Effects of individual differences in verbal skills on eye-movement patterns during sentence reading

Kuperman, V.[a,*], & Van Dyke, J. A.[b]

[a] Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 626, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M2

[b] Haskins Laboratories, 300 George Street, New Haven, CT 06511, United State

* Corresponding author. E-mail addresses: vickup@mcmaster.ca (V. Kuperman), jvandyke@ haskins.yale.edu (J.A. Van Dyke).

Abstract
This study is a large-scale exploration of the influence that individual reading skills exert on eye-movement behavior in sentence reading. Seventy-one non-college-bound 16– 24 year-old speakers of English completed a battery of 18 verbal and cognitive skill assessments, and read a series of sentences as their eye-movements were monitored. Statistical analyses were performed to establish what tests of reading abilities were predictive of eye movement patterns across this population and how strong the effects were. We found that individual scores in rapid automatized naming and word identification tests (i) were the only participant variables with reliable predictivity throughout the time-course of reading; (ii) elicited effects that superceded in magnitude the effects of established predictors like word length or frequency; and (iii) strongly modulated the influence of word length and frequency on fixation times. We discuss implications of our findings for testing reading ability, as well as for research of eye-movements in reading.

Keywords
Eye-movements; Individual differences; Reading ability; Rapid automatized naming; Psychometric tests

Introduction

Reading behavior is defined not only by the ability to recognize individual words and integrate them into a sentential discourse, but also by the ability to do so repeatedly in a rapid, sequential, and directional fashion, as determined by the conventions of a particular language's writing system. A close coupling between word recognition processes and higher-order processes of comprehension, on the one hand, and eye-movement control during reading, on the other hand, is frequently assumed in eye movement research (Boland, 2004; Just & Carpenter, 1980; Rayner & Pollatsek, 2006). A number of computational models of eye-movements in reading allocate a crucial role to processes of word identification during sentence comprehension. For instance, the E–Z Reader model focuses on word identification as the ''engine'' that drives the eyes forward in reading (e.g., Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Pollatsek, & Rayner, 2006; Reichle, Rayner, & Pollatsek, 2003) as well as recognizes the influence of post lexical sentence comprehension processes on the speed and quality of reading (Reichle, Warren, & McConnell, 2009). Another model, Glenmore, treats lexical properties of words as contributors to the pattern of activation that triggers saccade generation (Reilly & Radach, 2003, 2006), while the SWIFT model considers word identification as a major factor inhibiting the visuo-oculomotor process of stochastic saccade generation (Engbert, Longtin, & Kliegl, 2002; Engbert, Nuthmann, Richter, & Kliegl, 2005; Richter, Engbert, & Kliegl, 2006). Although these models differ in a variety of important ways, they have all been influenced by a large body of research pointing to properties of words as the major determiner of when or where eyes move during reading. For example, a very robust finding is that readers look longer at low-frequency words than at high frequency words (e.g., Altarriba, Kroll, Sholl, & Rayner, 1996; Inhoff & Rayner, 1986; Kliegl, Grabner, Rolfs, &Engbert, 2004; Rayner & Duffy, 1986). Other text variables that have been shown to determine fixation times include word length (Rayner (2009) and references therein), the extent to which words are predictable from context (e.g., Ashby, Rayner, & Clifton, 2005; Ehrlich & Rayner, 1981; Rayner & Well, 1996, and references in Rayner, 1998), how ambiguous a word is (e.g., bank = side of river vs. financial institution; Rayner & Duffy, 1986), and morphological complexity (e.g., Hyönä, Niemi, & Underwood, 1989)

While this large body of research makes it clear that characteristics of words and texts affect both lexical processing and eye-movements, a separate strand of research in the reading domain has focused on participant characteristics, specifically asking what abilities differentiate skilled from less-skilled readers. For example, Perfetti's (1985, 2007) Verbal Efficiency Theory suggests that skilled readers are those that have developed high quality lexical representations, characterized by (i) highly automatic associations of precise orthographic forms to the phonological representations learned during oral language acquisition; (ii) automatic associations of these same phonological representations to semantic representations; and (iii) a highly elaborated and redundantly specified semantics, encompassing the full variety of syntactic and semantic contexts for the word (cf. Harm & Seidenberg, 2001;

Seidenberg & McClelland, 1989 for a computational implementation of some aspects of this idea in a continuous-activation connectionist model). According to this approach, low-quality representations lead to comprehension difficulty because the lack of automatic and/or precise associations either at the junction of orthography–phonology or phonology-

semantics causes information necessary for integrating a word into its sentential context to be unavailable at the time when it is needed. Thus, poor comprehension can arise either from inefficient representation of the phonological procedures that allow the orthographic form of a word to be decoded, or else from imprecise or incomplete representations of meaning. Deficits in either area will lead to inefficient word identification, which in turn will lead to poor comprehension.

From this, we see that efficient word recognition is not simply about linguistic characteristics of words, but rather about linguistic characteristics of particular words as learned by particular individuals. Individuals will differ in the average lexical quality of their words, with this variation related to the stability of the knowledge a reader has about a particular word's form, sound and meaning. This intuition finds support within the eye-movement and other behavioral literature in a handful of studies that have investigated subjective aspects of lexical variables, such as subjective familiarity of a word (e.g., Chaffin, Morris, & Seely, 2001; Juhasz & Rayner, 2003; Williams & Morris, 2004), a word's subjective frequency (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), and age of acquisition (e.g., Balota et al., 2004; Juhasz, 2005; Juhasz & Rayner, 2006; Zevin & Seidenberg, 2002)

An important aspect of the lexical quality hypothesis is the notion of a qualitative, experience-dependent shift from deliberate constituent-based decoding (e.g., letter by-letter, or syllable-by-syllable reading) to a unitary, fully-specified representation of individual words (Andrews, 2008; Ehri, 1999; LaBerge & Samuels, 1974; Perfetti, 1992; Perfetti, 2007). Evidence for such a shift in high-skilled readers comes from a variety of sources. For example, Samuels, LaBerge, and Brener (1978), found a decreasing developmental trend in the size of the word length effect on semantic judgments about orthographically regular words in 2nd, 4th, and 6th grade samples, until it was completely absent for college-level students. The crucial role of experience for this effect is apparent in several related studies (Samuels, Miller, & Eisenberg, 1979; Terry, Samuels, & LaBerge, 1976) in which this interaction disappeared when words were presented in a degraded form (i.e., mirror-image or with partially erased letters). Thus, in the absence of experience with particular word-forms, adults will also resort to componential processing.

Additional evidence for the relationship between lexical quality and reading experience comes from the letter detection task (e.g., Cunningham, Healy, Kanengiser, Chizzick, & Willitts, 1988; Healy, 1994; Saint-Aubin and Klein; 2008). Here, readers have difficulty detecting particular letters within familiar words as compared to less familiar words, pointing to the use of unitary perceptual chunks for overlearned orthographic forms. Notably, this effect interacts with reading skill, even after controlling for grade level, such that better readers have more difficulty than poor readers, presumably due to interference from word-level perceptual chunks that are not present for poor readers. This result is consistent with eye movement studies investigating the size of the perceptual span in beginning and disabled readers (e.g., Häikiö, Bertram, Hyönä, & Niemi, 2009; Rayner, 1986; Rayner, Murphy, Henderson, & Pollatsek, 1989; Rayner, Slattery, & Bélanger, 2010). These studies are conducted using an eye-contingent display in which only an experimentally manipulated portion of text surrounding the readers' exact fixation point is visible to the reader. Reading rates are monitored as the size of the perceptual span is increased until asymptote is reached, at which point it is assumed that the readers' perceptual span during

natural reading is equivalent to the last measured perceptual window. A consistent finding from these studies is that beginning or less-skilled readers' perceptual spans are about 3–4 characters smaller than those of experienced readers. This reduction is generally understood as a function of the amount of resources required for processing the fixated word (e.g., Rayner et al., 2010) and can be related to lexical quality because low quality words—those which must be processed in a constituent-based manner—require greater processing resources than high-quality, lexicalized words.

Finally, evidence from neuroimaging suggests that a hallmark of reading skill is reduced activation in ventral regions of the occipital temporal cortex (dubbed the ''visual word-form area'') where orthographic, phonological, and lexical-semantic features of a word are thought to converge into full-form lexical representations (see McCandliss, Cohen, and Dehaene (2003) for a review). The fact that these skill-based reductions are accompanied by reductions in regions of the inferior frontal gyrus (IFG), which have been associated with sublexical phonological assembly (e.g., Poldrack & Wagner, 2004; Sandak et al., 2004; Shaywitz et al., 2002) further attests to a shift away from constituent-based processing.[1]
Understanding the participant-specific variables that enable individuals to shift from constituent-based reading to full-form word recognition has been a primary goal of applied reading research. A number of general ability measures such as IQ and memory have been examined, however measures of phonological processing have received the most attention (e.g., Brady, Braze, & Fowler, in press; McCardle & Pugh, 2009; Rieben & Perfetti, 1991) due to the importance of phonology as the mediating code between a language's writing system and oral vocabulary knowledge. A readers' ability to learn and use the grapheme-to-phoneme correspondence rules that link a word's orthography to its meaning (i.e., to decode the word) forms a core component of reading ability, and the repeated application of these rules has been argued to provide a self-teaching function, which promotes the shift from low to high-quality lexical representations on the orthographic dimension (Jorm & Share, 1983; Share, 1995).

A second well-documented component of reading ability, which together with phonological decoding has been argued to comprise a double deficit for severely disabled (dyslexic) readers (e.g. Wolf & Bowers,1999; Wolf & O'Brien, 2001; Wolf et al., 2002) is the ability to rapidly access orthographic information, usually measured via naming tasks. This deficit, which may relate to a broader speed of processing deficit (e.g. Catts, Gillespie, Leonard, Kail, & Miller, 2002; Kail, Hall, & Caskey, 1999; Savage, 2004) has a negative impact on text-level reading fluency (Manis & Freedman, 2001; Wolf & Katzir-Cohen, 2001) as well as word-level recognition processes.

Measures of decoding and speeded naming abilities serve as only two examples of the wide variety of skill measures that are examined in reading research; other examples include oral language ability, vocabulary, attention, and a variety of sub-lexical analytical skills (e.g., elision, blending, segmenting, etc.). The widely acknowledged import of individual skills such as these as determinants of successful reading acquisition points to the need for a thorough

[1] *A skill-related reduction in brain activity is consistent with studies of perceptual and motor skill learning in which initial (unskilled) performance is associated with increased activation in task-specific cortical areas, to be followed by task-specific decreases in activation in the same cortical regions after continued practice (e.g., Poldrack & Gabrieli, 2001; Ungerleider, Doyon, & Karni, 2002; Wang, Sereno, Jongman, & Hirsch, 2003)*

examination of how participant variables may affect the eye-movement record in sentence reading. So far, only a handful of studies in the current eye movement literature have taken up this challenge (for early investigations see Pavlidis, 1985; Rayner, 1985b; Schilling, Rayner, & Chumbley, 1998). These have examined a small set of skill measures, however, most of which are not specifically related to reading ability. For example, reader age is perhaps the most frequently studied variable, with comparisons made either between children and adult readers (e.g., Blythe, Liversedge, Joseph, White, & Rayner, 2009; Huestegge, Radach, Corbic, & Huestegge, 2009; Joseph, Liversedge, Blythe, White, & Rayner, 2009; Lefton, Nagle, Johnson, & Fisher, 1979; Rayner, 1986, 2009), or young adults with senior citizen readers (e.g., Kliegl et al., 2004; Laubrock, Kliegl, & Engbert, 2006; Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006). A second measure that has received extensive attention is working memory capacity, measured via the Daneman and Carpenter sentence span task. For example, Clifton et al. (2003), observed that low-span readers spent more time re-reading the ambiguous regions of reduced relative clauses than high span readers; but see Traxler (2009). Also, Traxler (2007) showed that low-span readers had more difficulties (showed longer regression path duration and total fixation time) with integrating relative clauses with preceding sentence fragments, if these clauses had an ambiguous attachment. Moreover, Kennison and Clifton (1995) and Osaka and Osaka (2002) tested whether differences in working memory capacity correlate with individual sensitivity to parafoveal preview in English and Japanese.

While measures of reading-specific skills have been used in some studies to match for the performance level across groups of participants (Radach, Huestegge, & Reilly, 2008), only a few studies have used such measures to contrast ability groups. Thus, Hyönä and Olson (1995) used the Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970; described further in Appendix A), as a criterion for dyslexia: they manipulated word length and frequency, and reported no group differences in the eye movements across dyslexic and nondyslexic younger readers as a function of either manipulated variable. Also, a test of phonological awareness (Pig Latin) was used as a predictor of eye-movement patterns in reading and non-reading tasks in Eden, Stein, Wood, and Wood (1994), revealing a correlation between poor phonological skills and difficulties in saccade planning.

The only other individual difference measure to receive considerable attention in eye-tracking studies has been the Nelson–Denny test, which provides a global measure of reading comprehension. Ashby et al. (2005) employed this measure to compare eye-fixation times in a cohort of average readers compared to a cohort of skilled readers, and found differential effects of word frequency and contextual predictability across the two cohorts. Both cohorts were slower when reading low-frequency and low-predictability words, but the average readers were slowed down more than skilled readers. Chace, Rayner, and Well (2005) also used this measure to reveal that less skilled comprehenders did not activate phonological codes of words visually available in the parafovea and generally obtained a lesser amount of parafoveal preview benefit than skilled readers did (see also Jared, Levy, and Rayner (1999) for individual differences in phonological processing during silent reading). While these studies offer valuable insights into the link between individual skills and strategies of reading, they are limited by the generality of the comprehension measure, which does not distinguish variation in particular reading sub-skills (e.g., word decoding, naming speed). Yet those sub-skills may directly impact particular stages of lower-level ocular-motor planning.

The present paper contributes to filling this lacuna by simultaneously considering multiple tests of reading and linguistic abilities as predictors of the eye-movement record in sentence reading.

The current research also seeks to address an additional limitation of previous research on individual differences in eye-movements of non-clinical populations—namely, that apart from research with children or the aging population, they have utilized undergraduate students in university subject-pools for their participants. While this is a conveniently available population, studies of reading abilities among high school and college students have revealed wide variability in a variety of basic reading-related skills, such as accuracy and fluency of word recognition and in non-word reading (e.g., Cunningham, Stanovich, & Wilson, 1990; Shankweiler, Lundquist, Dreyer, & Dickinson, 1996) – skills which are often assumed to have been mastered by the end of 2nd or 3rd grade. Indeed, the 2005 Nation's Report Card (Grigg, Donahue, & Dion, 2007) indicates that over five million adolescents in the United States are not able to adequately read or understand textbooks, teaching materials or assignments in their core academic classes (see also National Center for Education Statistics, 2005). This means that important participant-level determinants of eye-movements may be systematically understudied in previous research. Moreover, the focus on university students as participants only exacerbates the situation (Peterson, 2001): Although variation can be found, the range of this variation will be necessarily limited by college admission requirements and the self-selection of those who choose to matriculate.

We address this limitation by recruiting a community based sample of non-college-bound adolescents (ages 16–24), a population group which the National Center for Education Statistics in the US estimated as including approximately 39% of high school seniors in 2004, the most recently studied cohort (Ingels, Dalton, & LoGerfo, 2008). A variety of factors may lead to poor reading ability in the cohort under consideration (Barth, Catts, & Anthony, 2009). Some may suffer from deficits in the same basic skills as elementary students (i.e., poor phonological skills, inadequate decoding ability), while others may struggle because of the cumulative effect of poor experience (i.e., deficient vocabulary, insufficient comprehension skills). There is also the possibility that poor readers may be deficient in more general cognitive abilities, including memory span, processing speed, or general language ability. Thus, we seek to characterize our population via a broad battery of individual difference measures that includes many basic skills typically used to assess risk of reading failure in early grades, as well as measures of overall reading and language abilities and associated cognitive skills. This enables us to simultaneously investigate the correlates of poor reading skill in an understudied population and to obtain a range of skill variation that may be more comparable to that of the population at large. The battery of skill measures is comprised of normed standardized tests (with a few specified exceptions), enabling us to assess reading-grade level (and in some cases, reading-age level) for each participant based on his/her raw score. We provide the demographics for our population based on these tests in Table 1, although all analyses were conducted using raw scores only, rather than grade equivalents. For this population, we seek to establish (i) which skill measures are predictive of eye movements elicited when silently reading a sentence for comprehension, (ii) how strong the effects of those skill measures are, and how early (late) in the time-course of reading do they exert their influence; and (iii) how individual skills

modulate the effects of well-known predictors of eye-movements, such as frequency of occurrence and length of a fixated word.

Method

Participants

Our data source is the eye-movement record for 81 English sentences read silently in isolation by 71 participants. Prior to participating in the eye-tracking experiment, the participants undertook a battery of reading ability tests. Participants (43 females; 28 males) belonged to the age group of 16–24 (mean 20.8; SD = 2.6), and were not college-bound. They were recruited from the local community in a of ways, including: presentations at adult education centers; advertisements in local newspapers; posters/flyers placed on adult school and community college campuses, public transportation hubs, local retail and laundry facilities. All were native English speakers, and none had a diagnosed reading or learning disability. They were paid $12.50/h for 2.5 h of testing (1.5 h of reading skills testing and 1 h of eye-tracking time).

Materials

The sentences analyzed for the present report were selected post hoc from two sets of materials that were interleaved in a single eye-tracking study (i.e. separate experimental designs served as fillers for each other). Each of these sets contained 36 items, each with four conditions, which were counter-balanced across four-lists via a Latin square design, together with a third experiment from which no sentences were selected for the current analyses. To these 108 (3x36) sentences, an additional set of 36 filler sentences was added so that each participant read 144 sentences in a single session. Due to counterbalancing, no single participant saw all four versions of the experimental items comprising each of the 2 experiments, however, each participant did see the full set of filler sentences.

From this sentence corpus, we selected a set of 81 sentences, with each participant reading about 28 of these due to the effects of counterbalancing (i.e. a sentence chosen for analysis may appear in only 1 list out of 4). The goal of sentence selection was to enable us to examine eye movements in three types of syntactic structures: (i) syntactically non-complex sentences: The kind landlord improved the aging building; (ii) sentences with an embedded relative clause: The spy that encoded the message delivered the secret to the authorities; and (iii) sentences with a relative clause embedded in a complement clause: The husband knew that the wife who is devoted to the beautiful home would cry at the ceremony. Sentences of types (ii) and (iii) had no syntactic or semantic interference, defined in Van Dyke (2007) as the situation in which words occurring between a head and a dependent term (spy and delivered in example (ii)) do not overlap with either (or both) syntactic or semantic features of that dependent term (i.e., the NP message may not serve as a subject for delivered in (ii) and home cannot be the subject of cry in (iii)). We opted for including this variety of structures in our analysis as a representation of syntactic complexities that are abundant in naturalistic English texts. Because the items were not designed with the current analysis in mind, the number of data points associated with each sentence varies by its type of syntactic structure. Namely, all sentences with syntactic type (i) occurred in the filler list which was seen by all participants, while those of types (ii) and (iii) were part of separate

experimental designs and hence any particular sentence was seen by only 1 in 4 participants. In addition, sentences of type (i) occurred in the filler list together with sentences of a variety of other more complex syntactic types (e.g., passives, clefts) which were not chosen here. Hence, only 11 sentences of type (i) were available, while 35 instances of types (ii) and (iii), which occurred as part of structured experimental designs were available.[2] We provide the full list of sentences in Appendix B.

Skill measures

Skill measures were chosen based on previous research outlining the component factors of reading ability (e.g., Perfetti, 1985; Scarborough, 1998; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997; Vellutino, Scanlon, Sipay, Pratt, Chen, & Denckla, 1996). Four main areas were assessed, targeting ability at the sub-word, word, and sentence-level: (1) Phonological ability, including phonological awareness, phonological memory, and naming ability; (2) Word and non-word reading; (3) Sentence comprehension; and (4) Working Memory. A detailed description of the 18 tests that constituted the battery of skill measures is provided in Appendix A, while their descriptive statistics and labels used in statistical models are summarized in Table 1.

