INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Check for updates

Using Eye-Tracking Measures to Predict Reading Comprehension

Diane C. Mézière
Department of Psychology, University of Turku, Turku, Finland

Lili Yu

Erik D. Reichle
Macquarie Centre for Reading, Macquarie University, Sydney, New south Wales, Australia
School of Psychological Sciences, Macquarie University, Sydney, New south Wales, Australia

Titus von der Malsburg
Institute of Linguistics, University of Stuttgart, Stuttgart, Germany

Genevieve McArthur
Macquarie Centre for Reading, Macquarie University, Sydney, New south Wales, Australia
The Australian Centre for the Advancement of Literacy, Australian Catholic University, Sydney, New south Wales, Australia

ABSTRACT

This study examined the potential of eye-tracking as a tool for assessing reading comprehension. We administered three widely used reading comprehension tests with varying task demands to 79 typical adult readers while monitoring their eye movements. In the *York Assessment of Reading for Comprehension* (YARC), participants were given passages of text to read silently, followed by comprehension questions. In the *Gray Oral Reading Test* (GORT-5), participants were given passages of text to read aloud, followed by comprehension questions. In the sentence comprehension subtest of the *Wide Range Achievement Test* (WRAT-4), participants were asked to provide a missing word in sentences that they read silently (i.e., a cloze task). Linear models predicting comprehension scores from eye-tracking measures yielded different results for the three tests. Eye-tracking measures explained significantly more variance than reading-speed data for the YARC (four times better), GORT (three times better), and the WRAT (1.3 time better). Importantly, there was no common strong predictor for all three tests. These results support growing recognition that reading comprehension tests do not measure the same cognitive processes, and that participants adapt their reading strategies to the tests' varying task demands. This study also suggests that eye-tracking may provide a useful alternative for measuring reading comprehension.

The term "reading comprehension" is commonly used by reading researchers to refer to the sum total of processes that support the understanding of the meaning of a text, and the mental representations that are the product of those processes (Kintsch, 1998; LaBerge & Samuels, 1974; Perfetti & Stafura, 2014; Van Dyke, 2021). These processes and products likely include the perceptual, mental, and motoric operations and representations that are needed to understand individual words, constituents, phrases, sentences, and larger units of discourse (for a review of the computer models that have been developed to simulate and explain these operations, see Reichle, 2021).

Models of reading comprehension typically aim to explain reading comprehension as a process (e.g., Kintsch, 1988; Kintsch & van Dijk, 1978) or understand the cognitive processes and skills that support and explain variance in reading comprehension accuracy (e.g., Ahmed et al., 2016; Cromley et al., 2010; Cromley & Azevedo, 2007; Gough & Tunmer, 1986; Kim, 2017, 2020a, 2020b). Variance models of reading comprehension commonly assume that successful reading comprehension is supported by word reading skills (i.e., the ability to identify and sound out individual words) and oral language comprehension skills (i.e., higher-level cognitive skills necessary—such as syntax and inferences—to build a coherent representation of individual clauses, sentences, and passages of texts; Ahmed et al., 2016; Cromley & Azevedo, 2007; Cromley et al., 2010; Gough & Tunmer, 1986; Kim, 2017, 2020a, 2020b).

INTERNATIONAL LITERACY ASSOCIATION
Reading Research Quarterly, 58(3)
pp. 425–449 | doi:10.1002/rrq.498
© 2023 The Authors. *Reading Research Quarterly* published by Wiley Periodicals LLC on behalf of International Literacy Association. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Although models of reading comprehension vary in their complexity and the cognitive skills they include, reading comprehension is typically understood to rely on the fluid and coordinated operation of a large number of supportive cognitive processes. These processes include lower-level processes necessary for successful word identification (e.g., lexical access), as well as higher-level processes necessary for sentence- and discourse-level processing (e.g., syntactic processing, making inferences). If one then considers the actual *measurement* of reading comprehension using traditional measures, additional processes are introduced such as the capacity to remember the contents of a text, look for information in a text, or make predictions. This capacity, in turn, is influenced by motivation and willingness to exert effort to both generate and then "reconstruct" the meaning of a text. Critically, the relative importance of these processes that support reading comprehension has been shown to vary with development (e.g., Tilstra et al., 2009), text characteristics (e.g., Kim & Petscher, 2021), and assessment methods (e.g., Collins et al., 2019). This study focuses on the latter, and investigates the cognitive processes engaged by various reading comprehension measures using eye movements.

Given the aforementioned complexities associated with reading comprehension and its measurement, one might gain new appreciation for LaBerge and Samuels' (1974, p. 320) observation that "the complexity of the comprehension operation appears to be as enormous as that of thinking in general." This complexity, in turn, renders the development of accurate and reliable measures of reading comprehension particularly challenging. Indeed, recent research has brought the validity and reliability of standard methods of measuring reading comprehension into question, which has led researchers to investigate possible alternative ways of measuring reading comprehension (e.g., Sabatini et al., 2013, 2019). The work presented here falls within this line of research, and considers the possibility of using eye-movement behavior as an alternative to standard measures of reading comprehension.

