0.44	0.07	0.26	0.29	3.71	26.6

Table A4

Norms for the Spelling Recognition Test. Percentile ranks indicate the percentage of the sample with the same or lower score.

Spelling Recognition score	Percentile (N=799)
48	0.13
49	0.13
50	0.13
51	0.25
52	0.25
53	0.50
54	0.50
55	0.75
56	1.38
57	2.25
58	3.25
59	4.01
60	4.63
61	5.76
62	6.51
63	8.01
64	9.76
65	12.39
66	14.27
67	17.65
68	19.65
69	21.28
70	23.78
71	26.78
72	31.79
73	35.92
74	40.68
75	45.56

76	50.44
77	55.94
78	61.45
79	67.21
80	74.34
81	78.60
82	83.60
83	89.11
84	93.49
85	96.25
86	98.00
87	99.37
88	100.00

Table A5

Component loadings for the Full+Half sample for spelling dictation, spelling recognition, Nelson-Denny vocabulary, and comprehension

	Full-timed ND (<i>n</i> = 107)		Half-timed ND (<i>n</i> = 103)	
	Component 1	Component 2	Component 1	Component 2
Spelling Dictation	.94	.01	.92	.05
Spelling Recognition	.95	-.01	.97	-.05
NDT Vocabulary	-.03	.94	-.02	.91
NDT Comprehension	.04	.90	.02	.90
Total variance explained	63.54%	23.25%	60.88%	25.06%
Inter-component correlation	.46		.41	

Table A6

Component loadings for the Full+Half sample for spelling dictation, spelling recognition, Nelson-Denny vocabulary, comprehension, and reading rate

	Full-timed ND (<i>n</i> = 107)		Half-timed ND (<i>n</i> = 103)	
	Component 1	Component 2	Component 1	Component 2
Spelling Dictation	.02	.93	.05	.92
Spelling Recognition	-.02	.96	-.05	.97
NDT Vocabulary	.86	-.01	.80	.04
NDT Comprehension	.87	.04	.89	.02
NDT Reading Rate	.70	-.02	.82	-.05
Total variance explained	56.09%	19.32%	55.80%	22.25%
Inter-component correlation	.48		.42	