Procedure

Participants were seated in front of a 17-in. display with a refresh rate of 85.03 Hz with their eyes approximately 64 cm from the display. They wore an Eye Link II head mounted eye tracker (SR Research, Mississauga, Ontario, Canada), sampling at a rate of 250 Hz from both eyes. Sentences were presented one at a time on a single line, with a maximum of 90 characters, using a monospace font. Type size was such that 3.5 characters occupied 1 degree of visual angle. The eye tracker was calibrated using a series of nine fixed targets distributed around the display, followed by a 9-point accuracy test. Calibration was monitored throughout the experiment and was repeated after any breaks or whenever the experimenter judged necessary. Data were collected from both eyes, but analyses were done only on the right eye for all participants except one, whose right eye would not calibrate. Data from this participant's left eye were used for the analyses.

Prior to the experiment, participants were instructed to read each sentence for comprehension and told that they would be required to answer a comprehension question. Comprehension questions occurred on 55% of trials. Participants were also told that they could take a break at any time during the experiment. Each trial began with a screen containing a fixation point in the middle left of the display. While fixating on this point, participants were to press a button to bring up a sentence (the sentence would not appear unless participants fixated on the fixation point). After they had read the sentence, participants pressed the same button to view the comprehension question. The question appeared in the center of the screen; two possible answers appeared three lines below, one to the left of center and one to the right of center. Participants indicated their answer by pressing the associated button on a button box; for example, if the answer they chose appeared to the left of center they were to press the left button. The position of the correct

[2] *One item from the 36 in each experiment was dropped from the analysis due to typographic error.*

answer was counterbalanced throughout the experiment. Participants were limited to 10 s for reading the stimulus sentence and 30 s for answering the comprehension question. If participants had not signaled that they had completed reading the sentence before the 10-s limit, the computer moved onto the comprehension question automatically. This occurred in less than 5% of trials. Participants were told to make their best guess at the comprehension question if they were unsure of the answer. If they had not answered within the 30-s limit, the computer moved onto the next item. This occurred in less than 1% of trials.

Dependent variables

The eye-movement record is sufficiently fine-grained to allow for tracking the influence of factors of interest at different stages of visual, lexical and higher-level processing (Carreiras & Clifton, 2004; Rayner, 1998, 2009). Here we explored a broad range of eye-tracking measures to identify which stages are influenced by the various abilities of our participants, as reflected in their individual test scores. The accuracy of responses to comprehension questions did not shed any additional light on lexical processing and is not reported further. The dependent variables considered in this study are summarized in Table 2.

The earliest measure examined is the initial landing position of the first fixation on word N, which taps into the decision about where to move the eyes and is typically made before the word is foveated (i.e., prior to lexical access). First fixation duration, single fixation duration and refixation likelihood (an index of the decision made during the first fixation) are generally thought of as measures of the difficulty of the word's initial visual recognition and lexical access. Gaze duration on a word reflects the complexity of the early processing stage, namely, word identification and, especially for words that do not elicit rereading, the complexity of semantic integration of the word in the available context. Similarly, the regression likelihood points at how difficult it is to integrate the word into the sentence: the likelihood increases if there are syntactic, semantic or lexical ambiguities in the available fragment of the sentence, or if the context is difficult to memorize. Second pass likelihood, regression path duration (or go-past time), and second pass reading time are usually interpreted as late measures of the difficulty of integrating a word into a sentential context. Finally, total fixation time is a global measure of general comprehension difficulty. (See Tables 3 and 4 in Boston, Hale, Kliegl, Patil, and Vasishth (2008) for a detailed description of eye movement measures and relevant references). It is important to note that while some eye-movement measures can be attributed to earlier stages of lexical processing (e.g., initial landing position or duration of the first of multiple fixations) and others to later ones (e.g., second pass duration or likelihood), there is no one-to-one correspondence between eye-movement measures and the temporal order of word recognition in sentence reading.

Predictors

Our primary interest is in the influence of participant variables on the eye-movement record, which may be reflected in either main effects of the test scores on eye movements or in their interactions with other lexical predictors of behavioral measures. To meet this goal, we took into consideration individual scores of our participants in the tests summarized in Table 1 and described in detail in Appendix A. As other indices of individual differences, we

considered age and sex of participants. In addition to these participant variables, we included various text-level variables; these are summarized in Table 3 including their ranges, mean values, values after transformation (where applicable), and corresponding labels used in statistical models. These variables were chosen based on eye movement research of the last several decades, which has established lexical frequency and orthographic length (in characters) of words N, N -1 and N + 1 as ''benchmark'' predictors because they elicit robust effects on where and when the eyes move in the course of reading (for an overview see e.g., Kliegl, Nuthmann, & Engbert, 2006; Kliegl et al., 2004; Rayner, 1998, 2009). Word lengths were measured in characters. Lexical frequencies were obtained from the 320-million word HAL written corpus of US English (Lund & Burgess, 1996; available via the English Lexicon Project, http://elexicon.wustl.edu/). Nine words in our dataset did not occur in the HAL corpus: we assigned to these words the median word frequency of content words occurring in our dataset and in the HAL corpus. We further verified that removing these lexical items did not incur any qualitative change in our statistical models reported below.

We also took into consideration each word's relative position in the sentence, defined as the ordinal rank of the word in the sentence divided by the sentence length in words. It is a robust finding that reading times tend to be inflated on the final words of the clause (Rayner, Kambe, & Duffy, 2000; Rayner, Sereno, Morris, Schmauder, & Clifton, 1989). An explanation offered for this wrap-up effect is that the costly semantic integration of words within the clause and integration with prior discourse take place at the sentence-final words (Just & Carpenter, 1980; for alternative hypotheses also see Hill & Murray, 2000; Hirotani, Frazier, & Rayner, 2006; Warren, White, & Reichle, 2009). Finally, Kuperman, Dambacher, Nuthmann, and Kliegl (2010) proposed an oculomotor program of saccadic planning over the line of text that accounts for the steep increase in single fixation and gaze durations over the first few words of a sentence (see also Kennedy & Pynte, 2005 and references in Kuperman et al., 2010).

As the initial landing position (the position of the first fixation on the word) affects the number and durations of fixations on the word, we considered this variable as well. In typical readers, initial fixation on a word tends to land at the preferred viewing location, between the left boundary of the word and its center (Rayner, 1979). In dyslexic readers of German, however, Hawelka, Gagl, and Wimmer (2010) observed a tendency to position an initial fixation near the beginning of the word. Hawelka et al. interpreted this in terms of a dual-route word recognition model (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), suggesting that dyslexic readers are engaged in the serial sublexical grapheme–phoneme conversion (GPC) required to access the phonology of words with poor-quality lexical representations. This is in contrast to the automatized lexical processing route engaged by skilled readers.[3] We expected the initial landing position to vary as a function of verbal skill

3 Although not discussed by Hawelka et al. (2010), it is important to recognize that single-mechanism reading models also account for these effects (see Plaut, McClelland, Seidenberg, and Patterson (1996) for an extensive discussion of sublexical processing in these models; see Seidenberg (in press) for a review). Sublexical processes are invoked whenever full-form word representations are unavailable (i.e., when readers do not know the words, or when reading non-words). In single-mechanism models, like the "triangle" connectionist approach (e.g., Harm & Seidenberg, 1999; Plaut et al., 1996; Seidenberg & McClelland, 1989), sublexical processing entails a division of labor between orthographic, phonologic, and semantic neural-like units where associations between orthography and phonology become more important because direct associations between orthography and semantics are weak or unavailable. These models have demonstrated that the efficiency of sub-lexical processes is jointly determined by the precision of phonological representations and frequency of word occurrence. The latter influence is difficult to explain if words are recognized via a separate GPC route, which eschews storage of full-form lexical items.

in the present study as well. We also discuss the dependency of the initial landing position on the average saccade amplitude as a function of verbal skill. Furthermore, we utilize the initial landing position measure to examine the Inverted Optimal Viewing Position effect: fixations landing near the word's center tend to be longer and come with a lower likelihood of refixation than fixations landing at the word's extremes (for crosslinguistic reports of the Inverted Optimal Viewing Position effect in sentence reading see e.g., Kliegl et al., 2006; McDonald, Carpenter, & Shillcock, 2005; Nuthmann, Engbert, & Kliegl, 2005; Vitu, Lancelin, & d'Unienville, 2007; Vitu, McConkie, Kerr, & O'Regan, 2001).

Finally, we coded each sentence for its number in the experimental list to capture potential longitudinal effects of habituation or fatigue. The full list of continuous predictors, as well as their labels and distributional data are provided in Table 3. We also took into account the type of syntactic complexity of the sentence: the factor Type with three levels: S (Simple sentence), SE (sentence with a Single Embedded relative clause), and DE (sentence with Doubly Embedded relative clauses).

Statistical considerations

Several independent measures showed skewed distributions, which may lead to a disproportionate influence of outliers on the outcome of the statistical model. We log-transformed (base e) lexical frequencies of words N, N 1 and N + 1. Some of our predictors showed high correlations, e.g., word length and word frequency. This collinearity may give rise to inflated standard errors, i.e. a decrease in the confidence of the estimates of model parameters. To remove collinearity, we residualized log frequencies of words N, N -1 and N + 1 against the lengths of the respective words, by fitting a regression model for each of these three variables in which log frequency of the relevant word was predicted by its length. Residuals (distances between the observed and the fitted values) of these models were taken as the measure of word frequency. The residualized frequencies strongly correlated with the original frequency values (all Pearson's rs > .9; $p < .001$) and had an additional benefit of being orthogonal to the lengths of respective words: we added the prefix "r" to the labels of all residualized variables in our models and plots. As a result, collinearity was not a problem for the set of predictors that showed significant effects on reading times in our final statistical models: the values of the variance inflation factor (an index of how much the variance of an estimated regression coefficient is increased because of collinearity) were below 2 for each predictor in all models (Kutner, Nachtsheim, & Neter, 2004). Also, the condition number κ (the ratio of the largest to smallest eigenvalues in the matrix of numerical predictors, which positively correlates with collinearity) in the final models was below 10, which indicates relatively low collinearity (Baayen, 2008). To allow for comparability of regression coefficients across numerical predictors, we standardized all these predictors (where applicable, after log-transformation and residualization) by subtracting the mean value from the predictor's value and dividing the difference by one standard deviation: we added the prefix "s" to the labels of all standardized variables.

In view of the large number of predictors, we used mixed effects regression models with crossed random effects, implemented in package lme4 of the statistical software R 2.8.1 (Bates & Sarkar, 2007; R Development Core Team, 2007) that allow for the simultaneous consideration of multiple covariates, while keeping under statistical control the between participants and between-items variance (Baayen, Davidson, & Bates, 2008; Jaeger, 2008;

Pinheiro & Bates, 2000). We fitted multiple-regression models to the continuous dependent variables (i.e., durational measures and landing position), and logistic regression models to the binary dependent variables (e.g., whether or not the first fixation on a word was followed by a refixation, or the first pass was followed by a regression or a second pass). To attenuate the influence of atypical outliers on the outcome of multiple regression models, we trimmed the data sets again by removing outliers from the respective datasets, i.e., points that fell outside the range of 3.0 to 3.0 units of SD of the residual error of the model. Once outliers were removed, the models were refitted.

Eye-movement data

The original data pool contained 25,466 data points. We removed fixations that were shorter than 50 ms or longer than 1000 ms, as well as blinks and instances of track-loss (2976 data points altogether). We also excluded the initial and the final word from all sentences (3942 data points), as is a common practice in eye-tracking studies using corpus data (e.g., Kliegl et al., 2004, 2006). The word in the beginning of the string is a typical location for the initial fixation on the string and thus the processing of the sentence initial word differs from the visual uptake of sentence internal words. The processing of sentence-final words is also different due to semantic wrap-up effects and the position of the punctuation mark (Rayner et al., 2000).

At the next step of data preparation, we removed all function (closed class) words from the dataset: ‘‘for’’, ‘‘a’’, ‘‘that’’, ‘‘the’’, ‘‘in’’, ‘‘with’’, ‘‘to’’, ‘‘at’’, ‘‘from’’, ‘‘of’’, ‘‘on’’, ‘‘off’’, ‘‘his’’, ‘‘during’’, ‘‘by’’, ‘‘out’’, ‘‘up’’, ‘‘along’’, ‘‘about’’, ‘‘back’’, ‘‘near’’, ‘‘as’’, ‘‘and’’, ‘‘&’’, and ‘‘if’’. These words tend to be very high-frequency and very short, so readers show different behavioral patterns when encountering function words as opposed to content words (cf. Kliegl et al., 2006 for motivation of separate consideration of function and content words). The analyses reported below are for content (open class) words only, which comprised a set of 372 word types and averaged 4.6 content words per sentence. The exclusion of function words took out another 7835 data points from the pool, leaving us with the set of 10,613 data points. As noted above, most stimuli sentences were part of experimental lists to which only a small percentage of participants were assigned, resulting in only 11 out of 81 sentences in the data set being read by all 71 participants. In total, there were 1996 selected sentence trials, or about 28 sentences per participant. After sentence initial and final words were removed, the average number of fixations per sentence per participant was 12.7. When function words were additionally removed, the average number of fixations per sentence was 5.3, or 1.2 fixation per content word.

Results and discussion

Baseline models

For our initial analyses, we were interested in determining whether robust effects of text variables observed in the eye-movement literature with the college-student population held for our participant group. Consequently, no individual difference measures were included in this set of models. We fitted separate baseline models to each dependent variable listed in Table 2 while using the set of predictors described in Table 3 as fixed effects and

participant, word and sentence ID as random effects. Whenever warranted by the likelihood ratio test of model comparison (see Baayen et al., 2008), we modeled non-linear effects by using the orthogonal polynomials of degree 2 (as implemented in function poly in R). We opted for orthogonal polynomials rather than other approximation functions since the visual inspection of non-linear effects revealed that a quadratic parabola is an appropriate functional form to model the effects, and coefficients of orthogonal polynomial functions are readily interpretable.

Tables 7–16 in Supplementary materials 1 report specifications of the resulting baseline models throughout the entire time-course of visual processing. For linear mixed models fitted to continuous dependent variables (Tables 7–9, 11, 13, and 15–16), the output provides estimates of regression coefficients, Bayesian 95% confidence intervals and p-values obtained with the t-test and also estimated via Monte Carlo Markov chain simulations. Furthermore, we tested whether the effects of interest retained significance if the models are fitted to log-transformed, rather than raw values of continuous dependent variables: the patterns of results for baseline predictors, test scores and interactions of test scores with lexical properties were virtually identical to the patterns reported in this paper.

For binary dependent variables (Tables 10, 12, and 14), the mixed logistic regression models yield regression coefficients, standard errors, Wald's z-values and corresponding p-values; for details of model fitting and references on modeling techniques, see Supplementary materials 1. Below we only discuss the effects that had p-values below the 0.05 significance threshold in the baseline models: for linear mixed models, we required both the MCMC-based and the t-test based p-values to be below the threshold. We also report effect sizes defined in this paper as the difference between the values of dependent variables estimated for the highest score and for the lowest score by the respective regression model. The values are expressed either in milliseconds (for temporal eye-movement measures), characters (for initial landing position), or in percents of likelihood of an event (refixation, regression or second pass reading)

The baseline models for all temporal eye-movement measures replicated the well-established result that longer words N (sWordLength) elicited significantly longer reading times: [FirstFixDur: effect size = 17 ms; β = 3.47; SE = 1.40; p = .0097, SingleFixDur: effect size = 25 ms; β = 5.27; SE = 1.68; p = .0017, GazeDur: effect size = 190 ms; β = 39.6; SE = 2.73; p < .0001, RegrPathDur: effect size = 231 ms; β = 48.07, SE = 4.51; p < .0001, SecDur: effect size = 233 ms; β = 48.53; SE = 4.56; p < .0001, TotalTime: effect size = 352 ms; β = 73.35; SE = 4.63; p < .0001]. Initial landing positions were further to the right in longer words [effect size = 3.1 characters; β = 0.54; SE = 0.03; p < .0001]. Also, longer words elicited a higher rate of refixation [effect size = 65%; β = 0.92; SE = 0.05; p < .0001], of regressive saccades [effect size = 21%; β= 0.22; SE = 0.06; p = .0003], and of a second reading pass [effect size = 36%; β= 0.31; SE = 0.03; p < .0001].

Similarly, higher-frequency words N (srWordFreq) came with a reduction in all measures of reading times, except for second pass reading time: [FirstFixDur: effect size = 15 ms; β = 2.71; SE = 1.21; p = .020, SingleFixDur: effect size = 16 ms; β = 2.83; SE = 1.43; p = .046, GazeDur: effect size = 58 ms; β = 10.35; SE = 2.36, p < .0001, RegrPathDur: effect size = 103 ms; β = 18.36; SE = 3.93, p < .0001, TotalTime: effect size = 118 ms; β = 21.12; SE = 4.12; p < .0001]. Also higher-frequency words had an initial landing position further into the word [effect size = 0.40 characters; β = 0.07; SE = 0.03; p = .006] and were less likely to be refixated

[effect size = 10%; β = 0.17; SE = 0.04; p < .0001]. Word N frequency did not elicit a significant effect on either the regression or second pass likelihood.

A greater length of word N-1 (sPrevLength) influenced the processing of word N too. Thus, longer preceding words elicited longer first fixation duration [effect size = 22 ms; b = 4.49; SE = 1.178; p < .0001], and single fixation duration on word N [effect size = 21 ms; β = 4.46; SE = 1.393; p < .0011]. Finally, longer preceding words came with a slightly lower likelihood of regression after the first reading pass on word N is complete [effect size = 4%; b = 0.14; SE = 0.05; p = .012]. Similarly, a higher frequency of word N 1 (srPrevFreq) elicited an easier processing of word N: slightly shorter first fixation duration [effect size = 9 ms; β = 2.23; SE = 1.14; p = .048], regression path duration [effect size = 40 ms; β = 10.38; SE = 3.92; p = .006], and total fixation time [effect size = 41 ms; β = 10.62; SE = 3.85; p = .004]. We take these effects to indicate that the lexical processing of a particularly difficult (long or low-frequency) word N-1 does not always complete before the saccade is launched to word N. As a result, final stages of the lexical processing of word N-1 spill over to word N and affect early reading time measures on that word. Additionally, a less frequent word N -1 may lead to a greater comprehension difficulty for the sentence as whole, thus incurring the observed more frequent and longer regressions from word N, as well as a processing penalty in the global processing measure of total fixation time. Finally, the increased foveal load of processing a longer or less frequent word N-1 may lead to less parafoveal preprocessing of word N, resulting in more effortful processing when word N is foveated (Henderson & Ferreira, 1990).

An increase in (residualized) frequency of word N + 1 (srNextFreq) came with a sizeable decrease in gaze duration [effect size = 42 ms; β = 6.71; SE = 2.07; p < .001] – in accord with findings of Kliegl et al. (2006) for German and of Pynte and Kennedy (2006) for English (though not for French) – and additionally in regression path duration [effect size = 63 ms; β = 10.11; SE = 3.51; p = .007]. More frequent words N + 1 also elicited a lower refixation likelihood [effect size: 9%; β = 0.13; SE = 0.03; p < .001] and regression likelihood [effect size: 6%; β = 0.10; SE = 0.05; p = .048]. Length of word N + 1 (sNextLength) did not elicit statistically reliable effects on eye-movement measures. Similarly, in Pynte and Kennedy's English data this measure was only found to reliably affect the number of fixations on word N and not any of the reading times. The nature of parafoveal-on-foveal effects (i.e., effects of properties of word N + 1 on the processing of word N) is subject to dispute in the literature, especially when such an effect is observed in a corpus study, rather than a tightly controlled experiment (cf. discussion and references in Rayner, 2009). Here we confine ourselves to reporting the empirical finding and leave experimental validation of the parafoveal-on-foveal effect of word N + 1 frequency to future research.