The tracking of eye movements is widely used in reading research as it provides a non-invasive, online, and ecologically valid method of investigating the reading process at the word, sentence, and discourse levels (Rayner, Chace, et al., 2006). Unlike standard "pen and paper" measures of reading comprehension, eye-tracking does not require an overt comprehension task (e.g., answering comprehension questions), and it may therefore be possible to develop an online and ecologically valid measure of reading comprehension ability based on eye-movement behavior during reading. To date, while there is a plethora of research on eye-movement behavior during reading showing that there is a relationship between eye movements and reading comprehension processes (Rayner, Chace, et al., 2006), most studies have manipulated linguistic variables to investigate their effect on eye-movement behavior. Hence, the *predictive* relationship between eye-movement behavior during normal reading and reading comprehension is not yet well understood. To start to understand this relationship, it is important to first consider what is already known about (a) reading comprehension measures and their validity, (b) how eye movements can be used to investigate the processes involved in reading, and (c) the existing evidence for the predictive relationship between eye-movement measures and reading comprehension accuracy. We review this knowledge in the following three sections.

Reading Comprehension Measures and Their Validity

Previous studies have suggested that different tests of reading comprehension tax different cognitive skills to different degrees (Keenan et al., 2008; Kendeou et al., 2012). An early study by Nation and Snowling (1997) compared the performance of 184 children with typical development on two reading comprehension tests: the *Neale Analysis of Reading Ability* (*NARA*) and the *Suffolk Reading Scale*. In the NARA, children read passages of text aloud followed by comprehension questions. In the Suffolk Reading Scale, children were given a multiple-choice sentence completion task (i.e., cloze task). In a series of regression analyses, they found that word reading skills explained a significant amount of variance for both tests. However, oral language comprehension accounted for a significant amount of unique variance over and above word reading accuracy for the NARA. In contrast, it accounted for less than a 1% increase in explained variance for the Suffolk Reading Scale. This difference in the extent to which the two tests measure oral language comprehension might be attributed (at least in part) to differences in task demands between the two tests. Indeed, sentence completion tasks (i.e., cloze tasks) have been found to relate more strongly to word reading skills compared to comprehension questions (e.g., Spear-Swerling, 2004).

In a later study, Cutting and Scarborough (2006) investigated the relative contributions of word reading, oral language comprehension, as well as other cognitive skills (e.g., reading speed, attention) on three measures of reading comprehension with varying task demands. They tested 97 children on three widely used reading comprehension tests: the *Wechsler Individual Achievement Test* (*WIAT*), the *Gates-MacGinitie Reading Test—Revised* (*GM-R*), and the *Gray Oral Reading Test-3* (*GORT-3*). The tests differed from each other on three characteristics: (a) reading modality (aloud: GORT-3; silent: WIAT and GM-R); (b) task (multiple-choice question: GORT-3 and GM-R; open-ended questions: WIAT); and (c) availability of the text during the task (available: GM-R and WIAT; taken away: GORT-3). They found that word reading and oral language accounted for varying amounts of unique variance,

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and a large and significant amount of shared variance in test scores. Furthermore, the unique contribution of two aspects of oral language skills, lexical (e.g., vocabulary) and sentence-processing skills (e.g., syntax), varied between the three tests. Lexical skills accounted for unique variance in the GORT-3, sentence-processing skills accounted for unique variance in the WIAT, and both lexical and sentence-processing skills accounted for unique variance in the GM-R. These results suggest that test characteristics such as reading modality and question format can also influence the cognitive skills measured by reading comprehension tests.

In a recent study, Keenan et al. (2008) tested 510 children on four reading comprehension assessments (the *Peabody Individual Achievement Test*, *PIAT*; the *Qualitative Reading Inventory-3*, *QRI-3*; the GORT-3; and the *Woodcock-Johnson Passage Comprehension test*, *WJPC*), as well as oral language comprehension and word and nonword reading. The four tests were chosen such that they differed in reading modality (aloud: GORT-3 and QRI-3; silent: PIAT and WJPC), text length (sentences: PIAT and WJPC; passages: GORT-3, QRI-3, and WJPC), and comprehension task (picture selection: PIAT; cloze task: WJPC; multiple-choice question: GORT-3; open-ended questions and retell: QRI-3). They found that word and nonword reading explained the most unique variance in reading comprehension scores for the PIAT and WJPC, but that oral language comprehension was a better predictor for the GORT-3 and QRI-3. These findings further suggest that multiple characteristics of reading assessments, such as text length (e.g., sentence vs. passage) and format (e.g., reading aloud vs. silently), can influence the cognitive skills measured by reading comprehension assessment. To the best of our knowledge, similar studies on the cognitive skills measured by reading comprehension tests have not been carried out with adults, with the exception of studies investigating the impact of passage-independent questions (i.e., questions that can be answered correctly without the text) on the construct validity of reading comprehension tests for university students (Powers & Wilson Leung, 1995; Roy-Charland et al., 2017).