The sentence position in the experimental list (TrialNum) revealed effects of habituation, such that towards the end of the experiment participants showed a somewhat lower rate of refixation [effect size = 5%; β = 0.07; SE = 0.03; p = .04], regression [effect size = 7%; b = 0.09; SE = 0.03; p = .004], and especially second pass likelihood [effect size = 32%; b = 0.35; SE = 0.03; p < .0001]. Further into the experiment, the readers also showed slightly longer first fixation duration [effect size = 8 ms; β = 2.26; SE = 0.95; p = .01]. This finding is possibly explained by the well-established observation that single fixations on a word are generally longer than the first of multiple fixations (Rayner, Sereno, & Raney, 1996; Schroyens, Vitu, Brysbaert, & d'Ydewalle 1999; Vitu & O'Regan, 1995). Since the readers

were less likely to refixate words as the experiment progressed (see above), a larger percent of first fixations on words were single fixations, and hence the inflation in first fixation duration is as expected. Towards the end of the experiment, we observed substantially shorter regression path duration [effect size = 25 ms; β = 7.03; SE = 2.60; p = .007], second pass duration [effect size = 75 ms; β = 20.82; SE = 4.55; p < .0001], and total fixation time [effect size = 111 ms; β = 30.75; SE = 3.62; p < .0001]. The magnitude of the list effects on late eye-movement measures strongly suggest that they may confound results of experiments that do not control these effects statistically or experimentally.

Relative word position in the sentence (RelPos) elicited strong (linear or non-linear) effects across the entire eye movement record. Generally, there was an increase in reading times towards the end of the sentence: in FirstFixDur [linear: β = 9.39; SE = 1.42; quadratic: β = 1.22; SE = 0.15; p < .0001], SingleFixDur [β = 7.56; SE = 1.44, p < .0001], GazeDur [linear: β = 0.74; SE = 0.22; quadratic: β = 9.09; SE = 2.12; ps < .0001] and TotalTime [linear: β = 12.75; SE = 3.51; quadratic: β = 23.89; SE = 2.14; ps < .0001]. As the readers progressed to the right, there was also a greater likelihood of regression [β = 0.34; SE = 0.04; p < .0001], as well as lower likelihood of refixation [linear: β = 0.21; SE = 0.04; p < .0001; quadratic: β = 0.30; SE = 0.03; p < .0001] and of the second reading pass [β = 0.17; SE = 0.03; p < .0001]; see Tables 7–16 in Supplementary materials 1. These findings are consistent with earlier descriptions of the start-up and wrap-up inflation of reading times (Kuperman et al., 2010; Warren et al., 2009). The greater likelihood of regressive saccades as a function of the rightward word position is expected too, as the fragment of the sentence to regress to becomes increasingly larger and arguably more complex syntactically as the readers move towards the end of the sentence.

Initial landing position on word N (FirstFixPos) elicited a strong non-linear effect (an inverse-U shaped parabola) in all models in which it was included as a predictor: FirstFixDur [linear β = 4.42; SE = 1.71; p < .0001; quadratic = 4.75; SE = 1.80; p < .0001], SingleFixDur [linear: β = 5.66; SE = 1.21; p < .0001; quadratic = 4.15; SE = 0.80; p < .0001], RefixProb [linear β = 0.61; SE = 0.03; quadratic β = 0.26; SE = 0.03; both ps < .0001], and GazeDur [linear: β = 14.67; SE = 3.21; quadratic: β = 2.70; SE = 0.92; ps ps < .0001]; see Tables 8–13 in Supplementary materials 1. These effects are consistent with the Inverted Optimal Viewing Position (IOVP) effect in which initial fixations landing at the optimal viewing position – near the center of the word – elicit longest reading times and the lowest likelihood of refixation, while initial fixations close to the word boundaries come with shorter fixation durations and a higher likelihood of refixation (Vitu, McConkie, Kerr, & O'Regan, 2001; Vitu, Lancelin, & d'Unienville, 2007). These observations were first made using isolated word reading paradigms. Subsequent studies of continuous text reading confirmed the refixation component of the IOVP effect, but not the processing costs associated with a distance from the optimal position (Rayner et al., 1996; Vitu, O'Regan, & Mittau, 1990). Our data show evidence for this second aspect of the IOVP effect as well, suggesting that temporal processing costs related to the position of the initial fixation can be observed even in sentence reading.

The effect of syntactic type of the sentence (Type) was such that words in complex sentences were, on average, read longer than ones in syntactically simple sentences: the effect emerges in all but the earliest eye-movement measures, such as initial landing position, as well as single and first fixation duration; see Tables 7–16 (Supplementary

materials 1) for specifications of effects. Also, words in relatively complex sentences elicited a higher likelihood of refixation and regression, as well as a longer regression path duration.[4]

Modeling individual skill measures

The next step of the analysis involved augmenting the baseline models fitted to the dependent eye-movement measures with three additional participant-specific fixed effects: sex, age and score on one of the tests described in Appendix A and Table 1. Participant age and sex did not show consistent significant results in any of our models, and are not reported further. A separate model was fitted to gauge the influence of each test on the given eye-movement measure. In addition to evaluating the separate influence of each test, this procedure precluded inaccurate estimates of standard error for the regression coefficients that may arise due to the high collinearity of many of our tests. In order to avoid combinatorial explosion in the number of statistical models, we did not consider the interactions between individual tests. Table 4, Column 2 summarizes labels of skill measures that reached statistical significance at the 0.05 level in baseline models, as well as the polarity of their regression coefficients.5 Table 17 in Supplementary materials 2 also provides specifications of estimated effects for all skill measures that reached statistical significance (at the 0.05 threshold) when added to the baseline model for the respective eye-movement measure. The specifications for linear mixed models in Table 17 (Supplementary materials 2) include the regression coefficient for the standardized skill measure, the standard error and the t-value (an absolute t-value greater than 1.96 is roughly equivalent to $p < .05$). For logistic regression mixed models fitted to binary variables we report the regression coefficient for the standardized skill measure, the standard error, the Wald's z-value and the corresponding p-value.

As Table 4 Column 2 illustrates, many tests emerged as significant predictors of eye-movement measures at this stage of the statistical modeling process. Yet only some of the tests showed predictivity across the entire eye movement record: these include the rapid letter and digit naming tests (rln/rdn) and word identification test (wid). Moreover, a large number of tests were reliably predictive of earlier eye-movement measures (e.g., elision, blendw, blendnw, segmentw, segmentnw, phonrev, span, stan, piatl, and piatr), while the effects of simple-memory tests (digital memory span and non-word repetition) were confined to late measures, such as regression path duration, second pass likelihood and duration, and total fixation time. Perhaps the only surprising effect revealed above and in Table 4 was that relatively skilled readers had a higher likelihood of second pass reading than less skilled ones. Possibly, better readers invest more effort to get the right

[4] *To additionally account for the effect of sentence type on words depending on their syntactic role in the sentence, we considered the interaction between sentence type and the word's relative position. Given that sentences within each type are of a similar structure and a similar length, the word's relative position in a sentence is – to a first approximation – an index of its syntactic function. Sentence type interacted significantly with (linear or non-linear functions of) the word's relative position in all models for durational eye-movement measures. We opted, however, for reporting a simpler set of baseline models without this and several other tested interactions for two reasons. First, the inclusion of this interaction did not appreciably affect the estimates of model coefficients nor inferential statistics for our critical predictors, i.e. individual scores in tests of verbal ability, and second, there is no clear basis in earlier corpus studies with undergraduate participants against which an interaction we observe here can be compared. Outcomes of baseline models with the interaction are available from the authors on request.*

interpretation which might imply more rereading, especially because many of our sentences had a fairly complex syntactic structure. We leave this issue for future investigation.

We also observed that some of the tests that the models estimate as reliable co-determiners of eye-movements are in fact only predictive for a higher-scoring part of the population of participants, rather than for the population as a whole.

As an example, we consider the non-word repetition (nwrep) test, which reveals a strong correlation with gaze duration for the top half of the scorers, and no predictivity for the bottom half (Fig. 1; panel 1). This implies that in a cohort with better verbal skills than the one considered in the present study, these tests may be predictive of behavioral patterns across the entire population, however, they are not reliable for describing the behavior of the present cohort. There were also tests that correlated with eye movement measures across the entire participant population, see panels 2–4 in Fig. 1 for effects of wid, rln and piatr on gaze duration: These are of particular interest for our present purposes because these same tests were determined via procedures described below to predict unique variance in a broad range of eye-movement measures. Further treatment of the relationship between skill tests and their predictivity for various population subgroups is beyond the scope of this study.

Independence of tests as predictors of eye-movement measures

Our next step was to establish which tests that showed predictivity of eye-movement behavior would make unique contributions to explaining variance in data, when pitted against text variables and other predictive tests. This is particularly important given the strong intercorrelations between participants' performance in various tests (see Table 5). For instance, skill measure A may show a significant effect on the dependent variable Y by virtue of its high correlation with skill measure B that is also a significant predictor of Y, yet it may explain no unique variance in the model over and above test B.

To test the statistical independence of our skill measures, we applied two different methods of model selection: mediation analysis (MacKinnon, 1994) and the best-subset model selection technique that identifies models with the highest value of the selected goodness-of-fit criterion using several search algorithms. We provide descriptions of both methods of model selection in Appendix C. The outcomes of the model selection procedures for main effects of test scores on eye-movement measures are reported as Columns 3–4 in Table 4.

As Table 4 demonstrates, outcomes of the two methods showed remarkable convergence. Columns 3–4 of Table 4 highlight the fact that only a fraction of tests of verbal skills emerge as independent predictors of eye-movement measures. Importantly, visual inspection of correlation plots confirmed that effects of the tests that were identified as independent were predictive across the entire participant population, rather than its higher-scoring subrange (see, for example, panels 2–4 in Fig. 1). Moreover, these outcomes allow for distinguishing between the tests that only affect selected stages of visual comprehension of sentences and those that retain predictivity throughout the entire time-course of sentence processing. Across the methods, three tests surfaced as pervasive co-determiners of the eye-movement record: the strongly correlated rapid letter naming (rln) and rapid digit naming (rdn) tests, and the word identification (wid) test. More specifically, higher scorers in rapid naming tests (participants that showed shorter digit or letter naming latencies)

demonstrated a processing advantage in virtually every eye-movement measure (though in some measures rapid automatized naming scores were not independently predictive of dependent variables, see below). They showed reduced first fixation duration [rln: effect size = 33 ms; β = 7.512; SE = 3.014; p = .036], reduced single fixation duration [rln: effect size = 49 ms; β = 10.282; SE = 3.731; p = .013], a lower likelihood of refixation [rln: effect size = 22%; β = 0.260; SE = 0.091; p = .004], reduced gaze duration [rln: effect size = 91 ms; β = 20.076; SE = 5.544; p < .001], a lower likelihood of a regression [rln: effect size = 15%; β = 0.305; SE = 0.102; p = .003], reduced regression path duration [rln: effect size = 202 ms; β = 44.916; SE = 6.814; p < .001], reduced second pass duration [rln: effect size = 134 ms; β = 29.675; SE = 8.797; p = .004], and reduced total fixation time [rln: effect size = 124 ms; β = 27.670; SE = 11.081; p = .016]. The effects of the rapid digit naming (rdn) test on eye movement measures were similar in size and emerge in the same eye-movement measures (see Table 4 Column 2), except for initial landing position where a better performance in rln predicted the fixation position further to the right [rln: effect size = 0.4 characters; β = 0.094; SE = 0.041; p = .033], while rdn does not show a reliable effect. The rapid digit/letter naming tests do not reach statistical significance as predictors of the likelihood of second pass regression.

Higher scorers in the word identification test showed superior performance in every eye-movement measure (except for regression likelihood) as compared to lower scorers. Again, better performance in the word identification task came with the initial landing position further to the right [effect size = 0.3 characters; β = 0.086; SE = 0.040; p = .033], reduced first fixation duration [effect size = 52 ms; β = 13.353; SE = 2.646; p < .001], reduced single fixation duration [effect size = 71 ms; β = 17.725; SE = 3.250; p < .001], a lower likelihood of refixation [effect size = 21%; β = 0.392; SE = 0.082; p < .001], as well as reduced gaze duration [effect size = 124 ms; β = 31.108; SE = 4.625; p < .001], reduced regression path −2 −1 duration [effect size = 178 ms; β = 44.623; SE = 6.692; p < .001], reduced second pass duration [effect size = 105 ms; β = 26.256; SE = 8.551; p < .001], and reduced total fixation time [effect size = 86 ms; β = 21.449; SE = 11.057; p = .042]. Readers with higher word identification scores were also more likely to have a second reading pass on word N [effect size = 23%; β = 0.236; SE = 0.099; p = .017].

Outcomes of model selection techniques confirmed that rapid letter and digit naming scores were independent predictors of measures in both the first reading pass (initial landing position, refixation probability), as well as in measures associated with second pass reading and total word processing effort (regression probability, regression path time, second pass time, and total time). Similarly, individual performance in the word identification test pervasively predicted behavior in first reading pass (initial landing position, first and single fixation duration, gaze duration), as well as in later measures (regression path duration, second pass time).

In addition to these pervasive predictors of eye-movement measures, four tests showed independent predictivity for selected measures. The two versions of the Peabody Individual Achievement Test (listening: piatl and reading: piatr) showed significant effects on all eye-movement measures associated with the first reading pass, with benefits for higher scorers as follows: initial landing position further into the word [piatr (confirmed by the best-subset selection though not by mediation analysis): effect size = 0.3 characters; β = 0.090; SE = 0.041; p = .028], as well as reduced first fixation duration [piatr: effect size = 43 ms; β =

13.143; SE = 2.734; p < .001], single fixation duration [piatr: effect size = 57 ms; β = 17.461; SE = 3.313; p < .001], and gaze duration [piatr: effect size = 93 ms; β = 28.463; SE = 5.017; p < .001]. Unsurprisingly, effects of the listening version of the Peabody Individual Achievement Test comprehension test (piatl) were very similar to those of the reading version (piatr). For most dependent variables related to the first pass reading, either the listening or the reading version of the test came out as independent predictors (see Columns 3–4 in Table 4).

In contrast to the early influence of Peabody Individual Achievement Test comprehension tests, the digit memory span (digmem) test and the non-word repetition (nwrep) test only exerted influences on eye-movement measures after the first reading pass. Thus, higher scores in the digit memory span task (digmem) elicited significant effects on second pass duration [effect size = 92 ms; β = 20.726; SE = 8.683; p < .001] and total fixation time [effect size = 162 ms; β = 35.032; SE = 10.672; p < .001]. While the non-word repetition test came with reduced second pass duration [effect size = 105 ms; β = 26.423; SE = 8.206; p < .001] and total fixation time [effect size = 108 ms; β = 27.140; SE = 11.187; p = .013], its independent predictivity was only confirmed by the mediation analysis for total time fixation.

Finally, a few additional tests that survived our model selection techniques showed significant effects on a haphazard collection of eye-movement measures. For example, superior performance in the sentence span test (span) came with a rightward shift in the initial landing position on word N [effect size = 0.5 character; β = 0.095; SE = 0.040; p = .017], and refixation probability [effect size = 7%; β = 0.358; SE = 0.085; p < .001]. Also a higher score in the phoneme reversal test (phonrev) came with shorter first fixation duration [effect size = 49 ms; β = 13.021; SE = 2.714; p < .001] and second pass probability.

Interactions of test scores with other predictors of eye movement measures

The influence of individual differences may emerge not just in the main effects of test scores, but also in the differential sensitivity of participants to other predictors of eye movement control. To explore this possibility, we fitted a set of models to each of the dependent variables. In each model, we allowed one of the test scores to interact with exactly one of the following word properties: frequency or length of word N. We selected these variables out of the set of predictors because they serve as the strongest lexical predictors of foveal visual processing (Rayner, 1998, 2009). We report in Table 5 the outcomes of the mediation analyses (MacKinnon, Fairchild, & Fritz, 2007) that point at significant interactions that are not fully mediated by other interactions (see the description of the procedure in Appendix C), as well as the polarity of regression coefficients for the interaction terms in the statistical models. Since the best-subset model selection cannot be presently applied to regression models with interaction terms (see Appendix C), we only report the results of mediation analysis: we remind the reader that for main effects this analytical technique converged almost perfectly with the best-subset model selection. Furthermore, in consideration of space, we restrict our discussion in Table 5 and below in this section to interactions with spatial and temporal eye movement measures (rather than likelihoods of refixation, regression or second pass) as they provided most insight into the time-course of

the reading process. For completeness, we report the outcomes of statistical tests for all dependent variables and all significant ($p < .05$) interactions between test scores and selected lexical properties in Tables 18 and 19, Supplementary materials 2.

Similar to our observations of main effects of skill measures, only a few tests proved to consistently enter into significant interactions with length and frequency of words across the eye-movement record. As expected, these were also the tests that elicited ubiquitous main effects on the eye-movement record (see ''Modeling individual skill measures'' and Table 4): namely, rapid digit/letter naming, word identification, Peabody Individual Achievement Test (PIAT, reading and listening versions) and occasionally, phoneme reversal, non-word repetition and digit memory span tasks. This finding supports our observation above that a small subset of tests is predictive of the difficulty of word recognition across the entire time-course of sentence reading, while others are likely to show effects by virtue of their strong correlations with the tests from that subset.

Interactions with the length of word N

The model for initial landing position revealed strong interactions of a number of test scores with the length of word N (cf. Table 5, Column 2 and Table 18, Supplementary materials 2). The interactions indicate that higher scorers in several tests land their initial saccade further into word N than the lower scorers do. Moreover, the skill-related contrast between mean landing positions on word N increases with the length of word N: e.g., in 4-character words, there is only a 0.2 character contrast in mean initial landing positions between participants with the lowest and the highest score in the word identification task, but in 8-character words the contrast is about 0.6 characters, see Fig. 2 for mean initial fixation positions per word length in the top vs. bottom fourth of word identification scorers (the split into fourths is for illustrative purposes only; the statistical models are fitted to the entire population of participants).

The typical saccadic behavior of unimpaired readers of English is such that they tend to overshoot (fixate to the right of) the center of words up to 5 characters, land at the center of the 6–7 character words, and undershoot (fixate to the left of) words longer than 7 characters (McConkie, Kerr, Reddix, & Zola, 1988): arguably, this behavior reflects errors in saccade-planning aimed at the word center as the optimal viewing position (cf. references in Rayner, 2009). Fig. 2 reveals that our lowest scorers systematically undershot the words and fixated near the beginning of the word (at most 3 characters to the right for words up to 9 characters), except for 3 character-long words where they fixated near the center.6 Also, while word length modulated initial fixation position for both the lowest and highest scorers, the underperforming group showed limited sensitivity to word length in their distribution of fixation positions. This interaction of word length with skill (word identification and the reading version of the Peabody Individual Achievement test) was also observed when the dependent variable was transformed from initial landing position (in characters) to the ratio of initial landing position and word length, which characterizes the proportion of the way through the word that the initial fixation falls.

The reduced effect of word length on initial landing position as well as the leftward shift in initial landing position for low scoring participants is consistent with data on German dyslexics reported by Hawelka et al. (2010), who interpreted their results as support for a sublexical processing strategy. This was further supported by their observation that the

average size of forward saccades is 4.4 for the dyslexic group vs. 7.4 for their unimpaired control group. This is consistent with a large body of research attesting to smaller saccade sizes and increased refixations in poor, dyslexic, and beginning readers (Adler-Grinberg & Stark, 1978; Eden et al., 1994; Elterman, Abel, Daroff, Dell'Osso, & Bornstein, 1980; Lefton et al., 1979; Martos & Vila, 1990). A similar finding was also recently reported by Rayner et al. (2010), who observed smaller saccade sizes in slow (but unimpaired) readers as compared to fast readers. We likewise analyzed the size (in characters) of forward saccades as a function of skill, and observed a similar leftward shift in the saccade size distributions for less proficient readers: see Fig. 3.

The mean saccade sizes for the bottom, middle and top thirds of scorers in the word identification test were 6.3 (SD = 2.7), 6.9 (SD = 2.8), and 7.4 (SD = 2.7) characters respectively, the latter number being consistent with the estimate for unimpaired readers of German in Hawelka et al. (2010) and with the estimate of 7–9 characters as an average saccade size for English readers in Rayner (1998). We also note the tendency of poorer readers to have more fixations on a word, as is apparent in the negative correlation of test scores with refixation probability, see Table 4

The finding of shorter saccade sizes, leftward shifted initial landing position, and higher numbers of refixations on a word for low-skilled readers points to more effortful processing during word recognition. One possible explanation for this result focuses on ocular-motor control as the source of difficulty; shifted initial landing positions and (perhaps) increased refixations may arise simply because poor readers as a group make shorter saccades, leaving them no option but to fixate toward the beginning of words. While there have been suggestions in the literature that eye-movements are a causative factor in reading disability, the bulk of the evidence does not support this view. In his thorough review of the issue, Rayner (1998) concludes that the most likely explanation for these effects is that eye-movements are merely an index of an underlying linguistic processing deficit (p. 395; see also the discussion in Hutzler, Kronbichler, Jacobs, & Wimmer, 2006).7 We suggest therefore that a more plausible explanation of the data is that the eye-movement strategy that a reader adopts is simply an affordance of the overall quality of his or her lexical representations. Thus, a skilled reader who has full-form lexical representations available for the vast majority of words can use this information to guide eye-movements, employing a reading strategy which targets the eyes towards the center of words because this is the optimal viewing position for full-form word recognition. In contrast, readers who are frequently forced to rely on sublexical processing to identify words will adopt an eye-movement strategy that targets the beginning of a word (regardless of its length) since this is where grapheme-phoneme decoding must begin. Such a strategy will also produce multiple fixations on words because they are being processed in a piecemeal fashion. This suggestion is consistent with a number of other studies that have shown changes in eye-movement behavior as a function of the match between word difficulty and reader ability. For example, Pirozzolo and Rayner (1978) found that dyslexic readers who were given text appropriate for their reading level (in contrast to their age) showed eye movements that were like those of normal readers at that level (see also Olson, Kliegl, & Davidson, 1983); but when they were given texts appropriate to their age (but which were too difficult for them), their eye movement characteristics showed the usual indices of reading difficulty (shorter saccades, longer and more fixations, etc.) Similarly, Rayner (1986) showed that normal

children's eye movements (saccade lengths, fixation durations, and the size of the perceptual span) could be made to look characteristic of dyslexic readers' eye-movements when they were given text that was too difficult for them.

Regarding durational eye-movement measures, interactions of test scores with the length of word N (sWordLength) showed a consistent pattern across a range of tests for virtually all measures (except first fixation duration) with the same characteristics as discussed above. Higher scorers demonstrated much weaker positive correlations of sWordLength with respective eye-movement measures than lower scorers did; see Fig. 4 for mean gaze durations per word N length, computed for the bottom and top fourths of word identification scores.

The average contrast between a 4-character word and an 8-character word is estimated by the model at roughly 90 ms in gaze duration and 160 ms in total fixation time for participants with the lowest performance in the word identification task. The same contrasts amount to only 40 ms in gaze duration and 90 ms in total fixation time for best performers in the wid task. Evidently, superior verbal proficiency correlated with an advantage in reading all words, while the advantage increased when longer words are read. This is consistent with the ability of good readers to identify familiar words via association with previously stored full-form representations, producing only marginal effects of word length. Again, a much stronger effect of word length for poor readers is in line with the intuition that they are engaged in a sublexical processing strategy, which by its nature makes the number of letters a major codeterminer of processing effort.

Fig. 4 also hints at the difference in the reader's perceptual span as a function of reading skill. Poor readers showed a substantial increase in average gaze durations for words longer than 6 letters as compared to a more flat curve for words up to 6 letters. At the same time, for skilled readers a substantial increase in average gaze durations was associated with words that were longer than 9 letters (longer words typically elicit multiple fixations in skilled readers due to visual acuity constraints). The difference is consistent with the findings of smaller perceptual spans in beginning readers (Häikiö et al., 2009; Rayner, 1986), dyslexic readers (Rayner et al., 1989) and slow readers (Rayner et al., 2010) as compared to fast readers. Gauging the correlation between selected verbal skills and perceptual span using, for instance, the moving window technique (McConkie & Rayner, 1975; Rayner & Bertera, 1979) is a topic for future research.

Interactions with lexical frequency of word N

Lexical frequency of word N (residualized on length of word N, srWordFreq) entered into significant interactions with individual performance in the rapid naming task, word identification task and the non-word repetition test (cf. Table 5, Column 3 and Table 19, Supplementary materials 2). Across all durational eye-movement measures, the negative correlation of word frequency with the dependent variable was much weaker for higher scorers than it was for lower scorers in those tests, see Fig. 5 for the mean gaze durations per word N frequency, computed for the bottom and top fourths of word identification scores.

The advantage that participants with the lowest scores in the word identification task gained when reading the word with the highest residualized frequency in our data set as compared to the word with the lowest residualized frequency, was estimated by the

statistical model at about 110 ms in gaze duration and 230 ms in total fixation time. For top word identification scorers the estimated facilitation of word frequency was reduced to a mere 15 ms in gaze duration and 30 ms in total fixation time. Also the contrast in reading speed between the highest and lowest word identification scorers was estimated at about 50 ms in gaze duration and 80 ms in total fixation time for the most frequent word and at 170 ms in gaze duration and 250 ms in total fixation time for the least frequent word. Interactions of word frequency with rapid letter/digit naming tasks follows the same pattern.

This is consistent with a number of studies reporting that poor skilled readers show a greater benefit from higher-frequency words as compared to better readers in eye tracking measures (e.g., Ashby et al., 2005; Hawelka et al., 2010) and is related to the well-established finding that novel or less familiar words elicit longer reading times than relatively familiar words (Chaffin et al., 2001; Williams & Morris, 2004). The interaction has also been observed in measures of word naming, lexical decision, and brain activations as measured by BOLD responses (e.g., Frederiksen, 1978; Perfetti & Hogaboam, 1975; Pugh et al., 2008; Shaywitz et al., 2003). While a number of accounts for this effect are possible, the most widely accepted explanation is that frequency has its effect through facilitating an experience-dependent shift from componential to full-form processing (see ''Introduction''). Thus, readers with more education (Tainturier, Tremblay, & Lecours, 1992) and greater print exposure (Chateau & Jared, 2000) have been shown to be less sensitive to word frequency (as well as other lexical factors such as number of letters (Butler & Hains, 1979) and orthographic neighborhood size (Chateau & Jared, 2000)). Yet, an important limiting factor for the trajectory of this shift from componential-to-full-form recognition is a readers' ability to relate the orthographic form of a word to its meaning. Readers who have mastered the grapheme-to-phoneme correspondences that enable this decoding will need fewer learning exposures to achieve a level of automatic recognition, so that even low-frequency words can approach asympototic reading speeds fairly easily. Readers with poor phonological skills, however, will have much more difficulty acquiring unitized representations, even for high-frequency words with which they will have more learning opportunities. Low frequency words will be particularly problematic because their poor phonological skills will hinder their ability to make efficient use of the limited learning opportunities they will receive with these words. Hence, componential processing will persist longer for these words, which in terms of eye movement measures will require more fixations and longer fixation durations, which is consistent with the results observed here.

Summary of major results

To recap the findings above, we find evidence that multiple tests of verbal skills showed reliable effects on the eye-movement record (listed in Table 4, Column 3–4). Our model selection analyses were an important means of identifying a subset of skill measures that explained variance over and above other tests. Two types of tests – rapid naming tasks (rdn/rln) and word identification task (wid) – elicited ubiquitous and strong effects across the eye movement record, that is, in at least one measure associated with the first and with the second reading pass. Specifically, a superior performance in any of these tests translated into shorter inspection times in the first and second reading pass, as well as in cumulative durational measures; and a lower likelihood of a regressive saccade, a refixation in the first reading pass or a re-inspection of the word in the second reading pass.

The effects of individual variability on eye-movement measures were monumental in the examined cohort of participants compared to typical word-level predictors. To illustrate this point, Fig. 6 compares the magnitudes of main effects that skill measures and selected lexical predictors exert. The figure plots absolute values of regression coefficients for standardized scores in the rapid letter naming task (rln), word identification task (wid) and two lexical properties of word N that are acknowledged to be exceptionally strong predictors of eye-movements, namely, length and (residualized) frequency. Each coefficient indicates an (absolute value of the) amount of change in the dependent variable in milliseconds per one unit of standard deviation of the respective predictor: for comparability, we only produced plots for the durational eye movement measures.

Regression coefficients associated with rapid naming and word identification are both greater in their absolute values than those for word N length and frequency in early measures, such as first and single fixation duration, and they are second only to the effect of word N length in the rest of the durational measures. Furthermore, the contrast that extreme scorers show in the average word's total fixation time for a word of a median length (6) is the same as the contrast in total fixation time between a 3-character word (e.g., cat) and a 10-character word (e.g., university) of similar lexical frequencies. That is, the impact that individual abilities have on the eye-movement record is on par with or exceeds the effects of the strongest ''benchmark'' determiners of visual control in reading.

To further stress the variability indexed by these individual differences measures, we note that the estimated contrasts between a participant with the worst performance in the rapid letter naming task and one with the best performance (e.g., 49 ms in single fixation duration; 91 ms in gaze duration; and 124 ms in total fixation time) amount to 20–40% of the average values of respective reading times (listed in Table 2). Also, best-performing scorers in the rapid letter naming task differ from worst performing scorers by a factor of 2 in the likelihood of refixation or in the likelihood of regression: the participant with the lowest score is estimated to regress from about 1 word out of 4 (26%), while the best scorer will only regress from one word out of 8 (12%), which is well within the usual regression rate for adult readers estimated at 10– 15% (Rayner, 1998). The inclusion of individual scores in the rapid letter naming task in the regression models for durational dependent variables reduced the standard deviation of participant as a random effect by 7–15%, while the inclusion of word identification task scores led to a reduction of 5–23%. Finally, the amount of variance explained by individual word identification scores in the ordinary regression model fitted to by-participant mean gaze durations reached 32% for single fixation duration (10% for rapid letter naming); 40% for gaze duration (20% for rapid letter naming); and 6% for total fixation time (8% for rapid letter naming). Our previous discussion also illustrated the large main effects of individual differences scores; likewise their influence as it emerges through interactions with text-level variables is tremendous, often varying the effect size of word length on eye-movement measures by a factor of 3 and the effect of word frequency by a factor of 4, as estimated by regression models.

Our demonstration of the influence of individual variability on reading behavior is particularly significant in light of the import placed on accounting for variance due to text-level properties of words in both experimental investigations and computational models of eye movement control. The observation that participant-level variance can dwarf such widely recognized predictors as word length and word frequency, or drastically modulate

their effects, points to a need for a thorough screening of participants in an experiment, even when recruited from a seemingly homogenous pool in terms of educational level (see ''Introduction'' for references attesting to variability in the university undergraduate population). This observation also calls for developing experimental and statistical methods through which this variance can be better controlled. This is especially true for the many fields of psycholinguistic study that target subtle empirical effects. We further consider the broader implications of these findings in ''General discussion''.

While the rapid naming and word identification tests showed consistent predictivity throughout both early and late measures, the reading and listening versions of the Peabody Individual Achievement Test (piatr/piatl) showed predictivity confined to the early stages of visual processing. We restrict our discussion of this test, however, because its construct-validity has recently been challenged (Keenan, Betjemann, & Olson, 2008). Although the Peabody Individual Achievement Test has been widely used as a test of reading achievement since its introduction in 1970, when it was one of the first norm-referenced tests to include subtests for Reading Comprehension, factor analyses and hierarchical regressions have suggested that this test is more highly associated with word decoding than with listening comprehension for lower ability readers. For higher-skilled readers, the Peabody Individual Achievement Test appears to assess both comprehension and decoding ability. Our observation that this test is predictive of early eye-movement measures seems consistent with its association with word decoding, as opposed to comprehension, which would likely implicate later eye-movement measures as well.

Several other measures that were not ruled out by our mediation analysis affected only a small subset of eye movement measures: sentence span (initial landing position), and phoneme reversal (first fixation and single fixation duration), digit memory (total fixation time) and non-word repetition (regression path duration and total fixation time). As our interest is in identifying consistently reliable predictors of eye-movement behavior, we refrain from further discussion of these tests.

General discussion

The primary contribution of this report is to reveal the considerable effect of participant variables on the eye movement record. We have identified five specific skills, simple memory (digit span memory and non-word repetition), rapid naming (letters and digits), word identification (wid), phoneme reversal (phonrev) and broader comprehension (Peabody Individual Achievement Test, reading and listening versions), which significantly predicted eye movement behavior across the entire range of measures for the population of 71 non-college-bound young adults considered here. Although subsequent analyses revealed that unique variance was consistently explained by only rapid naming and word identification, we consider it noteworthy that tests with good predictivity across the full range of ability in our sample stem from several major components of reading ability that have been identified in the literature (Olson, Forsberg, Wise, & Rack, 1994). This suggests that the eye-movement record is sensitive to the entire spectrum of verbal skills, rather than any single ability. Further, our analyses suggest that among tests that are argued to tap similar skills, only some emerge as signifi- cant co-determiners of reading behavior in the current population: e.g., word identification (wid) rather than the word attack test (watt) for

word decoding; Peabody Individual Achievement Test (piatr) rather than Stanford Fast Reading test (stan) for reading comprehension; rapid letter/digit naming (rln/rdn) rather than rapid object/color naming (ron/rcn) for rapid naming, etc. Hence, the major theoretical question yet to be answered involves understanding what it is about the ability measured in these tests that is relevant for eye-movement control. Here, we consider this issue for the two most robust predictors of unique variance in the eye-movement record: word identification and rapid naming.

The emergence of word identification (wid), a measure of word decoding skill, as a pervasive predictor of eye movements follows from the fundamental role of decoding in the reading task. While it is clear that the goal of reading is comprehension, a complex process that requires the integration of multiple levels of linguistic and real-world knowledge (Kintsch, 1998), the written modality imposes an additional constraint on comprehension processes that is not present in the oral domain; namely, the need to unlock the phonological form of individual words which have been encoded within an orthographic writing system. In some cases, orthographies may have quite arbitrary and inconsistent means for encoding the sounds of oral language. For example, languages like English, which are considered to have ''deep orthographies'' present a beginning reader with a formidable task: readers must learn that a single orthographic pattern such as ''int'' has multiple pronunciations (e.g., pint, mint), and, to make matters worse, they must also master cases where the same pronunciation is encoded by multiple orthographic patterns (e.g., mile, aisle, dial, guile, style). In contrast, languages like German, Italian, Serbian or Finnish have more predictable orthographies: the sound sequence /ajl/ will always be encoded via the same pattern of vowels and consonants, and those patterns will always translate into the same oral pronunciation. Thus, the primary task for a beginning reader is to learn the set of correspondence rules that map letters and letter-clusters to the phonemes and syllables that will enable him or her to decode orthographic forms, translating them into representations already learned during oral language acquisition. The large variability on the decoding measures in our data set (word identification ranging from the 4th to 18th grade (second year of graduate school) level and word attack ranging from the 2nd to 18th grade level) supports the intuition that this task is particularly difficult in a language like English, and is consistent with other findings indicating that these skills are often not mastered in college-aged readers (e.g., Cunningham et al., 1990; Shankweiler et al., 1996).

Readers with high decoding skill (as measured by one of our pervasive predictors, an individual's score in word identification task) have mastered the grapheme-to phoneme correspondence rules that—together with adequate experience—facilitates a shift away from sub lexical processing strategies. For these readers, the majority of words will be read via rapid, automatic activation of unitary lexical representations (Pugh & McCardle, 2009). In contrast, low decoding ability suggests poor-quality lexical representations, a consequence of a poorly tuned phonological coding system and limited experience with written language. Word recognition for these readers must rely on sub lexical processing, which is noisy by definition due to incomplete knowledge of grapheme-to-phoneme correspondence rules, and insufficient feedback arising from variable phonological forms, and shallow meaning representations (Perfetti, 2007). This situation has important implications for eye-movement control. Highly integrated representations are accessed very quickly, with simultaneous availability of their orthographic, phonologic, morpho-syntactic,

and semantic attributes. The eye movement patterns demonstrated by highly performing participants corroborate this observation, as such participants are less sensitive to factors such as word length and frequency, resulting in generally fast lexical access and less re-reading overall. In contrast, the eye-movement record of individuals with poor decoding skill (i.e., low scorers in the word identification task) displays clear indicators of effort, including initial landing positions that are more towards the beginning of a word, where sub lexical processing must begin, shorter saccades and greater sensitivity to word length, and increased fixations, especially on infrequently encountered words.

A particularly novel finding in the current study was the prominence of the rapid letter/digit naming tasks (rln/rdn), which arose as equally strong predictors (see Fig. 6) of eye movements as the word identification task. Rapid automatized naming (RAN) tasks have received an enormous amount of attention in the reading literature (see Wolf & Bowers, 1999; and the special issue of the Journal of Learning Disabilities, vol. 33(4), July/August 2000), however their relationship to eye-movements in natural reading has not previously been investigated. Below we discuss three potentially related hypotheses accounting for the substantial magnitude of RAN effects on the eye-movement record. First, RAN has been reported to account for independent variance in both phonological and orthographic processing tasks (e.g., Bowers, Golden, Kennedy, & Young, 1994; Bowers & Newby-Clark, 2002; Bowers & Wolf, 1993; Cutting & Denckla, 2001; Manis, Doi, & Bhadha, 2000). Sensitivity to phonological and/or orthographic components of reading ability makes RAN a particularly strong predictor of the sublexical decoding process. Indeed, Misra, Katzir, Wolf, and Poldrack (2004) provided neuroimaging evidence that the RAN letter naming task activated brain regions that have been implicated in phonological assembly procedures (Price, 2000; Pugh et al., 2000) and in regions associated with letter-specific processing (e.g., Polk et al., 2002). However, there was no apparent activation in the visual word-form area where activations occur when highly familiar words are accessed. Thus, performance in RAN tasks appears to be especially sensitive to the controlled sub lexical assembly processes that poorer readers employ for word identification, which, we have seen here, is also indexed in the eye-movement record.

Secondly, the rapid naming tasks were the only timed tasks in our battery. Since they produce a measure of response latency, RAN tasks have been interpreted as an index of general processing speed in cognitive tasks (e.g., Cutting & Denckla, 2001; Kail & Hall, 1994; Powell, Stainthorp, Stuart, Garwood, & Qunlan, 2007; Savage, Pillay, & Melidona, 2007; Wolf & Bowers, 1999). This speed is crucial for the efficient lexical access that is assumed by at least some models of eye-movement control (e.g., Reichle et al., 2003) as the engine that drives eye-movements forward during reading. Another link between the speed of performance in RAN tasks and the speed of lexical processing relates again to phonological and orthographic components of reading (as discussed above). Since the orthographic form must be decoded prior to accessing previously stored phonological and semantic representations, uptake of lexical (particularly semantic) information will necessarily be limited by how quickly the orthographic and phonological connections can be made. Even aside from issues related to lexical representations, research in the perceptual domain suggests that visual information reaches the visual cortex about 70 ms post-stimulus onset (Schmolesky et al., 1998), while auditory information reaches the auditory cortex about 30 ms following stimulus onset (Heil et al., 1999). The need to integrate visual and phonological

representations during reading means that these two sources of information must be appropriately synchronized, and any differences in the speed of processing visual or auditory information that a poor reader may have may jeopardize this synchrony (Breznitz, 2006; Carello, LeVasseur, & Schmidt, 2002; Denckla, 1985; Gladstone, Best, & Davidson, 1989; Nicolson & Fawcett, 1990; Pavlidis, 1985; Wolff, Michel, Ovrut, & Drake, 1990). Indeed, Breznitz and Misra (2003) observed that adult dyslexics had a greater time gap between the processing of orthographic and phonological information than did unimpaired readers, and that the size of this gap was the best predictor of word identification. Although it is not clear how this result transfers to our population of ''garden-variety'' poor readers, it is likely that, like dyslexics, our poor readers will have a greater proportion of low quality lexical representations than unimpaired readers. The fact that RAN tasks provide a means of estimating the speed with which readers may associate visual with phonological information may enable this task to provide an indicator of the gap between phonological vs. orthographic processing, which could affect individual performance in word recognition and associated eye-movements.

Finally, the predictive role of RAN may originate from a central component of the rapid naming task—namely, serial ocular-motor programming. Seriality is a defining feature of reading skill: readers must be able to disengage from one stimulus and move to another, rapidly programming saccades as the eyes move across the printed page. Since rapid naming tasks involve speeded serial visual inspection and subsequent naming of letters, digits, colors or objects arranged in a grid on a computer screen or paper, the oculomotor component of this task is very similar to that required in natural reading. The RAN task is the only major assessment of reading skill (both in our battery and in the broader collection of typically used assessments) that directly measures this attribute of reading behavior. Consequently, it is not surprising that this task surfaced as one of the strongest and most predictive tests in our battery.