It has been suggested that differences in task demands between reading comprehension tests could impact performance on these tests (Andreassen & Bråten, 2010; Best et al., 2008; Collins et al., 2019; Davey & Lasasso, 1984; In'nami & Koizumi, 2009; Ko, 2010; Shohamy, 1984; Wolf, 1993), which, in turn, could explain why correlation coefficients differ across tests and studies (e.g., 0.64 to 0.79, Cutting & Scarborough, 2006; 0.75, Nation & Snowling, 1997; 0.31 to 0.70, Keenan et al., 2008; and 0.45 to 0.68 Keenan & Meenan, 2014).

More importantly, the fact that there is task variance in the comprehension tests also means that different cognitive skills and cognitive processes are engaged by and support reading comprehension in the various tests. This is reflected by the variance in the explanation power of the different cognitive skills. For example, oral language comprehension seems to be a more important predictor for comprehension tests using reading aloud than reading silently as the task modality (Keenan et al., 2008; Nation & Snowling, 1997); in contrast, nonword reading (Keenan et al., 2008) or sentence-processing skills (Cutting & Scarborough, 2006) explain more variance in silent than oral reading task. In an ideal world, it would be possible to measure reading comprehension without the confounding influence of such task demands. This may be possible by tracking readers' eye movements while they read text. Indeed, while eye movements are not entirely insensitive to differences in task demands (e.g., Kaakinen & Hyönä, 2010), eye-tracking does not require overt comprehension responses (e.g., comprehension questions) such that it may be possible to develop a test that only requires readers to read naturally with no additional comprehension tasks, thereby diminishing the effects of task demands. In this study, we investigate the cognitive skills and processes engaged and measured by reading comprehension assessments with varying task demands. In addition, we examine the relationship between eye-movement behavior during reading comprehension tasks and reading comprehension scores and explore the possibility of using eye movements to predict reading comprehension accuracy.

Using Eye Movements to Investigate Reading Processes

The tracking of eye movements is a widely used, noninvasive method to index online cognitive processing during reading (Rayner, Chace, et al., 2006). Eye-tracking provides two primary measures of eye-movement behavior: *saccades* (i.e., rapid ballistic movements of the eyes from one viewing location to the next) and *fixations* (i.e., the pauses between saccades where the eyes are relatively stationary). Although most saccades move the eyes forward through a text, 10–15% of saccades are *regressions*, which move the eyes back to a previous part of the text.

According to the *cognitive-control hypothesis* (e.g., Rayner & Reingold, 2015), eye-movement measures can be used to index the cognitive processing of linguistic properties of words or texts, such as a sentence's syntactic complexity (Staub, 2010), regions of lexical or syntactic ambiguity (Leinenger et al., 2017; Sturt, 2007), and word frequency and predictability (Rayner et al., 2011; Schilling et al., 1998). Eye-movement measures have thus been used to investigate cognitive skills and processes involved in reading comprehension, including processes that support word reading (e.g., lexical processing: Rayner et al., 2011; Schilling et al., 1998) and oral language comprehension (e.g., making inferences: Cunnings et al., 2014; Kreiner et al., 2008; Sturt, 2003). In addition, eye-movement

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

measures have been shown to reflect individual differences in overall reading skills (e.g., children vs. adults: Mancheva et al., 2015; Reichle et al., 2013; dyslexia: Hyönä & Olson, 1995; Jones et al., 2007), as well as cognitive skills that support reading comprehension such as word reading (e.g., Kuperman et al., 2018; Kuperman & Van Dyke, 2011), making inferences (e.g., pronoun resolution: Eilers et al., 2018; Murray & Kennedy, 1988), or working memory capacity (e.g., Clifton et al., 2003; Kuperman & Van Dyke, 2011). Eye movements therefore provide an online measure to investigate the cognitive processes that support reading comprehension, as well as individual differences in reading comprehension ability during natural reading.

Eye-movement measures can be divided into "global" and "local" measures. *Global measures* are aggregated over regions within sentences or multiple sentences that form texts. Some typical global measures include *mean fixation duration* (i.e., mean duration of all fixations in a sentence or text) and *mean saccade length* (i.e., mean length of all saccades in a sentence or text). Global measures are primarily informative about overall reading behavior such as text-level effects of linguistic manipulations (e.g., overall passage difficulty; Rayner, Chace, et al., 2006), or general individual differences between groups of readers with varying reading skills (e.g., children: see Blythe & Joseph, 2012, for a review; older adults: Rayner, Reichle, et al., 2006; dyslexic readers: Hyönä & Olson, 1995). For example, studies show that reading more difficult texts results in longer average fixation durations (Rayner, Chace, et al., 2006), or that less skilled readers tend to make more and longer fixations compared to skilled readers (e.g., children vs. adults; Blythe & Joseph, 2012; Reichle et al., 2013) with longer fixations indicating longer processing times and/or processing difficulties. Although global measures are informative about overall reading behavior and differences in processing, they are not very informative as to the specific cognitive processes that are affected by linguistic manipulations or reading ability. For example, children's longer average fixation durations suggest that they require additional processing time compared to adult readers, but does not tell us whether these longer processing times result from differences in low-level cognitive processes (e.g., lexical access), higher-level processes (e.g., syntactic processing), or non-linguistic processes (e.g., working memory capacity). To investigate such cognitive processes more directly, researchers typically calculate *local measures*.