To be clear, we are not suggesting here that ocular-motor programming in and of itself plays a role in determining reading ability, as our previous discussion (and footnote 6) indicates that evidence does not support this view. Rather, we suggest that the RAN task may be capturing important variance associated with the processing of rapidly occurring serial information, which requires the reader to repeatedly engage and disengage attention as the eyes move through the text. Although previous research investigating the RAN-reading relationship has not employed eye-movement measures (but see Jones, Obregon, Kelly, and Branigan (2008) for the single exception in a non-reading task), there have been a number of studies pointing to the importance of the task's serial format. For example, a number of studies that have compared naming of words in isolation vs. those in a serial list, have shown that it is the latter task that is more strongly associated with reading ability (e.g., Bowers & Swanson, 1991; Jones, Branigan, & Kelly, 2009; Perfetti, Finger, & Hogaboam, 1978; Stanovich, Feeman, & Cunningham, 1983; Wagner, Torgesen, & Rashotte, 1994; Wolff, Michel, & Ovrut, 1990). The fact that rapid naming tasks preserve the characteristic requirement of serial eye-movements during reading affords them the possibility of indexing the same eye-movement patterns that poor and dyslexic readers demonstrate in normal reading: namely these readers make longer fixations, have shorter saccades, and more fixations and regressions than normal readers (Biscaldi, Gezeck, & Stuhr, 1998; Olson et al.,

1983; see Rayner (1998) for a review). Consequently, even without a test of actual reading ability, this task may afford a measure of important correlates of reading skill.

Conclusions

An important contribution of the present study is to reveal the extent of individual variability that may be present in the eye-movement record. The cohort investigated here was reasonably matched in terms of age to that of populations that traditionally supply behavioral data to researchers, namely, undergraduate university students. This, together with the fact that a number of studies point to considerable variability in reading ability among undergraduate populations, at least in the US (e.g., Cunningham et al., 1990; Shankweiler et al., 1996) suggests that methods which enable researchers to account for individual variability in the eye-movement record will have an important impact on advancing our current understanding of eye-movement patterns in sentence reading. The magnitude of effects associated with our participant variables on the eye-movement record as compared with routinely controlled text-level variables such as word length and frequency, as well as their strong modulation of effects of text-level variables, makes a strong case for incorporating at least some indices of individual differences in eyetracking studies of reading. While our sample of non-college-bound young adults is relatively large (71 participants), it is understood that additional experimental replication is required before one can reach a definitive conclusion on which tests excel and which fail as predictors of eye-movements in reading. Nevertheless, our current results point very clearly to two such measures: word identification and rapid naming (see Appendix A for description). The fact that these tests, which are simple to administer and require less than 5 min each to complete, appeared to be robust predictors across the entire eye-movement record, makes them highly suitable for garnering data that will enable researchers to account for one of the strongest sources of variance in the data and improve predictions of statistical models.

Acknowledgments

This work was supported in part by the Rubicon grant of the Netherlands Organization for Scientific Research (NWO) to Victor Kuperman, and the following grants from the NIH National Institute of Child Health and Human Development: HD 058944 to Julie Van Dyke (PI), HD 056200 to Brian McElree (PI), HD 040353 to Donald Shankweiler (PI). Partial support also came from the NIH National Institute on Deafness and Other Communication Disorders via Grant DC 07548 to Carol Fowler (PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are grateful to David Braze for assistance with eye-tracking software and experiment set-up, and to Jessica Grittner and Kim Herard for administration and scoring of battery tests. The manuscript was greatly improved by the careful reading of Bernhard Angele, Clint Johns, Roger Levy, Keith Rayner, and an anonymous reviewer.

Appendix A

This appendix provides detailed descriptions of the 18 tests of verbal skills summarized in Table 1.

Phonological processing

A deficit in some aspect of phonological processing ability is viewed as a cause of the most common form of reading disability (e.g., Olson, Wise, Conners, & Rack, 1990; Olson et al., 1994; Shankweiler & Liberman, 1989; Wagner & Torgesen, 1987), which is characterized by difficulty in translating individual words into their phonetic components (i.e., decoding) in order to make contact with oral vocabulary knowledge. Confirmatory factor analysis (e.g., Wagner et al., 1997) suggests three related, but separable phonological abilities: (a) phonological awareness (i.e., the ability to break words into phonemes and synthesize words from discrete sounds); (b) phonological memory (i.e., the ability to encode and store information phonologically); and (c) rapid naming (i.e. the ability to efficiently retrieve phonological encodings). To test these abilities, we employ the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999), which includes multiple subtests in each of these three areas, summarized below.

Phonological awareness with words and nonwords

Elision

Participants are instructed to say ''bold''. Now say ''bold'' without saying the /b/. The correct answer is ''old''. In all cases, both the initial and resulting utterances are actual words. Twenty trials are given, with elisions occurring at any position within a word. All words are 1 or 2 syllables long. Reliability was a = .87; test duration is 3–4 min.

Blending words

Participants are asked ''What word do these sounds make?'' /nVm/ – /bEr/. Correct response = ''number''. All answers are words, ranging in length from 2 phonemes to 10, presented as an audio recording in order of increasing length. Twenty trials are given. Reliability was a = .84; test duration is 3–4 min.

Blending nonwords

Same procedure as above, although responses are nonwords. For example, participants are asked ''What madeup word do these sounds make?'' /mO/ – /tEb/. Correct response = ''motab''. Non-words containing between 3 and 8 phonemes are presented. Eighteen trials are given. Reliability was a = .81; test duration is 3–4 min.

Segmenting words

Participants are instructed ''I'm going to say a word and I want you to repeat it. Then say the word one sound at a time. Say each sound that you hear in the order that your heard it. For example, if I say ''me'', you would say the word ''me'' and then you would say m-ee.''

Twenty trials are given, beginning with ''no'' and ending with ''graduate''. Reliability was a = .88; test duration is 3–4 min.

Segmenting nonwords

Participants are instructed ''You are going to hear a made-up word. First I want you to repeat the made-up word, then say it one sound at a time. Say each sound in the order that you heard it. For example, if you hear ''jp'' you would repeat the word ''jp'' and then say /F/-/p/''. Twenty trials are given, beginning with 2-phoneme nonwords and ending with 8-phoneme non-words. Reliability was a = .88; test duration is 3–4 min.

Phonological awareness with memory load
Phoneme reversal

Like previous tests, this test was also taken from the CTOPP. Participants were required to say a non-word, then say the same non-word backwards. To complete the task, it was necessary to maintain the non-word in memory while re-ordering the phonemes to create the response, which always turned out to be a word. For example, in the first trial, participants said /ta/ and the transformed response was /at/. Eighteen trials were presented, ranging from one to two syllables, and increasing in the number of phonemes from 2 to 7. The final trial asked participants to say /zmitmus/, which they were supposed to transform into the word ''sometimes''. Reliability was a = .85.

Phonological memory: simple memory span
Digit span

Participants hear a list of numbers presented as an audio recording, and must say them back in the order that they were heard. Twenty-one trials are presented, beginning with lists of length 2 and ending with length 8. Reliability was a = .77; test duration is 3–4 min.

Non-word repetition

Participants are instructed ''You will hear some made up words. After you hear each made-up word, say the word exactly as you heard it. Recordings begin with 1 syllable non-words (e.g., /joop/) and end with 6-syllable non-words (e.g., /sha-bur-i-hu-voi-mush/''. Eighteen trials are given, with increasingly longer non-words. Reliability was a = .76; test duration is 3–4 min.

Rapid naming
Naming of letters, digits, colors, and objects

Four tests were presented, which we refer to as rln, rdn, rcn, and ron, differing only in the form of the objects to be named. Together, these are known as RAN (rapid automatized naming) tasks. In all tests, participants saw a 4 9 grid containing either letters (rln), digits (rdn), color patches (rcn), or pictures of objects (ron), and they were instructed to ''say the names of the letters/digits/colors/objects on this page as fast as you can.'' Responses were timed by the experimenter beginning from the first name uttered to the pronunciation of the last form on the grid. The number of errors was recorded as well as the response time. Two

trials were conducted for each participant in each version of the test, unless participants made more than four errors on the first trial. Reliability was a = .82, .86, .75, .85 for letters, digits, colors, and objects, respectively. Test duration for each version is 2–3 min.

Word and non-word decoding

Both word and non-word reading were assessed using subsets of the Woodcock-Johnson-III Tests of Achievement (Woodcock, McGrew, & Mather, 2001).

Word identification

Word reading was assessed via the Word Identification subtest, Form A (WordID). Participants read aloud a list of individual words, divided into sets of 6–8 words, each set increasing in difficulty. No contextual support is provided, making this a test of decoding skill; participants were not expected to know the meaning of the words, but simply to pronounce them correctly using their knowledge of letter-sound correspondences. Seventy-six trials are possible, however participants begin in the middle of the list and move backward if errors occur in order to establish their baseline. Participants then advance through the list until they make 6 errors in a row. Difficulty ranges from words like ''achieved'', ''tremendous'', ''systematic'' in the initial set (from the middle of the list), to ''homogenization'', ''indissolubly'' and ''ubiquitous'' in the final set. Average reliability of this task across the age range of our study participants is reported as .90 (McGrew & Woodcock, 2001); test duration is 5–8 min.

Word attack

Non-word reading was assessed via the Word Attack subtest, Form A. Participants read aloud individual pseudo-words (i.e., non-words adhering to the phonotactic constraints of English) presented in list form, divided into sets of 6 words, each set increasing in difficulty. This test is a relatively pure measure of skill in orthographic-phonological decoding, as there is no possibility that participants had previously learned pronunciations for these words. Thirty-two trials are possible, however participants begin in the middle of the list and move backward until 6 items are read correctly, then moving forward from the starting point until six items are read incorrectly. Difficulty ranges from ''tiff '' and ''zoop'' as the first items in the list, to ''phigh'', ''deprotenation'', and ''doitbility'' in the last list. Average reliability of this task across the age range of our study participants has been reported as .82 (McGrew & Woodcock, 2001); test duration is 3–5 min. For both tests, raw scores were converted to W scores, which are equalinterval scaled scores based on a transformation of the Rasch ability scale (Rasch, 1960; Wright & Stone, 1979) that are generally considered more appropriate for statistical analyses.

Comprehension

As noted in the Introduction, a frequently used measure of individual differences in comprehension is the Nelson– Denny Comprehension test. This test requires participants to read expository prose passages and answer a series of questions about the passage, drawing on vocabulary knowledge and inferencing skill. Normal administration allows 20 min to complete the test. The first of the 8 passages is 640 words long and requires answering 8

comprehension questions, while the next 7 passages are closer to 200 words long with 4 comprehension questions. While the test is well-suited for college-level readers, our previous experience with this population led us to expect a high proportion of readers at grade-school level (despite their older age). We therefore expected that many readers in our sample would be unable to complete the test in reasonable time (not even getting through the first passage), or would guess on a majority of questions, so we opted to assess reading comprehension via two other standardized tests, which are equally popular in educational settings.

Stanford fast reading

Reading comprehension was assessed using the Fast Reading subtest of the Stanford Diagnostic Reading Test (SDRT; Karlson & Gardner, 1995). This test consists of a short expository passage containing 30 choice points (similar to a cloze format) at which the participant is required to select the appropriate word from among three alternatives. For example, the test begins ''San Francisco, California is known as the city with cable cars. These are essentially street...'' (here the participant must choose between ''cleaners'', ''cars'', or ''lights'') ''that are pulled up or down the steep hills of that city by means of a...'' (again participants must choose, choices are ''horse'', ''sail'', or ''cable''). From here the story continues to describe how cable cars were constructed, including the history behind their use in the city. The final two test items are as follows: ''These two eighteen-foot-long cars travel in opposite directions and carry up to eight (choice point: ''cars'', ''passengers'', ''invitations''). There is not much need for seats, since the entire two-and-a-half-mile ride is (choice point: ''long'', ''bumpy'', ''brief '').'' In previous work with this population, we determined that the proportion of correct responses was a better indicator of their ability than their absolute score, thus all analyses reported here are based on the proportion correct measure. Reported reliability coefficients in the technical manual were greater than .85; test duration is 4 min.

PIAT-reading

An abridged version of the Peabody Individual Achievement Test-Revised (PIAT-R; Markwardt, 1998) also assessed reading comprehension. In this test, participants read a list of increasingly difficult sentences and then choose a picture, from an array of four, that best matches the meaning of the sentence. Notably, difficulty is increased mainly with respect to vocabulary items in the sentences, and not complexity of the sentence structure. While the most difficult sentence structures have 2 clauses, there are few embedded clauses represented. Examples of initial sentences are ''There is the sun.'', ''The bird is on the horse's back''; intermediate examples are ''The aquarium provides an entertaining display in the barbershop.'' and ''The inexperienced senator was now being bombarded with numerous inquiries.''; the most difficult sentences (last in the test) are ''The pollster is perturbed by the flippant observation of the minstrel she is interrogating.'' and ''A raconteur, the picture of sartorial resplendence, has a ponderous and histrionic barrister cajoling an adjudicator for clemency.'' Odd-numbered items from the subtest were administered, with a stop condition of 5 errors in 7 consecutive items. For the abridged form, we found a reliability of $\alpha = .90$; Leach, Scarborough, and Rescorla (2003) reported a reliability of $\alpha = .89$ for a similarly

abridged form of the task administered to fourth and fifth-grade students. Testing time is approximately 8–12 min.

Working memory

Sentence span

As discussed in the Introduction, the most commonly employed individual difference measure is the Daneman and Carpenter sentence span task, which has been argued to assess verbal working memory (Daneman & Carpenter, 1980; Turner & Engle, 1989). This task differs from measures of phonological memory (e.g., Digit Span described above) by adding a task-switching dimension to the task of remembering a list of items, thus testing participants' ability to allocate resources between the two components of the tasks. This is thought to provide a more accurate assessment of the working memory capacity that supports complex behavior. Indeed, a meta-analysis of 77 studies found that complex span tasks predicted language comprehension better than simple span tasks such as digit span (Daneman & Merikle, 1996). We used a listening version of the task, so as to avoid any confound with reading ability. Participants were required to judge sentences in an increasingly long series (2–5 items) as true or false and then, at the end of each series, to verbally recall the final words of every sentence in the series; words did not have to be recalled in the order presented. Sentences were designed to have relatively simple syntactic structure and for their truth-value to be easy to assess. Sample sentences are as follows: ''Mathematics deals with the study of plants and their environment.'', ''Dentists attend three years of law school before seeing patients.'', ''The elbow is the joint connecting the leg to the ankle.'', ''Traffic lights are helpful for controlling cars at busy corners.'' Scores corresponded to the total number of items correctly recalled. This test is not a standardized test, but was included in our battery due to substantial influence in reading research. No published reliability statistics are available; we calculated reliability in our sample as $\alpha = .85$. Testing time is 8–12 min.

Scoring

In all tests, except for rapid naming tests, a higher score corresponds to better performance. Rapid naming (RAN) tests, however, assess the performance of participants as the time it takes them to complete a task: so a higher score in those tests stands for a longer performance time and hence a worse performance. For easier interpretability of these tests, we transformed the original scores of rapid naming tests using the formula: $x_{transformed} = x_{max} + x_{min} - x_{original}$. This transformation ensures that the highest score is associated with the shortest time it took to complete the test, so the best performance, and the lowest score with the worst performance, as in other tests. Also the linear transformation does not alter the intervals between individual scores in rapid naming tests. All analyses reported here are based on transformed scores of rapid naming tests, coded as rln, rdn, ron and rcn, see Table 1.

Correlations between skill measures

We report the pairwise correlation matrix based on scores of 71 participants in tests of interest, see Table 6. The non-parametric Spearman rank correlation q, rather than the

Pearson's r, was used because the distributions of many scores were non-symmetrical and might lead to the disproportionate influence of outliers. We note that the population that we consider here is under researched, so that an examination of relationships between tests on the basis of strengths and directions of correlations between individual test scores is warranted. In particular, it is of interest to determine which tests reliably explain variation in reading skill for this age and ability group, and whether relationships between subtests (esp. phonological processing tests) suggested by previous research (e.g., Wagner et al., 1999) hold for this population. We explore this question and offer a cluster analysis of the correlational data in Table 6 elsewhere (Kuperman & Van Dyke, in preparation).

Appendix B

Stimuli sentences

Below is the complete list of sentences analyzed in the current study. Note from Materials section that each participant did not read this entire list, but rather only approximately 28 of these sentences, including all 11 of the simple sentences. Codes are as follows (S = simple; SE = single embedding; DE = double embedding).

Simple sentences

S1. The young doctor with a temperature diagnosed the patient.
S2. The well-known singer thanked his fans during intermission.
S3. The large hospital with budget problems fired the doctor.
S4. The informed citizen elected the most experienced candidate.
S5. The creative young architect preferred the complicated design.
S6. The loving grandmother made the cake for breakfast.
S7. The police officer arrested the boys.
S8. The principal of the high school suspended the undisciplined teen.
S9. The kind landlord improved the aging building.
S10. The police inspector with the funny raincoat questioned the worker for three full hours.
S11. The complicated scenarios capture real circumstances.

Single embeddings

SE1. The scientist that studies the climate recognized the writer from the magazine.
SE2. The spy that encoded the message delivered the secret to the authorities.
SE3. The psychologist that printed the notes asked the secretary for new copies.
SE4. The woman that reported the accident noticed the bicycle on the crosswalk.
SE5. The journalist that composed the article discovered a scandal in the government.
SE6. The child that loaded the revolver called the babysitter from the kitchen.
SE7. The man who blackmailed the school lived near the police station on the corner.
SE8. The climber who managed to reach the rope thanked the guide on the cliff.
SE9. The customer that consumed the meat lives in a country with few resources.
SE10. The golfer that mastered the game ignored the fans in the stands.
SE11. The trucker who knocked over the barricade bragged to the guys at the bar.

SE12. The contestant that misplaced the prize requested a replacement from the Academy.

SE13. The senator who criticized the report complained to the chairman of the committee.

SE14. The woman that coveted the jewelry avoided the man at the counter.

SE15. The salesman that examined the product mentioned the benefits in the newsletter.

SE16. The teacher that watched the play suspended a few of the students.

SE17. The reporter who described the article asked the editor about the details.

SE18. The worker who replaced the machine thanked the boss at Pratt & Whitney.

SE19. The pilot that crashed the plane requested a review by the safety board.

SE20. The dieter that desired the dessert requested the meal as a treat.

SE21. The fireman that fought the fire reported the damage in the cellar.

SE22. The director that watched the movie expected a prize at the film festival.

SE23. The soldier who shot at the tank shouted at the villagers during the battle.

SE24. The boys that drank the water invited the neighbors for a taste.

SE25. The doctor who cited the memo sued the hospital for the error.

SE26. The fish that attacked the boat fought the fisherman quite a lot.

SE27. The banker that refused the loan requested a meeting with the mayor.

SE28. The musician that witnessed the accident asked the policeman for a reward.

SE29. The actor that rehearsed the play expected first prize at the awards dinner.

SE30. The driver who passed the taxi yelled to the policeman on the corner.

SE31. The child who burned the matches called the babysitter in the next room.

SE32. The farmer that purchased the tractor called from the showroom of the store.

SE33. The student that practiced the instrument scheduled a performance in the auditorium.

SE34. The convict who hit the car yelled at the pedestrians during the escape.

SE35. The gardener that trimmed the plants described the house as more attractive.

Double embeddings

DE1. The neighbor wondered if the mailman who was afraid of the angry dog would knock on the door.

DE2. The lady forgot that the client who had come for the important meeting was waiting in the hall.

DE3. The vendor noticed that the boy who was eating up the hot hamburger was sitting on the curb.

DE4. The clerk anticipated that the teller who was working in the boring room would quit in a huff.

DE5. The girl saw that the woman who had brought 66 V. Kuperman, J.A. Van Dyke / Journal of Memory and Language 65 (2011) 42–73 along the cute teddy bear was playing with the toy.

DE6. The driver saw that the peddler who had begged for a kind handout was paying with a ten.

DE7. The owner hoped that the woman who had looked at the interesting painting would buy at the show.

DE8. The priest said that the woman who had stolen from the strict church was living by the rectory.

DE9. The man thought that the customer who had ordered up the new casserole was acting very badly.

DE10. The man knew that the merchant who had sold off the creative artwork was lying about the price.

DE11. The teen said that the boy who had asked for the popular number was asking too many favors.

DE12. The husband knew that the wife who is devoted to the beautiful home would cry at the ceremony.

DE13. The lady thought that the man who had set up the illegal fire was fighting a lost cause.

DE14. The ranger noticed that the hiker who was hiking along the unclear trail was afraid of bears.

DE15. The woman saw that the hostess who had cleaned up the messy table was talking to some friends.

DE16. The man yelled that the owner who had driven out the new undergrowth had made a big mistake.

DE17. The star saw that the reporter who had looked for the black sedan was watching at all times.

DE18. The soldier shouted that the enemy who had shot at the hidden tank was lying in the bushes.

DE19. The senior hoped that the aid who had picked up the lost ticket would come to the door.

DE20. The guide knew that the tourist who was searching for the old city was asking for big trouble.

DE21. The pilot knew that the man who is sitting in the smelly seat would argue with the crew.

DE22. The editor knew that the critic who had laughed at the memorable play would praise the cast.

DE23. The guard figured that the burglar who had given back the precious jewel was scared off the job.

DE24. The husband predicted that the woman who was paying off the new car would spend all the cash.

DE25. The boy believed that the elf who had talked about the amazing cave was telling a white lie.

DE26. The lady said that the boy who had dumped out the rich soil should apologize for the mistake.

DE27. The man noticed that the lady who was sitting in the new seat was talking on the phone.

DE28. The guard saw that the lady who had screamed about the dangerous fire was looking for a cop.

DE29. The parent asked if the teacher who had written up the new curriculum would come to the office.

DE30. The player knew that the coach who had sought out the helpful playbook was hoping for a win.

DE31. The grocer wondered if the customer who was holding up the heavy bag could carry the big load.

DE32. The man asked if the manager who had trained for the difficult program had gotten a new job.

DE33. The child hoped that the mother who had yelled at the bad dog would explain the real reason.

DE34. The kid knew that the woman who had yelled about the dirty room would forget about the mess.

DE35. The man said that the teller who had worked on the new account was wasting time and money.

Appendix C

Model selection

One of our goals was to identify a subset of tests that serve as reliable and independent predictors of the many eye-movement-dependent variables we considered. We addressed this issue by exploring the amount of unique variance explained by critical predictors (test scores) over and above other critical predictors and multiple controls, as estimated by multiple-regression models.[8] The two techniques we employed were mediation analysis (results reported in Column 3 of Tables 4 and 5) and the best-subset model selection procedure (Column 4 of Table 4).

Mediation analysis

Mediation analysis (Judd & Kenny, 1981; MacKinnon, 1994) applies to situations in which an independent variable (X) influences a dependent variable (Y) through one or multiple mediating variables (M). As outlined in (Preacher & Kayes, 2008), causal mediation analysis allows partitioning the causal effect of X into the indirect effect of X on Y via the proposed mediator M (which can be further apportioned into the effect of X on M, and the effect of M on Y partialling out the effect of X), and the direct effect of X on Y. We applied the mediation analysis to each dependent variable by testing all possible pairs of critical predictors (test scores) that showed significant main effects in the regression models fitted to that dependent variable (see Table 4, Column 2). In each pair, one of the test scores served as a mediator (M) and the other as a treatment (X): as we used the power set of pairs, every combination of the mediator and the treatment was tested. Causal mediation analysis was conducted using function mediate in library mediation in the statistical software R (Imai, Keele, Tingley, & Yamamoto, 2010). The function took as its input two linear regression models: one with X as an dependent variable and M as a predictor; and the other with an eye-movement measure as a dependent variable, and the baseline set of predictors (described in ''Baseline models''), as well as X and M as critical predictors. The output of the function included estimates and confidence intervals for the direct effect of X on Y, the indirect or mediation effect of X on Y via M, the total effect, as well as the proportion of the total effect accounted for by mediation. We considered the latter statistic, the proportion of the total effect via mediation, as an index of the contribution that a critical predictor makes to

explaining variance in the dependent variable over and above mediators: the lower the proportion, the stronger the contribution. For example, for gaze duration this proportion amounted to 53% when sentence span was the treatment variable X and word identification skill was the mediation variable. The proportion was down to a mere 1% when word identification was the treatment and sentence span was the mediator. Thus, word identification skill was a much stronger mediator for the sentence span skill than the other way around, and the former skill had more independent influence on gaze duration than the individual's performance in the sentence span task had. We computed the average proportion of the total effect by mediation for each treatment X (test score), as a statistic of the relative weakness of the treatment as an independent predictor of the dependent variable. Column 3 in Table 4 reports three tests that showed the lowest values (in the increasing order) of this statistic, and thus showed their relative independence from mediating variables. For gaze duration, the average proportion of total effect via mediation was 0.11 for word identification; 0.20 for the reading version of the Peabody Individual Achievement Test; and 0.21 for the listening version of the same test: we concluded that these tests, in this order, have the most independence as predictors of the behavioral measure. The maximum average proportion was observed for the digit memory span that had 52% of its total effect accounted for by the mediating variables. Using the harmonic mean for computing the average proportion led to the same results. We further applied causal mediation analysis to establish which test scores are independent and strong predictors of eye-movements when they enter into interaction with either word length or its lexical frequency (MacKinnon et al., 2007), see Table 5. We note that causal mediation analysis is not available for the mixed-effects models which we based our conclusions on in the body of the paper: rather it is based on linear (for continuous dependent variables) or generalized (for binary dependent variables) multiple regression models. To ensure that the present results of model selection are not specific to the methodology of mediation analysis, we validated them against the outcome of another model selection technique.

The best-subset model selection

This model selection algorithm selects the best subset of models based on the models' goodness-of-fit criteria rather than on the estimates of the amount of unique variance explained by an individual predictor over and above other predictors in the model. In its most comprehensive ''exhaustive'' format, this algorithm creates a power set of 2^k models with all possible combinations of k critical predictors, including the model where none are present and the model where all predictors are present. The model selection algorithm selects n best models of each size (with size ranging from 0 to k critical predictors) and ranks the resulting subset of models by their goodness-of-fit. The alternatives for navigating the model space implemented in this package include backward stepwise elimination (the method starts off with all candidate variables, tests them one-by-one against a selection criterion, and removes the candidate(s) that do not meet the criterion), forward model selection (the method starts off with no candidate variables, tests them one-by-one against a selection criterion and includes the candidate(s) that meet the selection criterion) or sequential replacement (both exclusion and inclusion are considered at each step).

Goodness-of-fit can be assessed using one of several established criteria that include the (adjusted or non-adjusted) amount of explained variance of the model, the model's

Bayesian Information Criterion, and others. The technique is implemented as function *regsubsets* in package *leaps* in the statistical software package *R.* We used the function with the parameters that allowed for the least restricted search through the model space: 10 best models selected for each model size; intercept added; and the default Mallows' Cp criterion for the model's goodness-of-fit. We also ran this model selection technique for all available search algorithms: exhaustive, backward, forward and sequential replacement.

As the input, we provided the full list of control predictors as defined in the baseline models, as well as the full list of 18 test scores as critical predictors: the controls were set to be present as constant terms in all models that the function fitted. Notably, function *regsubsets* can only operate with continuous dependent variables and does not allow testing interactions between predictors: thus, some cells in Column 5 of Table 4 are empty, and we could not use the technique for verifying the results in Table 5. (Function *step* in library *stats* does fit models to categorical and continuous dependent variables, and can handle interactions; we found, however, that the outcomes of this function are overly inclusive as compared to the outcomes of our mediation analysis and the best-subset model selection: we opted for not pursuing this option further.) We note that the best-subset selection algorithms, just like causal mediation analysis, are only implemented for linear regression models, rather than linear mixed effects models. Thus, the results of the best-subset selection, if used as the only selection procedure, should be treated with caution: we used it in conjunction with mediation analysis and were reassured by the converging pattern of results across the two techniques (see below).

The outcomes of the procedure are reported in Column 4 of Table 4 and they show excellent convergence between search algorithms: the slightly deviating outcomes of the backward elimination are reported in parentheses. Even though the best-fitting model may have had more than one (up to 18) critical predictors in it, all models that made it into the top three rank based on their goodness-of-fit performance had exactly one critical predictor. Even more importantly, there was a remarkable convergence between the model selection done with the help of causal mediation analysis and the model selection done using the best subset method, even though these techniques are based on different assumptions and selection criteria. The discrepancies are confined to the inclusion of sentence span as one of the three best-performing predictors of first fixation position in mediation analysis, and of the reading version of PIAT in the best-subset model selection, as well as the inclusion of the non-word repetition task as one of the three best-performing predictors of total fixation time in mediation analysis, and rapid digit naming in the best subset method. Otherwise, the outcomes of the two techniques are identical, modulo the occasional different ranking of the three variables chosen by each technique.

Crucially, both techniques confirm our conclusions that word identification and rapid letter/digit naming tasks are pervasive predictors of eye-movement behavior, and that the scores in the Peabody Individual Achievement Test and (less consistently) phoneme reversal are reliable predictors of early (first pass-related) eye-movement measures and that memory-related tasks of late and cumulative eye-movement measures. This convergence in the testing of main effects suggests that the results that we report in Table 5 for interactions based solely on mediation analysis are well-founded as well.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/ j.jml.2011.03.00

References

Adler-Grinberg, D., & Stark, L. (1978). Eye movements, scanpaths and dyslexia. American Journal of Optometry and Physiological Optics, 55, 557–570.

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. Memory & Cognition, 24, 477–492.

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.

Ashby, J., Rayner, K., & Clifton, C. Jr. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. The Quarterly Journal of Experimental Psychology Section A, 58A, 1065–1086.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390–412.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition for single-syllable words. Journal of Experimental Psychology: General, 133, 283–316.

Barth, A., Catts, H. W., & Anthony, J. (2009). The component skills underlying reading fluency in adolescent readers: A latent variable analysis. Reading and Writing, 22, 567–590.

Bates, D. M. & Sarkar, D. (2007). The lme4 library. <http:// lib.stat.cmu.edu/R/CRAN/>

Biscaldi, M., Gezeck, S., & Stuhr, V. (1998). Poor saccadic control correlates with dyslexia. Neuropsychologia, 36, 1189–1202.

Blythe, H. I., Liversedge, S. P., Joseph, H. S. S. L., White, S. J., & Rayner, K. (2009). Visual information capture during fixations in reading for children and adults. Vision Research, 49, 1583–1591.

Boland, J. (2004). Linking eye movements to sentence comprehension in reading and listening. In M. Carreiras & C. Clifton, Jr. (Eds.), The on-line study of sentence comprehension: Eyetracking, ERP, and beyond (pp. 51–76). Brighton, England: Psychology Press.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research, 2, 1–12.

Bowers, P., Golden, J., Kennedy, A., & Young, A. (1994). Limits upon orthographic knowledge due to processing indexed by naming speed. In V. Berninger (Ed.), The 45 varieties of orthographic knowledge I: Theoretical and developmental issues (pp. 173–218). Dordrecht: Kluwer.

Bowers, P., & Newby-Clark, E. (2002). The role of naming speed within a model of reading acquisition. Reading and Writing, 15, 109–126.

Bowers, P. G., & Swanson, L. B. (1991). Naming speed deficits in reading disability: Multiple measures of a singular process. Journal of Experimental Child Psychology, 51, 195–219.

Bowers, P. G., & Wolf, M. (1993). Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. Reading and Writing, 5, 69–85.

Brady, S., Braze, D., & Fowler, C. A. (Eds.). (in press). Explaining individual differences in reading: Theory and evidence. Psychology Press.

Breznitz, Z. (2006). Fluency in reading: Synchronization of processes. Erlbaum.

Breznitz, Z., & Misra, M. (2003). Speed of processing of the visualorthographic and auditory-phonological systems in adult dyslexics: The contribution of "asynchrony" to word recognition deficits. Brain and Language, 85, 486–502.

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. Memory & Cognition, 7, 68–76.

Carello, C., LeVasseur, V., & Schmidt, R. (2002). Movement sequencing and phonological fluency in (putatively) nonmpaired readers. Psychological Science, 13, 375–379.

Carreiras, M., & Clifton, C. Jr., (2004). The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond. New York: Psychology Press.

Catts, H. W., Gillespie, M., Leonard, L. B., Kail, R. V., & Miller, C. A. (2002). The role of speed of processing, rapid naming, and phonological awareness in reading achievement. Journal of Learning Disabilities, 35, 509–524.

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.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. Journal of Experimental Psychology: Learning, Memory and Cognition, 27, 225–235.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. Memory & Cognition, 28, 143–153.

Clifton, C., Jr., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. Journal of Memory and Language, 49, 317–334.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. Psychological Review, 108, 204–256.

Cunningham, T. F., Healy, A. F., Kanengiser, N., Chizzick, L., & Willitts, R. L. (1988). Investigating the boundaries of reading units across ages and reading levels. Journal of Experimental Child Psychology, 45, 175– 208.

Cunningham, A. E., Stanovich, K. E., & Wilson, M. R. (1990). Cognitive variation in adult college students differing in reading ability. In T. H. Carr & B. Levy (Eds.), Reading and its development: Component skills approaches (pp. 129–159). San Diego, CA, US: Academic Press.

Cutting, L. E., & Denckla, M. B. (2001). The relationship of rapid serial naming and word reading in normally developing readers: An exploratory model. Reading and Writing, 14, 673–705.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. Psychonomic Bulletin & Review, 3, 422–433.

Denckla, M. B. (1985). Eye Motor coordination in dyslexic children: Theoretical and clinical implications. In F. Duffy & N. Geschwind (Eds.), Dyslexia: A neuroscientific approach to clinical evaluation (pp. 187–196). Boston: Little, Brown.

Dunn, L., & Markwardt, F. (1970). Peabody individual achievement test. Circle Pines, MN: American Guidance Service.

Eden, G. E., Stein, J. E., Wood, H. M., & Wood, E. B. (1994). Differences in eye movements and reading problems in dyslexic and normal children. Vision Research, 34, 1345–1358.

Ehri, L. C. (1999). Phases of development in learning to read words. In J. Oakhill & R. Beard (Eds.), Reading development and the teaching of reading (pp. 79–108). Malden, MA: Blackwell.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. Journal of Verbal Learning and Verbal Behavior, 20, 641–655.

Elterman, R. D., Abel, L. A., Daroff, R. B., Dell'Osso, L. E., & Bornstein, J. L. (1980). Eye movement patterns in dyslexic children. Journal of Learning Disabilities, 13, 312–317.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. Vision Research, 42, 621–636.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. Psychological Review, 112, 777–813.

Frederiksen, J. R. (1978). Assessment of lexical decoding and lexical skills in their relation to reading proficiency. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), Cognitive psychology and instruction. New York: Plenum.

Gladstone, M., Best, C. T., & Davidson, R. J. (1989). Anomalous bimanual coordination among dyslexic boys. Developmental Psychology, 25, 236–246.

Grigg, W., Donahue, P., & Dion, G. (2007). The nation's report card: 12thgrade reading and mathematics 2005 (NCES 2007-468). US Department of Education, National Center for Education Statistics, Washington, DC.

Häikiö, T., Bertram, R., Hyönä, J., & Niemi, P. (2009). Development of the letter identity span in reading: Evidence from the eye movement moving window paradigm. Journal of Experimental Child Psychology, 102, 167–181.

Harm, M. W., & Seidenberg, M. S. (1999). Reading acquisition, phonology, and dyslexia: Insights

from a connectionist model. Psychological Review, 106, 491–528.

Harm, M. W., & Seidenberg, M. S. (2001). Are there orthographic impairments in phonological dyslexia? Cognitive Neuropsychology, 18, 71–92.

Hawelka, S., Gagl, B., & Wimmer, H. (2010). A dual-route perspective on eye movements of dyslexic readers. Cognition, 115, 367–379.

Healy, A. F. (1994). Letter detection: A window to unitization and other cognitive processes in reading text. Psychonomic Bulletin & Review, 1, 333–344.

Heil, M., Rolke, B., Engelkamp, J., Roesler, F., Ozcan, M., & Hennighausen, E. (1999). Event-related brain potentials during recognition of ordinary and bizarre action phrases following verbal and subject-performed encoding conditions. European Journal of Cognitive Psychology, 11, 261–280.

Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 417–429.

Hill, R., & Murray, W. (2000). Commas and spaces: Effects of punctuation on eye movements in sentence processing. In A. Kennedy, D. Heller, & J. Pynte (Eds.), Reading as a perceptual process (pp. 565–589). Amsterdam: Elsevier.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. Journal of Memory and Language, 54, 425–443.

Huestegge, L., Radach, R., Corbic, D., & Huestegge, S. M. (2009). Oculomotor and linguistic determinants of reading development: A longitudinal study. Vision Research, 49, 2948–2959.

Hutzler, F., Kronbichler, M., Jacobs, A. M., & Wimmer, H. (2006). Perhaps correlational but not causal: No effect of dyslexic readers' magnocellular system on their eye movements during reading. Neuropsychologia, 44, 637–648.

Hyönä, J., Niemi, P., & Underwood, G. (1989). Reading long words embedded in sentences: Informativeness of word halves affects eye movements. Journal of Experimental Psychology: Human Perception and Performance, 15, 142–152.

Hyönä, J., & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 1430–1439.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. In H. D. Vinod (Ed.). Advances in social science research using R (Vol. 196, pp. 129–154). New York: Springer.

Ingels, S.J., Dalton, B.W., & LoGerfo, L. (2008). Trends Among High School Seniors, 1972–2004 (NCES 2008-320). National Center for Education Statistics, Institute for Education Sciences, US Department of Education, Washington, DC.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. Perception & Psychophysics, 40, 431–439.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59, 434–446.

Jared, D., Levy, B. A., & Rayner, K. (1999). The role of phonology in the activation of word meanings during reading: Evidence from proofreading and eye movements. Journal of Experimental Psychology: General, 128, 219–264.

Jones, M. W., Branigan, H. P., & Kelly, M. L. (2009). Dyslexic and nondyslexic reading fluency: Rapid automatized naming and the importance of continuous lists. Psychonomic Bulletin & Review, 16, 567–572.

Jones, M. W., Obregon, M., Kelly, M. L., & Branigan, H. P. (2008). Elucidating the component processes involved in dyslexic and nondyslexic reading fluency: An eye-tracking study. Cognition, 109, 389–407.

Jorm, A. F., & Share, D. L. (1983). Phonological recoding and reading acquisition. Applied Psycholinguistics, 4, 103–147.

Joseph, H. S. S. L., Liversedge, S. P., Blythe, H. I., White, S. J., & Rayner, K. (2009). Word length and landing position effects during reading in children and adults. Vision Research, 49, 2078–2086.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. Evaluation Review, 5, 602–619.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. Psychological Bulletin, 131, 684–712.

Juhasz, B., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. Journal of Experimental Psychology: Learning, Memory and Cognition, 29, 1312–1317.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. Visual Cognition, 13, 846–863.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329–354.

Kail, R., & Hall, L. K. (1994). Processing speed, naming speed and reading. Developmental Psychology, 30, 949–954.

Kail, R., Hall, L. K., & Caskey, B. L. (1999). Processing speed, exposure to print, and naming speed. Applied Psycholinguistics, 20, 303–314.

Karlson, B., & Gardner, E. (1995). Stanford diagnostic reading test (4th ed.). San Antonio, TX: Psychological Corp.

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

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45, 153–168.

Kennison, S. M., & Clifton, C. Jr., (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: Implications for eye movement control. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 68.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16, 262–284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present and future words on fixation durations. Journal of Experimental Psychology: General, 1, 12–35.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. The Quarterly Journal of Experimental Psychology, 63(9), 1838–1857.

Kuperman, V., & Van Dyke, J. A. (in preparation). Construct-validity of tests of verbal ability: Cluster analyses for the non-college-bound young-adult population.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied linear regression models (4th ed.). Boston: McGraw-Hill Irwin.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6, 293–323.

Laubrock, J., Kliegl, R., & Engbert, R. (2006). SWIFT explorations of age differences in eye movements during reading. Neuroscience and Biobehavioral Reviews, 30, 872–884.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. Journal of Educational Psychology, 95, 211–224.

Lefton, L. A., Nagle, R. J., Johnson, G., & Fisher, D. F. (1979). Eye movement dynamics of good and poor readers: Then and now. Journal of Reading Behavior, 11, 319–328.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behaviour Research Methods, Instruments, and Computers, 28, 203–208.