Local measures focus on smaller units of text, usually single words. These word-level measures can be further divided into "early" measures that reflect rapid processes involved in reading, such as lexical access, versus "late" measures that reflect subsequent reading processes, such as syntactic integration (Clifton et al., 2007; Vasishth et al., 2013). These word-level measures are typically used to investigate specific cognitive processes, typically by manipulating word-level linguistic variables. For example, one of the most robust findings in the eye movement literature is the *word frequency* effect (i.e., words that occur more frequently in text tend to receive shorter fixations than less frequent words; Schilling et al., 1998), which can be interpreted as indicative of the ease of lexical processing with longer fixations reflecting lexical processing difficulty. This effect typically appears early (e.g., first fixation on a word) indicating that word frequency affects early cognitive processes in reading. Importantly, these word-level effects can also be indicative of individual differences in processing. For example, studies have shown that the word frequency effect is larger for children with dyslexia (Jones et al., 2007) and can be impacted by task demands (Kaakinen & Hyönä, 2010), thus reflecting differences in lexical processing between readers and tasks. Note that these dichotomies between global/local and early/late measures are not strict; however, because word-level measures have been used to study post-lexical integration (e.g., Warren et al., 2009) and other higher-level linguistic variables (e.g., violations of semantic plausibility; Rayner et al., 2004; Warren & McConnell, 2007), as well as non-linguistic processing (e.g., gender stereotypes; Sturt, 2003).

In sum, eye-movement measures can be used to successfully investigate the cognitive processes that support reading comprehension, as well as individual differences in these processes and their relation to successful reading comprehension. The combination of both global and local measures of eye-movement behavior provides a more detailed picture of readers' cognitive processing during natural reading, and can therefore be used to investigate the predictive relationship between eye-movement behavior and reading comprehension accuracy. In addition, the relationship between individual local measures and reading comprehension scores from assessments with varying task demands will be informative as to the cognitive processes engaged by and therefore measured by standardized reading comprehension assessments. We expect that the importance of individual eye-movement measures as predictors of comprehension scores across assessments will be informative as to both differences and similarities in cognitive processes across tasks, and aim to identify the possible common useful predictors across tasks. In this study, we investigate whether eye-movement measures can successfully predict reading comprehension accuracy. Specifically, we investigate the usefulness of measures indicative of both overall reading behavior (i.e., global measures) and individual cognitive processes that support reading comprehension (i.e., local measures). While it is likely that differences in task demands between tasks will be apparent in the relative importance of individual predictors, it is also

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

plausible that one or more predictors will be commonly useful across tasks.

Predicting Comprehension Accuracy from Eye-Movement Behavior

Predicting reading comprehension ability from eye-movement behavior is no easy task. Indeed, while the relationship between reading comprehension and eye-movement behavior is well established, most eye-movement studies of reading comprehension to date have systematically manipulated linguistic variables (e.g., syntactic complexity) to determine their effect on eye-movement measures. Few studies have directly investigated how eye-movement behavior relates to reading comprehension accuracy, with varying and sometimes conflicting results. For example, some studies on the relationship between eye-movement patterns and reading comprehension suggest that more efficient eye-movement behavior (e.g., shorter fixations, fewer regressions) tend to be associated with better reading comprehension (e.g., Kim et al., 2019; Parshina et al., 2022). On the other hand, studies of the relationship between regressions and comprehension accuracy find the opposite pattern, suggesting that making more regressions is associated with better reading comprehension (Schotter, Tran, & Rayner, 2014; Wonnacott et al., 2016). And still other researchers find no relationship (Christianson et al., 2017). To date, eye-movement markers of successful reading comprehension have not yet been clearly identified.

Early attempts at using eye movements to predict reading comprehension accuracy come from studies using machine learning methods (e.g., neural networks) to investigate the potential of eye-gaze data to predict performance on comprehension assessed during or immediately after reading. Copeland et al. (2014) had participants read short slides of text from a university course tutorial followed by two comprehension questions (one multiple choice, one cloze task). They used artificial neural networks to predict performance on the comprehension questions from multiple global eye-movement measures (e.g., average fixation duration), with 79–89% accurate classification rate. Although this type of method does not allow for a clear link to be established between specific eye-movement measures and comprehension, the results do suggest that eye movements can be used to successfully predict comprehension scores (see also Copeland et al., 2016; Copeland & Gedeon, 2013; Martínez-Gómez & Aizawa, 2014).

Inhoff et al. (2018) investigated the predictive relationship between subsets of eye movements and comprehension ability more directly. They collected eye-movement data from 37 participants reading single sentences. Comprehension was measured separately with comprehension questions (yes/no answers) following filler items (not included in the eye-movement data), and multiple-choice questions following a short text. They grouped the eye-movement measures into two latent variables to reflect two processes of reading: "acquisition" (i.e., early measures such as first-fixation duration) and "correction" (i.e., measures of returning to previous parts of the text to correct reading or parsing errors, such as regression rate). They found that the correction variable best predicted comprehension, and that acquisition was correlated with correction but had no direct effect on comprehension. These results further suggest that eye-movement measures can be used to predict comprehension scores, and that some eye-movement measures may be more useful in predicting comprehension than others.

In a recent study, Southwell et al. (2020) predicted reading comprehension scores in three datasets using global eye-movement measures. In all three datasets, participants were given long passages of text to read silently, followed by multiple-choice questions. To predict comprehension, they fit linear models with cross-validation (i.e., the dataset was partitioned, and the model run on part of the data then used to predict the left-out data). Prediction accuracy was calculated as the correlation between the model-predicted scores and the observed scores. For all three datasets, eye moments predicted comprehension scores with correlations ranging from 0.35 to 0.40. In addition, the relationship between eye movements and comprehension was highly similar across datasets, with more fixations and shorter fixations associated with better comprehension scores across datasets (see D'Mello et al., 2020 for similar results).

Taken together, these results suggest that eye movements can be used to successfully predict reading comprehension. However, most studies have either grouped measures into latent variables or focused only on global measures. Hence, it is unclear whether other individual local eye-movement measures (e.g., gaze duration) can also be useful in predicting reading comprehension accuracy. In addition, comprehension in these studies was typically measured with multiple-choice questions. Because differences in task demands have been shown to affect both performance on comprehension tasks and eye-movement behavior (Bax & Chan, 2019; Kaakinen & Hyönä, 2010; O'Reilly et al., 2018; Radach et al., 2008; Schotter, Bicknell, et al., 2014), it is important to investigate whether these results can be replicated across comprehension measures (e.g., open-ended questions). In this study, we examine whether eye-movement measures can successfully predict reading comprehension accuracy across reading assessments with varying task demands.[1] In addition, we examine the influence of task demands on reading behavior and the predictive relationship between eye-movement measures and reading comprehension accuracy.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The Present Study

As justified above, this study had three specific aims. The first is to investigate the cognitive processes engaged by and measured by reading comprehension assessments with different task demands. The second is to investigate whether eye-movement measures can successfully predict reading comprehension accuracy. The third is to examine the influence of task demands on reading behavior and the predictive relationship between eye-movement measures and reading comprehension accuracy. We addressed these aims by measuring global and local eye movement measures while 79 undergraduates completed three widely used reading comprehension tests for adults: The York Assessment of Reading for Comprehension (YARC), the GORT-5, and the sentence comprehension subtest of the Wide Range Achievement Test (WRAT-4). We chose this combination of tests because they are widely used in both clinics and research, and are representative of the different ways that reading comprehension assessments typically measure comprehension and thus of the differences in task demands usually found between standardized assessments of reading comprehension. We conducted two analyses of the comprehension test scores and eye movements. First, we calculated descriptive statistics of eye-movement behavior and ran correlations between eye movements and test scores to compare eye-movement behavior between the three tests. In line with previous studies, we expect that the three reading assessments do not measure the same cognitive skills to the same extent, and that participants may thus perform differently across the three tests. In addition, it is plausible that the strength of the correlations between eye movements and comprehension scores will vary across tasks. Second, we ran linear regression models within the Bayesian framework to investigate the predictive relationship between eye-movement measures and test scores. The models were then compared using cross-validation to identify the best predictors of performance on the three comprehension measures. Based on previous research, we expect that eye movement measures can be used to successfully predict reading comprehension accuracy in all three tests. In addition, we expect that differences in task demands between the three tests will influence participant's reading behavior, and may impact the usefulness of individual predictors across assessments. Nevertheless, we expect that one or more eye movement measures may be identified as a useful predictor of comprehension accuracy across the three tests. The results of these two analyses are discussed with regards to the cognitive processes measured by reading comprehension assessments, the usefulness of eye-movement measures as predictors of reading comprehension accuracy, and the influence of task demands on the relationship between eye-movement measures and reading comprehension accuracy. Finally, the implications of our findings for reading theories are discussed.

Method

Participants

In all, 79 undergraduate students participated in our experiment (65 females, mean age 22 years) for course credit, as approved by the Macquarie University ethics committee. The sample size for this study was determined based on previous research investigating individual differences with eye-tracking data (e.g., Kuperman et al., 2018; Staub, 2021). In all, 60 participants were monolingual native speakers of English, 8 were birth bilinguals with English as one of their native languages, and the remaining 11 were non-native speakers of English. Native and non-native speakers were included to ensure that we would have a range of reading comprehension abilities. All participants lived and studied in Australia at the time of testing.

Reading Comprehension Tests

The reading comprehension tests in this study were selected to be a representative sample of the most commonly used test formats and tasks for such tests. Specifically, we chose tests with differences in text length (sentences vs. passages), reading modalities (aloud vs. silent), availability of the test items (can vs. cannot return to the text), and comprehension tasks (questions vs. cloze procedure). The reasons behind choosing tests that vary in the way that they measure comprehension were two-fold. First, it allowed for possible differences in the cognitive skills measured by reading comprehension assessments with different task demands to be investigated in relation to both test scores and eye-movement behavior. Second, it allowed us to investigate the usefulness of eye-movement measures as predictors of comprehension scores across reading activities, and to uncover possible commonalities and/or differences in this relationship across tasks. The administration procedure for each test is described below, based on the test manuals.