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. In A. Cazares & L. A. Be atty (Eds.), Scientific methods in prevention research (pp. 127–153). Washington, DC: US Government Printing Office.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. Annual Review of Psychology, 58, 593–614.

Manis, F. R., Doi, L. M., & Bhadha, B. (2000). Naming speed, phonological awareness and orthographic knowledge in second graders. Journal of Learning Disabilities, 33, 325–333.

Manis, F., & Freedman, L. (2001). The relationship of naming to multiple reading measures in disabled and non-disabled normal readers. In M. Wolf (Ed.), Dyslexia, fluency and the brain (pp. 65–92). MD: York Press.

Markwardt, F. C. Jr., (1998). Peabody individual achievement test – Revised. Circle Pines, MN: American Guidance Service.

Martos, E. J., & Vila, J. (1990). Differences in eye movement control among dyslexic, retarded and normal readers in the Spanish population. Reading and Writing, 2, 175–188.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. Trends in Cognitive Sciences, 7, 293–299.

McCardle, P., & Pugh, K. (2009). How children learn to read: Current issues and new directions. In The integration of cognition, neurobiology and

genetics of reading and dyslexia research and practice. Psychology Press.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. Perception & Psychophysics, 17, 578–586.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial fixations in words. Vision Research, 28, 1107–1118.

McDonald, S. A., Carpenter, R. H. S., & Shillcock, R. C. (2005). An anatomically-constrained, stochastic model of eye movement control in reading. Psychological Review, 112, 814–840.

McGrew, K. S., & Woodcock, R. W. (2001). Technical manual: WoodcockJohnson III. Itasca, IL: Riverside.

Misra, M., Katzir, T., Wolf, M., & Poldrack, R. A. (2004). Neural systems for rapid automatized naming in skilled readers: Unraveling the RANreading relationship. Scientific Studies of Reading, 8, 241–256.

National Center for Education Statistics (NCES) (2005). A first look at the literacy of Americas adults in the 21st century. Government Printing Office, Washington, DC.

Nicolson, R. I., & Fawcett, A. J. (1990). Automaticity: A new framework for dyslexia research? Cognition, 35, 159–182.

Nuthmann, A., Engbert, R., & Kliegl, R. (2005). Mislocated fixations during reading and the inverted optimal viewing position effect. Vision Research, 45, 2201–2217.

Olson, R. K., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyon (Ed.), Frames of reference for the assessment of learning disabilities: New views on measurement issues (pp. 243–277). Baltimore: Paul H. Brookes.

Olson, R. K., Kliegl, R., & Davidson, B. J. (1983). Dyslexic and normal readers' eye movements. Journal of Experimental Psychology: Human Perception and Performance, 9, 816–825.

Olson, R. K., Wise, B., Conners, F., & Rack, J. (1990). Organization, heritability, and remediation of component word recognition and language skills in disabled readers. In T. H. Carr & B. Levy (Eds.), Reading and its development: Component skills approaches (pp. 261–322). San Diego, CA, US: Academic Press.

Osaka, N., & Osaka, M. (2002). Individual differences in working memory during reading with and without parafoveal information: A moving-window study. American Journal of Psychology, 115, 501–513.

Pavlidis, G. T. (1985). Eye movements in dyslexia: Their diagnostic significance. Journal of Learning Disabilities, 18, 42–50.

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

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), Reading acquisition (pp. 145–174). Hillsdale, NJ: Erlbaum.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. Scientific Studies of Reading, 11, 357–383.

Perfetti, C. A., Finger, E., & Hogaboam, T. W. (1978). Sources of vocalization latency differences between skilled and less skilled young readers. Journal of Educational Psychology, 70, 730–739.

Perfetti, C. A., & Hogaboam, R. (1975). Relationship between single word decoding and reading comprehension skill. Journal of Educational Psychology, 67, 461–469.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. Journal of Consumer Research, 28, 450–461.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. Statistics and computing. New York: Springer.

Pirozzolo, E. J., & Rayner, K. (1978). Disorders of oculomotor scanning and graphic orientation in developmental Gerstmann syndrome. Brain and Language, 5, 119–126.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103, 56–115.

Poldrack, R. A., & Gabrieli, J. D. E. (2001). Characterizing the neural mechanisms of skill learning and repetition priming: Evidence from mirror reading. Brain, 124, 67–82.

Poldrack, R. A., & Wagner, A. D. (2004). What can neuroimaging tell us about the mind? Current Directions in Psychological Science, 13, 177–181.

Polk, T. A., Stallcup, M., Aguirre, G. K., Alsop, D. C., D'Esposito, M., Detre, J. A., et al. (2002). Neural specialization for letter recognition. Journal of Cognitive Neuroscience, 14, 145–159.

Powell, D., Stainthorp, R., Stuart, M., Garwood, H., & Qunlan, P. (2007). An experimental comparison between rival theories of rapid automatized naming performance and its relationship to reading. Journal of Experimental Child Psychology, 98, 46–68.

Preacher, K. J., & Kayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40, 870–891.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. Journal of Anatomy, 197, 335–359.

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., et al. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). Mental Retardation & Developmental Disabilities Research Reviews, 6, 207–213.

Pugh, K. R., Frost, S. J., Sandak, R., Landi, N., Rueckl, J. G., Constable, R. T., et al. (2008). Effects of stimulus difficulty and repetition on printed word identification: An fMRI comparison of nonimpaired and reading-disabled adolescent cohorts. Journal of Cognitive Neuroscience, 20, 1146–1160.

Pugh, K., & McCardle, P. (2009). How children learn to read. New York: Psychology Press.

Pynte, J., & Kennedy, A. (2006). An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. Vision Research, 46, 3786–3801.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. Psychological Research, 72, 675–688.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Paedagogiske Institut.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. Perception, 8, 21–30.

Rayner, K. (1985a). Do faulty eye movements cause dyslexia? Developmental Neuropsychology, 1, 3–15.

Rayner, K. (1985b). The role of eye movements in learning to read and reading disability. Remedial and Special Education, 6, 53–60.

Rayner, K. (1986). Eye movement and the perceptual span in beginning and skilled readers. Journal of Experimental Child Psychology, 41, 211–236.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology, 62, 1457–1506.

Rayner, K., & Bertera, J. H. (1979). Reading without a fovea. Science, 206, 468–469.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. Memory & Cognition, 14, 191–201.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. Quarterly Journal of Experimental Psychology A, 53, 1061–1080.

Rayner, K., Murphy, L. A., Henderson, J. M., & Pollatsek, A. (1989). Selective attentional dyslexia. Cognitive Neuropsychology, 6, 357–378.

Rayner, K., & Pollatsek, A. (2006). Eye-movement control in reading. In M. Traxler & M. Gernsbacher (Eds.), Handbook of psycholinguistics (pp. 613–658). Amsterdam: Elsevier.

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

Rayner, K., Sereno, S., Morris, R., Schmauder, A., & Clifton, C. (1989). Eye movements and on-line language comprehension processes. Language and Cognitive Processes, 4, 21–49.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: Comparison of two types of models. Journal of Experimental Psychology: Human Perception and Performance, 22, 1188–1200.

Rayner, K., Slattery, T. J., & Bélanger, N. N. (2010). Eye movements, the perceptual span, and reading speed. Psychonomic Bulletin & Review, 17(6), 834–839.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. Psychonomic Bulletin & Review, 3, 504–509.

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.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E–Z Reader: A cognitive control, serial-attention

model of eye-movement behavior during reading. Cognitive Systems Research, 7, 4–22.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E–Z Reader model of eye-movement control in reading: Comparisons to other models. Behavioral and Brain Sciences, 26, 445–526.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E–Z Reader to model the effects of higher-level language processing on eye movements during reading. Psychonomic Bulletin & Review, 16, 1–21.

Reilly, R. G., & Radach, R. (2003). Foundations of an interactive activation model of eye movement control in reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), The mind's eye: Cognitive and applied aspects of eye movements (pp. 429–455). Amsterdam: North-Holland.

Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye-movement control in reading. Cognitive Systems Research, 7, 34–55.

Richter, E. M., Engbert, R., & Kliegl, R. (2006). Current advances in SWIFT. Cognitive Systems Research, 7, 23–33.

Rieben, L., & Perfetti, C. A. (Eds.). (1991). Learning to read: Basic research and its implications. Hillsdale, NJ: Erlbaum.

Saint-Aubin, J., & Klein, R. M. (2008). The influence of reading skills on the missing-letter effect among elementary school students. Reading Research Quarterly, 43, 132–146.

Samuels, S. J., LaBerge, D., & Brener, C. D. (1978). Units of word recognition: Evidence for developmental changes. Journal of Verbal Learning and Verbal Behavior, 17, 715–720.

Samuels, S. J., Miller, N. L., & Eisenberg, P. (1979). Practice effects on the unit of word recognition. Journal of Educational Psychology, 71, 514–520.

Sandak, R., Mencl, W. E., Frost, S. J., Rueckl, J., Katz, L., Moore, D. L., et al. (2004). The neurobiology of adaptive learning in reading: A contrast of different training conditions. Cognitive, Affective, and Behavioral Neuroscience, 4, 67–88.

Savage, R. S. (2004). Motor skills, automaticity and dyslexia: A review of the research literature. Reading and Writing: An Interdisciplinary Journal, 17, 301–325.

Savage, R., Pillay, V., & Melidona, S. (2007). Deconstructing rapid naming: Component processing and the prediction of reading difficulties. Learning and Individual Differences, 17, 129–146.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid serial naming, and IQ. Annals of Dyslexia, 48, 115–136.

Schilling, H., Rayner, K., & Chumbley, I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. Memory & Cognition, 26, 1270–1281.

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Lentgeb, S., Schall, J. D., et al. (1998). Signal timing across the macaque visual system. Journal of Neurophysiology, 79, 3272–3278.

Schroyens, W., Vitu, F., Brysbaert, M., & d'Ydewalle, G. (1999). Eye movement control during reading: Foveal load and parafoveal processing. Quarterly Journal of Experimental Psychology, 52A, 1021–1046.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96, 523–568.

Seidenberg, M. S. (in press). Computational models of reading: Connectionist and dual-route approaches. In M. Spivey, K. McRae, & M. Joanisse (Eds.), Cambridge handbook of psycholinguistics. Cambridge University Press.

Shankweiler, D. P., & Liberman, I. Y. (1989). Phonology and reading disability: Solving the reading puzzle. Ann Arbor, MA: University of Michigan Press.

Shankweiler, D., Lundquist, E., Dreyer, L. G., & Dickinson, C. C. (1996). Reading and spelling difficulties in high school students: Causes and consequences. Reading and Writing, 8, 267–294.

Share, D. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. Cognition, 55, 151–218.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P., et al. (2002). Disruption of posterior brain systems for reading in children with developmental dyslexia. Biological Psychiatry, 52, 101–110.

Shaywitz, S. E., Shaywitz, B. A., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., et al. (2003). Neural systems for compensation and persistence: Young adult outcome of childhood reading disability. Journal of Biological Psychiatry, 54, 25–33.

Solan, H. A. (1985). Deficient eye movement patterns in achieving high school students: Three case histories. Journal of Learning Disabilities, 18, 66–70.

Spring, C., & French, L. (1990). Identifying children with specific reading disabilities from listening and reading discrepancy scores. Journal of Learning Disabilities, 23, 53–58.

Stanovich, K. E., Feeman, D. J., & Cunningham, A. E. (1983). The development of the relation between letter-naming speed and reading ability. Bulletin of the Psychonomic Society, 21, 199–202.

Tainturier, M. J., Tremblay, M., & Lecours, A. R. (1992). Educational level and the word frequency effect: A lexical decision investigation. Brain and Language, 43, 460–474.

Terry, S., Samuels, S. J., & LaBerge, D. (1976). The effects of letter degradation and letter spacing on word recognition. Journal of Verbal Learning and Verbal Behavior, 15, 577–585.

Tinker, M. A. (1946). The study of eye movements in reading. Psychological Bulletin, 43, 93–120.

Tinker, M. A. (1958). Recent studies of eye movements in reading. Psychological Bulletin, 55, 215–231.

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

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. Memory & Cognition, 35, 1107.

Traxler, M. J. (2009). A hierarchical linear modeling analysis of working memory and implicit prosody in the resolution of adjunct attachment ambiguity. Journal of Psycholinguistic Research, 38, 491–509.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? Journal of Memory and Language, 28, 127–154.

Ungerleider, L. G., Doyon, J., & Karni, A. (2002). Imaging brain plasticity during motor skill learning. Neurobiology of Learning and Memory, 78, 553–564.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 407–430.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. Journal of Educational Psychology, 86, 601–638.

Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: An inverted optimal viewing position effect. Vision Research, 41, 3513–3533.

Vitu, F., Lancelin, D., & d'Unienville, V. (2007). A perceptual-economy account for the inverted-optimal viewing position effect. Journal of Experimental Psychology: Human Perception and Performance, 33, 1220–1249.

Vitu, F., & O'Regan, J. K. (1995). A challenge to current theories of eye movements in reading. In J. M. Findlay, R. Walker, & R. W. Kentridge (Eds.), Eye movement research: Mechanisms, processes, and applications (pp. 381–393). Amsterdam: North Holland.

Vitu, F., O'Regan, J. F., & Mittau, M. (1990). Optimal landing position in reading isolated words and continuous text. Perception & Psychophysics, 47, 583–600.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. Psychological Bulletin, 101, 192–212.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. Developmental Psychology, 30, 73–87.

Wagner, R., Torgesen, J., & Rashotte, C. A. (1999). Comprehensive test of phonological processing. Austin, TX: Pro-Ed.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. Developmental Psychology, 33, 468–479.

Wang, Y., Sereno, J. A., Jongman, A., & Hirsch, J. (2003). fMRI evidence for cortical modification during learning of Mandarin lexical tone. Journal of Cognitive Neuroscience, 15, 1019–1027.

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z Reader. Cognition, 111, 132–137.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. European Journal of Cognitive Psychology, 16, 312–339.

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. Journal of Educational Psychology, 91, 415–438.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. Scientific Studies of Reading, 5, 211229.

Wolf, M., & O'Brien, B. (2001). On issues of time, fluency and intervention. In A. Fawcett (Ed.), Dyslexia: Theory and good practice (pp. 124–140). London: Whurr Publishers.

Wolf, M., O'Rourke, G. A., Gidney, C., Lovett, M., Cirino, P., & Morris, R. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. Reading and Writing: An Interdisciplinary Journal, 15, 43–72.

Wolff, P. H., Michel, G. F., & Ovrut, M. (1990). Rate variables and automatized naming in developmental dyslexia. Brain and Language, 39, 556–575.

Wolff, P. H., Michel, G. F., Ovrut, M., & Drake, C. (1990). Rate and timing precision of motor coordination in developmental dyslexia. Developmental Psychology, 26, 349–359.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III tests of achievement. Itasca, IL: Riverside.

Wright, B., & Stone, M. (1979). Best test design. Chicago: Mesa Press.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. Journal of Memory and Language, 47, 1–29.

Table 1 Summary of tests of verbal skills. Distributional data (in quartiles), grade equivalents and labels used in statistical models

Domain	Skill	Label	Raw scores							Grade equivalents				Range (standardized)
			Mean	SD	Min.	25%	50%	75%	Max.	Mean	SD	Min.	Max.	
Phono. awareness	Age	age	20.8	2.6	16	18.6	21.1	22.6	25.7	-	–	-	–	
	1. Elision	elision	15.1	4.6	6	12.5	17	19	20	6.8	3.3	1.7	>9.7	-1.98:1.06
	2. Blending words[a]	blendw	13.5	4.6	4	10	13	18	20	5.9	4	k.7	>9.7	-2.09:1.46
	3. Blending nonwords[a]	blendnw	8.8	3.1	2	7	9	10	16	5.7	3.3	k.7	>9.7	-2.18:2.34
	4. Segmenting words	segmentw	11.4	4.3	2	8	12	15	19	6.4	3.8	2	>9.7	-2.11:1.77
	5. Segmenting nonwords	segmentnw	13.7	5.4	0	8	16	18	20	7.3	3.3	2	>9.7	-2.56:1.20
	6. Phoneme reversal[a]	phonrev	10.4	4.6	1	6	11	14	18	7.2	3.2	k.7	>9.7	-2.06:1.70
Simple memory span	7. Digit span[a]	digmem	16.3	2.6	9	14.5	16	18	21	8.7	2.7	k.2	>9.7	-2.82:1.84
	8. Non-word repetition[a]	nwrep	10.3	2.5	5	8	10	13	15	4	3.6	k.0	9.7	-2.13:1.85
Rapid Naming: Print[e]	9. Digit naming	rdn	23.7	4.7	14.8	23.3	26.2	30.1	35.4	9.1	1.5	4.7	>9.7	-2.48:1.92
	10. Letter naming	rln	25.2	4.9	13.2	21.8	24.3	29.2	37.2	9.1	1.5	5	9.7	-2.15:2.37
Rapid naming: non-print[e]	11. Color naming	rcn	38.7	8.3	25.2	33.5	37	42.1	62.6	9	1.6	4.4	>9.7	-2.96:1.65
	12. Object Naming	ron	43	7.6	29.5	37.6	41.6	48.6	62.9	9.1	1.5	4.7	>9.7	-2.71:1.81
Word and non-word reading	13. Word identification [b,c]	wid	549	22.6	500	533	545	568	588	12.9	5.1	4.4	>18	-2.22:1.78
	14. Word attack [b,c]	watt	518	16.5	468	510	520	530	545	10.9	5.4	1.9	>18	-3.15:1.66
Reading comp.	15. Stanford fast reading[c]	stan	0.9	0.1	0.5	0.8	0.9	1	1	10.2	3.3	3.4	13	-2.79:1.06

	16. PIAT reading [c,d]	piatr	33.9	5.8	22	29	36	38.5	41	9.8	3.1	4.3	13	-2.02:1.27
	17. PIAT listening[c]	piatl	34.4	4.9	22	31.5	34	38.5	41	9.9	2.7	4.3	13	-2.50:1.44
Complex memory span	18. Sentence span	span	76.9	15.2	26.7	67.8	78.3	88.3	100	-	-	-	-	-3.40:1.58

[a] Grade levels prior to first grade are indicated as k [month], where [month] indicates months into the kindergarten year.
[b] W-scores reported here. See note in Appendix A for explanation.
[c] Grade Equivalent of ''post high school'' is indicated by 13; grade equivalent of "second year of graduate school'' is indicated as 18.
[d] Actual mean = 85.8, actual min = 62, actual max = 100. See note in Appendix A for description of revised administration and scoring procedure. [e] For comparability with other tests, we transformed the original scores of rapid naming tests so that the highest score is associated with the shortest time it took to complete the test, so the best performance, and the lowest score with the worst performance (see section Scoring in Appendix A for details). All analyses reported here are based on transformed scores of rapid naming tests, coded as rln, rdn, ron and rcn.

Table 2 Summary of eye-movement measures as dependent variables

Eye-movement measure	Code	Mean (SD)	Range
Initial landing position (characters): the landing position of the first fixation on the word, measured in characters from the word's left boundary	FirstFixPos	2.5(1.9)	0–11
First fixation duration (ms): duration of the first fixation on the word	FirstFixDur	238(97)	52–948
Single fixation duration (ms): duration of the only fixation in the first reading pass (i.e. before the eyes move away from the word for the first time)	SingleFixDur	242(100)	52–948
Refixation likelihood: index of a refixation on the word in the first reading pass	RefixProb	0.24(0.43)	0–1
Gaze duration (ms): the summed duration of fixations on the word before the eyes move away from the word	GazeDur	299(153)	52–2128

Regression likelihood: index of a regression (a saccade to the left of the word) after the first reading pass	RegrProb	0.19(0.36)	0–1
Regression path duration (Go-past time, ms): the summed duration of all fixations on the word and words to its left after the first fixation on the word is made and before the eyes moved to the right of the word for the first time	RegrPathTime	417(372)	52–6044
Second pass likelihood: index of whether there was a second pass on the word (regardless of whether the second pass is preceded by a progressive or a regressive saccade).	SecPassProb	0.42(0.49)	0–1
Second pass reading time (ms): the summed duration of fixations on the word in the second reading pass	SecPassTime	395(298)	52–2752
Total fixation time (ms): the summed duration of all fixations that landed on the word	TotalTime	464(311)	52–3232

Table 3 Summary of continuous text-level predictors, including labels used in statistical models. Mean values, standard deviations and ranges are reported for raw values of predictors and for values of predictors after transformation.