Three existing reading comprehension tests were administered to participants while their eye movements were monitored: (a) the *YARC—Passage Reading Secondary*, Australian Edition (Snowling et al., 2009); (b) the *GORT*—5th edition (*GORT-5*; Wiederholt & Bryant, 2012); and (c) the word reading and sentence comprehension subtests of the *WRAT*—4th edition (*WRAT-4*; Wilkinson & Robertson, 2006). All three tests have two sets of forms for test–retest purposes. In this study, only items from the YARC Form A, GORT-5 Form A, and WRAT-4 Green Form were used.

For eye-tracking purposes, the test items were all presented on a computer screen. However, the tests were

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

administered and scored according to the test manual procedures with one exception: to obtain consistent eye-movement data across participants, the baseline and discontinue rules (e.g., stop participant after 7 consecutive wrong answers) were not employed during testing. Instead, each participant started with the baseline item recommended for adults and ended with the last item on the test. During scoring, participants were given full marks for all items prior to the baseline item. Additional information about scoring procedures and test reliability can be found in Data S1.

YARC

In this test, participants read two passages silently. The starting point for this test is based on participant's grades. Because our participants were adults, they were given passages 2.1 and 2.2. The passages were spread over four screen pages, and participants could move forward and backward between the pages (via buttons on a response box) during reading. At the end of each passage, participants were asked 13 open-ended comprehension questions about the text, and were able to return to the text to answer them. In addition, participants were given a summary question at the end of each passage but were not able to return to the text for this question. All answers were transcribed before they were scored. For this test, participants' eye movements were tracked both while they read the passages and while they answered the questions. The comprehension questions were scored for accuracy, with a maximum score of 13. The summary was scored separately, based on the number of key events from the text participants provided in their summary. Reading time was calculated as the time participants took to read the text. Final scores for reading time,[2] comprehension, and summary were obtained through the YARC online score conversion tool (https://rgt.testwise.net/YARC_Aust_Pri_index.htm). This tool provided standard scores only for comprehension and reading time. Standard scores have a mean of 100 and a standard deviation of 15. The authors of the test manual indicate that correlations between comprehension scores and summary scores in their sample were low, and that the summary scores may not be entirely reliable. For this reason, we chose not to include the summary scores in our comprehension measures in later analyses.

GORT-5

This test comprises 16 passages of text that increased in length and difficulty as items progress. The passages are adapted from works of fiction and non-fiction. The starting item for each participant is based on their grade. Because all participants in this study were adults, they started at item 6 and continued to the final item 16.

Participants were instructed to read passages aloud as quickly and accurately as possible and then answered five open-ended questions. Each passage was presented on a computer screen over 1–3 pages. Participants used a button on a response box to move to the next page of text; they were not able to move backwards to a previous page of text.

Raw scores were calculated for each participant's reading time (i.e., how long it took them to read the passage), reading accuracy (i.e., total number of reading errors they made while reading aloud; e.g., incorrect pronunciations, hesitations), reading fluency (i.e., the sum of their reading rate and reading accuracy scores), and reading comprehension (i.e., the total number of comprehension questions answered correctly). Raw scores for time, accuracy, fluency, and comprehension were converted into scaled scores provided by the manual of the GORT-5, with those scores having a mean of 10 and standard deviation of 3.

WRAT-4

Participants were first given the word reading subtest of the WRAT-4, because the score on this test is used to determine participants' starting point in the sentence comprehension subtest. This subtest comprises of 55 words of increasing difficulty presented on a single screen that participants are instructed to read out loud. Participants were scored on the number of words that they could read correctly. This score was then used to determine the starting point on the sentence comprehension subtest for scoring purposes only.

The sentence comprehension subtest comprises 50 items of increasing difficulty. Each item consists of one or two sentences with one word missing. The sentence comprehension subtest starts with two example items to familiarize participants with the task. Participants were instructed to read the sentences carefully to themselves and say what they thought the missing word was. Each item was scored for accuracy according to the answers provided by the manual. Correct answers ranged from only one possible answer to "anything denoting concept X." An early starting point (starting point D) was chosen for all participants to ensure that participants read the same number of items, and for comparing participants' eye movements on the same items. Participants thus saw items 20–50. The raw total score was the sum of all correctly answered items, and ranged from 20 to 50, because all items prior to the starting point were scored as correct. The raw scores were then converted into standard scores provided by the test manual. These scores have a mean of 100 and a standard deviation of 15.

Eye-Tracking Procedure

All the texts were presented in Courier New font with a size of 24, and in black color on a gray background (RGB: 204, 204, 204) on a BenQ Zowie XL2540 screen with a screen resolution of 1920 × 1080 pixels and a refresh rate of 240 Hz. The items from the YARC and GORT (i.e.,

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

passages) were spread over the whole screen, and the items from the WRAT (i.e., sentences) were presented in the middle of the screen. Participants were instructed to press a button to move along the pages, and when they had finished reading. They were then asked to answer the comprehension questions. In the WRAT, the experimenter moved to the next trial as soon as the participant gave the missing word. The three tests were administered in random order.