Predictor	**Code**	**Mean(SD): original**	**Range: original**	**Range: transformed**
Length of word N, characters	sWordLength	6(2)	2:13	-1.8:3.0
Frequency of word N (residualized)	srWordFreq	201,522(472,855)	101;6.4x 10^6	-3.2:2.4
Length of word N - 1, characters	sPrevLength	4(2)	1:11	-1.5:3.3
Frequency of word N -1 (residualized)	srPrevFreq	9.1 x 10^6(1.1 x 10^7)	101:2.3x 10^7	-2.7:1.2
Length of Word N + 1, characters	sNextLength	5(2)	1:13	-1.5:3.6
Frequency of word N + 1 (residualized)	srNextFreq	4.8 x 10^6(7.8 x 10^6)	110:2.3 X 10^7	-3.5:2.7

Relative word position in sentence	sRelPos	0.51(0.23)	0.11:0.94	-1.7:1.9
Sentence position in experimental list	sTrialNum	73.9(40.3)	1:144	-1.8:1.8
Initial landing position, characters	sFirstFixPos	2.5(1.9)	0:11	-1.3:4.4

Note: the baseline set of predictors additionally includes the factor Type with three levels: S (Simple sentence), SE (sentence with a Single Embedded relative clause), and DE (sentence with Doubly Embedded relative clauses). Frequency counts are based on the 320-million HAL written corpus of US English. Prefix "s" indicates that the predictor was standardized (the mean subtracted from the raw value and the difference divided by 1 unit of standard deviation), e.g., sWordLength. Prefix "r" indicates predictors that were residualized prior to standardization due to considerations of statistical modeling, e.g., srNextFreq.

Table 4 Summary of main effects of skill measures on eye-movement measures, and of results of the mediation analysis, and the best-subset model selection (Appendix C). All main effects listed in Column 2 reached statistical significance at the 0.05 level: The polarity of the correlations between tests and a given dependent variable were identical across all tests and are reported in parentheses. Column 3 reports the three tests with the strongest direct effect on the dependent variable, sorted in the decreasing order. Column 4 reports three best subsets based on 18 tests, obtained using the exhaustive, backward, forward and sequential replacement algorithms: unless noted otherwise, the results were identical across algorithms for a given dependent variable.

1. Eye-movement measure	2. Skill measures	3. Mediation analysis	4. Best subset
FirstFixPos	span, rln, wid, piatr (+)	rln, span, wid	rln, wid, piatr
FirstFixDur	all, except digmem and nwrep (-)	wid, piatl, phonrev	wid, phonrev, piatl
SingleFixDur	all, except digmem and nwrep (-)	wid, piatr, piatl	wid, piatr, piatl
RefixProb	all, except digmem and nwrep (-)	rln, wid, span	–
GazeDur	all, except digmem and nwrep (-)	wid, piatr, piatl	wid, piatr, piatl (wid, piatl, watt in backward selection)
RegrProb	rdn, rln, piatl (-)	rln, nwrep, rdn	–

RegrPathTime	all, except nwrep (-)	wid, rln, rdn	wid, rln, rdn (wid, watt, rln in backward selection)
SecPassProb	segmentnw, phonrev, span, digmem, wid, watt, stan, piatr, piatl (+)	phonrev, wid, piatl	–
SecPassTime	digmem, nwrep, rdn, rln, ron, wid, watt (-)	wid, rdn, digmem	digmem, wid, rdn
TotalTime	digmem, nwrep, rdn, rln, wid, watt (-)	digmem, nwrep, rln	digmem, rln, rdn (digmem, wid, piatl in backward selection)

Table 5 Summary of interactions of skill measures by lexical properties of words N and N 1 as predictors of eye-movement durations, obtained via mediation analysis (Appendix C). All reported interactions are significant at the 0.05 level. Directions of coefficients for all reported interaction terms were identical within each type of interaction and are reported in parentheses.

1. Eye-movement measures	2. Test scores by word N length	3. Test scores by residual word N frequency (all +)
FirstFixPos	span, piatr, wid (+)	span, wid, piatr
FirstFixDur	wid, piatl, phonrev	rln, rdn, wid
SingleFixDur	wid, piatr, piatl (-)	rln, rdn, wid
GazeDur	wid, piatl, piatr (-)	wid, rln, piatr
RegrPathTime	wid, rdn, rln (-)	rdn, rln, wid
SecPassTime	wid, rdn, digmem (-)	–
TotalTime	wid, digmem, nwrep (-)	nwrep, rln, digmem

Table 6 Correlation matrix of tests. The Spearman correlation index q is indicated above the diagonal, while the p-value of the correlation below the diagonal

	elision	**blend**	**blendnw**	**segment**	**segmentnw**	**phonrev**	**span**	**digmem**	**nwrep**	**rdn**	**rln**	**rcn**	**ron**	**wid**	**watt**	**stan**	**piatr**	**piati**
elision	*****	0.842	0.837	0.938	0.931	0.918	0.809	0.325	0.538	0.816	0.834	0.763	0.704	0.882	0.905	0.83	0.797	0.7
blend	<0.001	*****	0.95	0.911	0.91	0.918	0.88	0.472	0.726	0.84	0.882	0.871	0.86	0.913	0.851	0.881	0.879	0.824
blendnw	<0.001	<0.001	*****	0.911	0.899	0.913	0.869	0.414	0.726	0.795	0.831	0.861	0.825	0.891	0.822	0.859	0.856	0.818
segment	<0.001	<0.001	<0.001	*****	0.988	0.95	0.879	0.352	0.58	0.8	0.845	0.844	0.767	0.929	0.887	0.929	0.896	0.826
segmentnw	<0.001	<0.001	<0.001	<0.001	*****	0.965	0.902	0.317	0.558	0.777	0.827	0.862	0.789	0.94	0.901	0.938	0.914	0.849
phonrev	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.919	0.378	0.598	0.845	0.877	0.884	0.843	0.962	0.94	0.921	0.905	0.844
span	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.349	0.576	0.804	0.845	0.938	0.922	0.953	0.87	0.892	0.928	0.882
digmen	0.188	0.048	0.088	0.152	0.2	0.122	0.156	*****	0.495	0.605	0.597	0.346	0.438	0.392	0.238	0.296	0.34	0.361
nwrep	0.021	0.001	0.001	0.012	0.016	0.009	0.012	0.037	*****	0.702	0.697	0.666	0.687	0.593	0.522	0.569	0.573	0.512
rdn	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.008	0.001	*****	0.979	0.73	0.756	0.882	0.83	0.767	0.798	0.728
rln	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.009	0.001	<0.001	*****	0.804	0.798	0.906	0.862	0.804	0.824	0.741
rcn	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.159	0.003	0.001	<0.001	*****	0.953	0.891	0.853	0.833	0.836	0.764
ron	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.069	0.002	<0.001	<0.001	<0.001	*****	0.885	0.815	0.797	0.833	0.789
wid	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.108	0.009	<0.001	<0.001	<0.001	<0.001	*****	0.928	0.94	0.959	0.908
watt	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.341	0.026	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.84	0.827	0.733
stan	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.233	0.014	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.97	0.924
piatr	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.168	0.013	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*****	0.967
piatl	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.14	0.03	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	*****

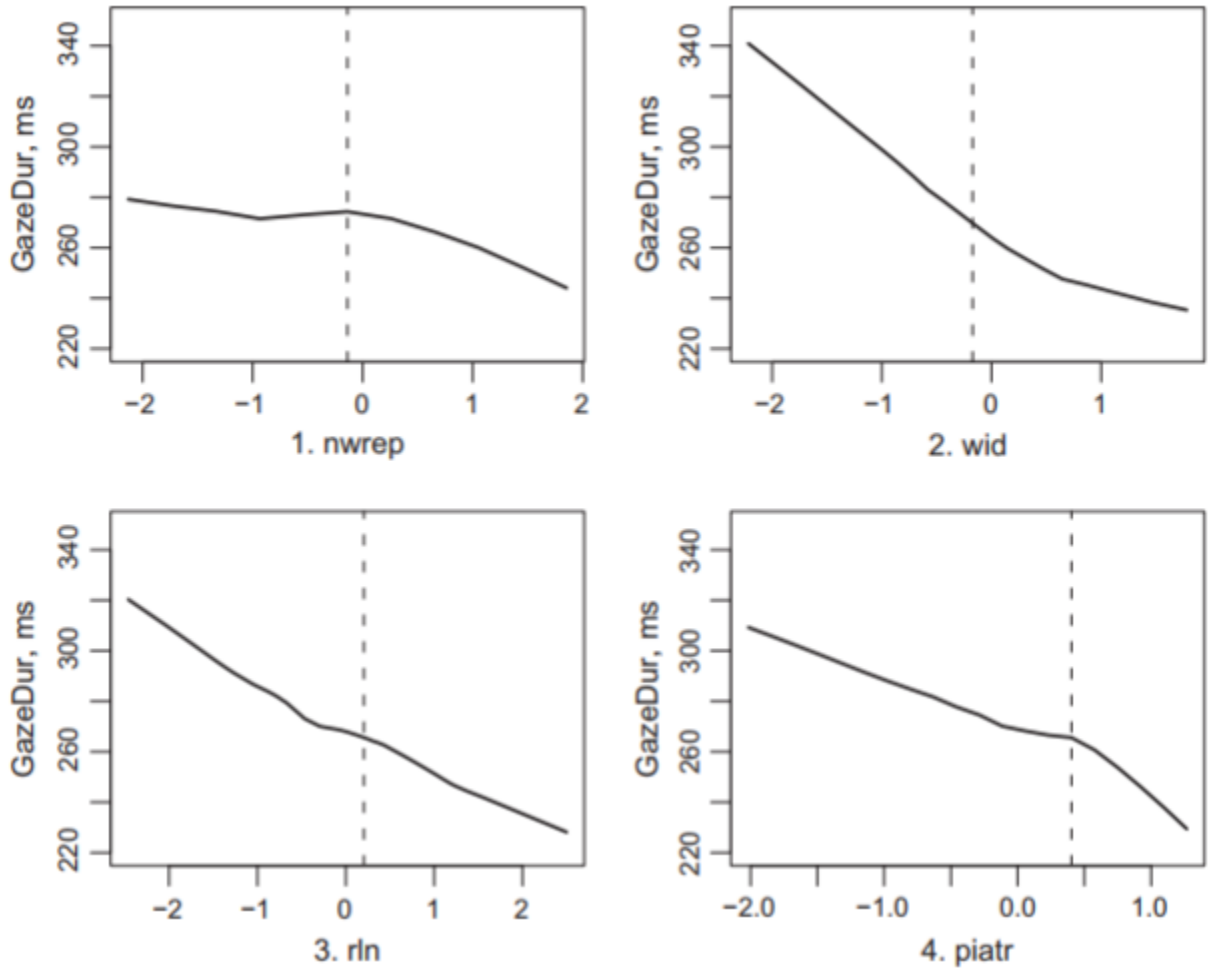

Fig. 1. Scatterplots of gaze duration as a function of the participant scores in (panel 1) the non-word repetition task (nwrep), (panel 2) word identification task (wid), (panel 3) rapid letter naming task (rln), and (panel 4) the Peabody Individual Achievement Test, reading version (piatr). The trend lines are lowess (locally weighted smoother) lines. Vertical dashed lines indicate the median values of respective scores. The scores of nwrep (panel 1) are only predictive of gaze duration for higher scorers, while scores in other tests (panels 2–4) correlate with gaze duration across the entire range of participants.

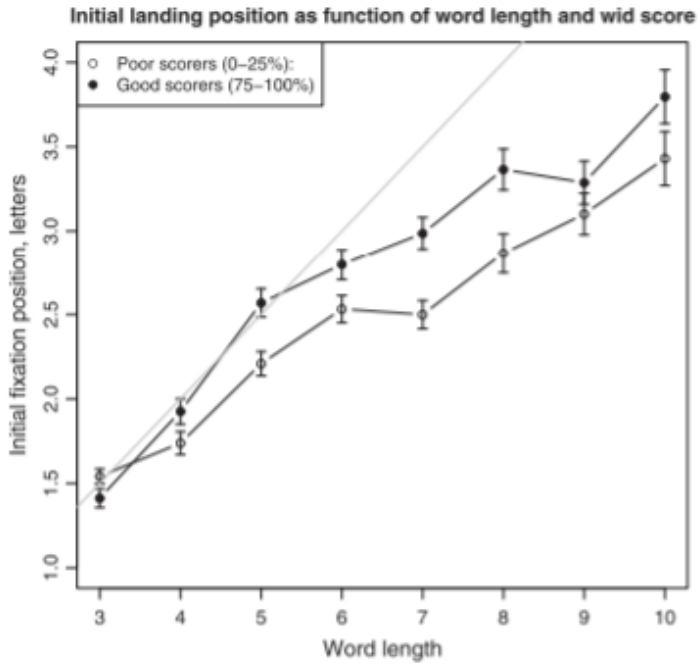

Fig. 2. Mean initial landing positions per word length, computed for the bottom fourth and the top fourth of the scorers in the word identification task. Word lengths 2–3 are binned together as 3, and lengths 10–13 are binned as 10. The gray line indicates the position at the center of the word. Error bars represent 1 standard error.

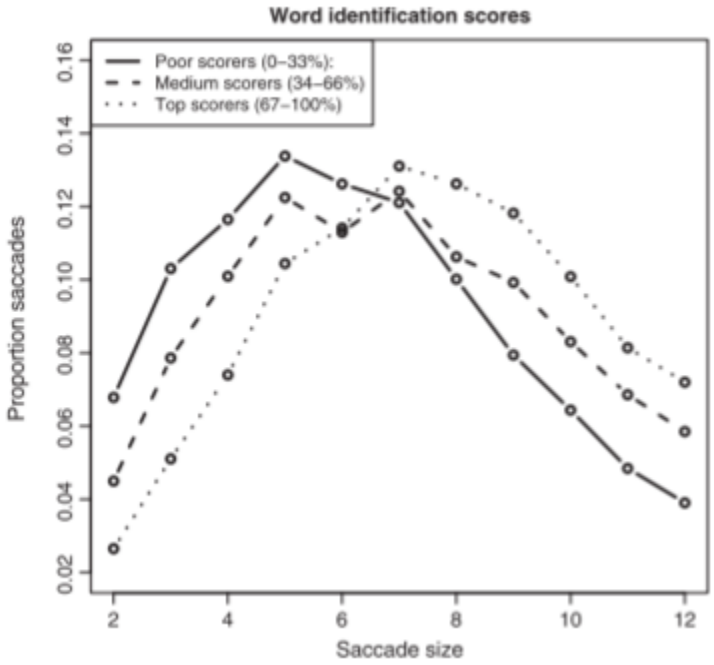

Fig. 3. Distribution of saccade sizes in letters for the bottom, middle and top thirds of scorers in the word identification test

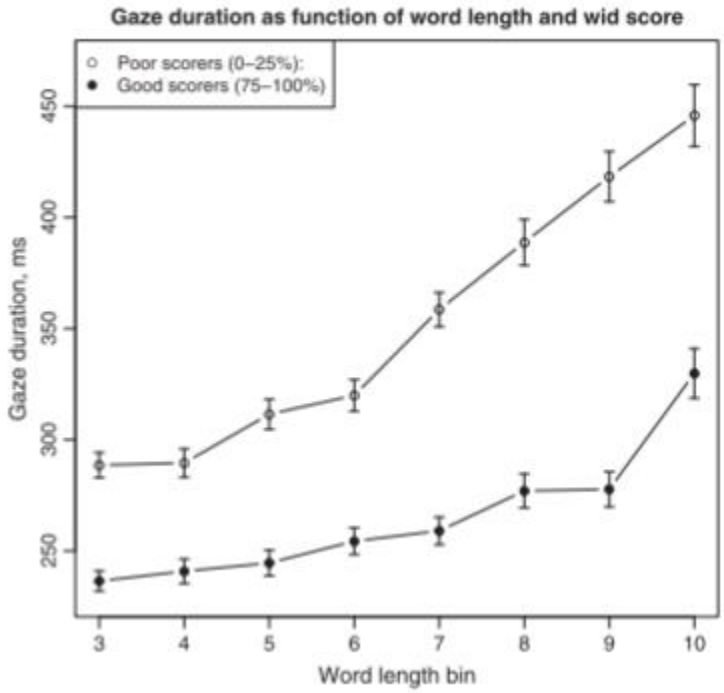

Fig. 4. Mean gaze durations per word length, computed for the bottom fourth and the top fourth of the scorers in the word identification task. Word lengths 2–3 are binned together as 3, and lengths 10–13 are binned as 10. Error bars represent 1 standard error.

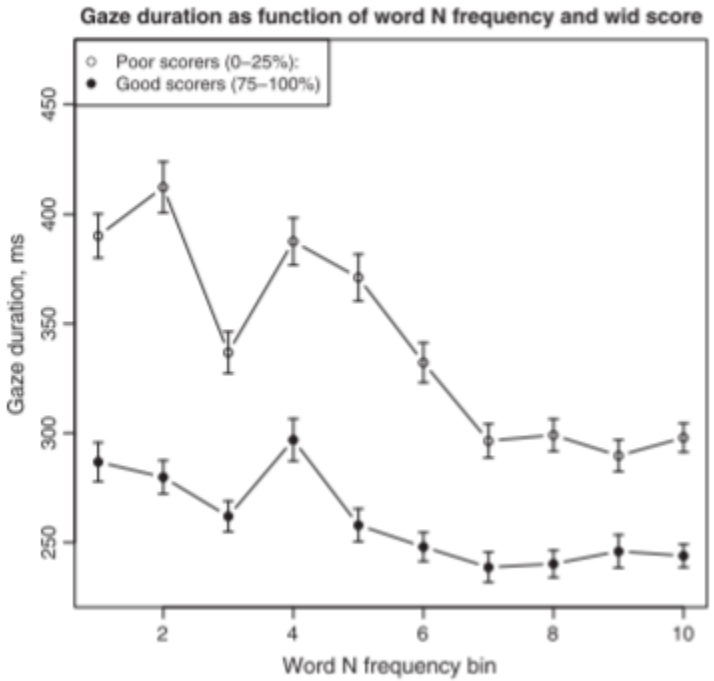

Fig. 5. Mean gaze durations per residualized log-transformed word frequency, computed for the bottom fourth and the top fourth of the scorers in the word identification task. Word frequencies were grouped into 10 bins. Error bars represent 1 standard error.

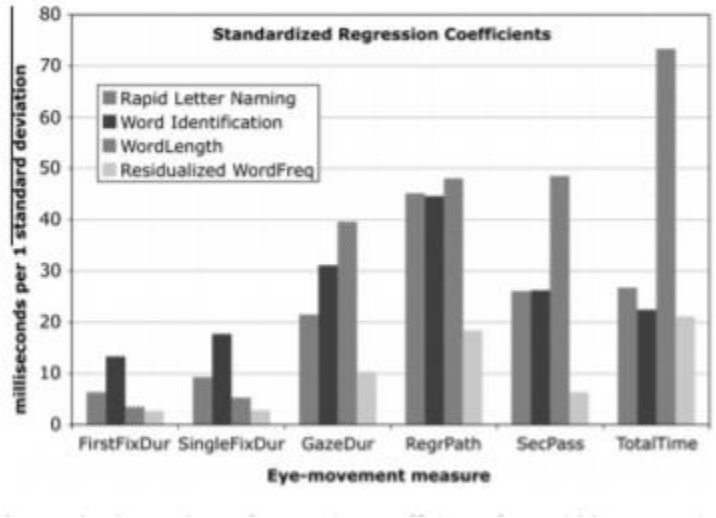

Fig. 6. Absolute values of regression coefficients for rapid letter naming (rln) scores, word identification (wid) scores, word N length and residualized frequency, as estimated by models fitted to durational eye movement measures. All regression coefficients are for standardized predictors and indicate the (absolute) change in the dependent variable (in ms) per one unit of SD in the predictor. Individual differences (participant) variables are indicated by solid fill; text variables are indicated by patterned fill.