Eye movements were recorded using an EyeLink 1000+ eye tracker (SR Research, Toronto, Ontario, Canada) located in a sound-proof lab. Participants were seated approximately 95 cm from the display screen such that each letter occupied approximately 0.24° of visual angle on the screen. The experimenter sat behind the participant to give instructions and ask the comprehension questions throughout the experiment. A headrest was used to minimize head movements.

Data Collection

A 9-point calibration process was used at the beginning of each test to ensure the tracking accuracy of participants' eye movements. Participants were also re-calibrated as necessary (e.g., if they moved, or if the calibration became poor) at the end of a given item. The maximum allowance for the calibration error for all points was 0.45°, with only one participant exceeding this cutoff with a maximum of 0.48°. This calibration process was repeated at the start of each reading test, and between the two items of the YARC test. Each test item also started with a drift correction point, placed at the very beginning of the first sentence. The eye-tracker collected fixation positions and durations. This information was then used to calculate various eye-movement measures for data analysis.

Data Pre-Processing

Tests

Items for which participants did not read the whole text were excluded from analysis and thus not scored. For these participants, final test comprehension scores could not be calculated accurately and were treated as missing data (7 GORT scores, and 2 WRAT scores). Scoring was done based on the test manual guidelines, as described above.

Eye Movements

The eye-movement data were pre-processed in Data Viewer (SR Research, Toronto, Ontario, Canada). We excluded participants and trials with poor calibration from this analysis based on visual inspection of the data. This resulted in the data exclusion of two participants, and a total loss of 7% of trials. In addition, all words around punctuation marks were excluded from analysis (except for calculating the wrap-up effect). For each participant, fixations shorter than 80 ms or longer than 800 ms were excluded. In addition, forward saccades longer than the perceptual span (20 characters; Rayner, 2009) were excluded from analysis (3% of all forward saccades). For the YARC test, only eye movements collected during the initial reading of the text was included; any eye movements collected while answering the questions were excluded from analysis.

Data Analysis

We ran two sets of analyses. Because the nature of this work is exploratory, we first investigated differences between the tests both in terms of the test scores and participants' eye movements using descriptive statistics and correlations. As the correlations between eye movements and test scores are descriptive and not used for making inferences, we only looked at the correlation coefficients and confidence intervals. In the second analysis, we used eye-movement measures to predict test scores. For both analyses, local eye-movement measures were first calculated for each word in the text and then aggerated over all words in the text prior to analysis. We included nine variables: (a) reading speed (i.e., number of words read per minute) and eight eye-movement measures: (b) average fixation duration (i.e., mean duration of all fixations in a given text); (c) average forward saccade length (i.e., mean length of all rightward saccades in a text, in character spaces); (d) first-pass skipping rate (i.e., the proportion of words skipped in a text during first-pass); (e) first-fixation duration (i.e., the duration of the initial fixation on a word); (f) gaze duration (i.e., the sum of all first-pass fixations on a word); (g) regression rate (i.e., the proportion of all regressions made in a text); (h) go-past time (i.e., the sum of fixations on a word up to when it is exited to its right, including all regressions to the left of the word); and (i) total-reading time (i.e., the sum of all fixation durations on a word). The latter three measures, contrary to first-pass measures (d, e, and f), are posited to reflect higher-level processes such as syntactic processing and the integration of word meanings. Although measures (d) to (i) were calculated based on each word within the tests, these measures were aggregated per participant per test. In addition, we calculated two linguistic effects on eye movements: (a) *word-frequency effects* on gaze duration (Schilling et al., 1998) and (b) *wrap-up effects* on total-reading time (i.e., words tend to be fixated longer when they are at the end of a clause or sentence than when they are in the middle; Just & Carpenter, 1980). The wrap-up effect is argued to reflect integration processes that occur at the end of clauses/sentences. In this study, the effect was calculated as the difference between total-reading time on words at the end of clauses compared and the average total-reading time of all "middle" words in a sentence (excluding the first word of a sentence). Because the items

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from the WRAT-4 contained missing words, which were often at the end of the sentence, we did not calculate the wrap-up effect for this test.

Transparency and Openness

We report how we determined our sample size and all data exclusions, manipulations, and measures in the study. The materials of this study are not available, as they are copyrighted. The data and analysis code for this study are available by contacting the corresponding author. All analyses were conducted in the R system for statistical computing, version 4.0.2 (R Core Team, 2020), and the packages *brms* version 2.15.2 (Bürkner, 2017, 2018), *lme4* version 1.1.26 (Bates et al., 2015), *tidyverse* version 1.3.1 (Wickham et al., 2019), and *patchwork* version 1.1.1 (Pedersen, 2020). This study was pre-registered on the Open Science framework. Pre-registration of the design and analysis plan can be found here: https://osf.io/d7apz. The analysis presented in this study deviates partly from the pre-registration on the Open Science framework, as the planned analyses were less suited to answer our research questions.

Results

Correlations Between Reading Comprehension Test Scores

In the first analysis, we investigated correlations between the three comprehension scores, calculated descriptive statistics, and looked at correlations between test scores and eye-movement measures. The results of correlations between comprehension scores showed that participants could receive quite different scores on the three tests, as shown in Figure 1. Specifically, while most participants received scores within the average range (i.e., within one standard deviation from the mean: 85–115) in each test, many participants performed differently in at least two of the tests. Of our 78 participants,[3] only 29 received scores in the same range for all three tests (37%). These participants all performed within the average range (85–115) on all tests except for one participant who performed below average on all tests (<85). Of the remaining 49 participants, 31 (63%) performed below the average range on one test but within the average range or higher on another.We investigated the differences in participants' test performance statistically by running Pearson *r* correlations between individuals' scores on each test. These correlations are shown in Figure 2. Participants' scores on the GORT were transformed from scaled scores into standardized scores to allow for easy visual comparisons between the three tests (mean = 100; standard deviation = 15). Summary statistics for the three tests are shown Table 1. All the correlations were statistically significant ($p < 0.001$) but moderate in strength (GORT-WRAT: $r = 0.62$; GORT-YARC: $r = 0.57$; YARC-WRAT: $r = 0.57$).

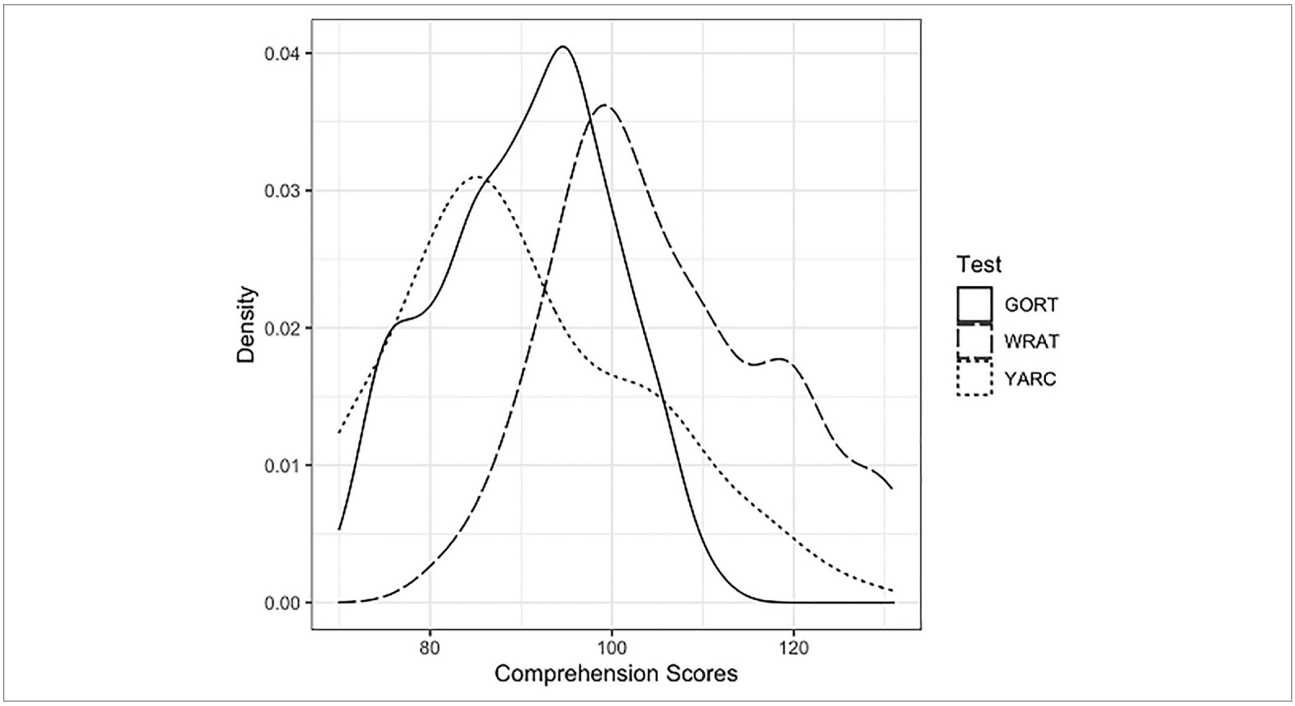
Differences in Eye Movements Between Tests

We then calculated descriptive statistics for the eye-movement measures. Table 2 shows the mean values for reading speed, the eye-movement measures, and word frequency. Values for the word-frequency effect on gaze duration are estimates of mixed linear models. In these models, gaze duration was log-transformed to control for differences between tests, and the effect of frequency was controlled for length, with a random effect for participants. The wrap-up effect was calculated as the difference between the average total-reading time on words within the sentence (excluding the first word of the sentence) and the total-reading time on words at the end of clauses and sentences. As we did not find evidence for a wrap-up effect, we do not discuss this effect further. We also investigated differences in eye-movement behavior between the three tests. For this purpose, we examined four types of eye-movement measures: global measures (average fixation duration and forward saccade length), first-pass measures (skipping rate, gaze duration, and first-fixation duration), late measures (regression rate, go-past time, and total-reading time), and word frequency. Participants tended to read at a slower pace in the GORT, with slower reading speed, longer fixations, and larger word frequency effects on gaze duration. This is in line with previous findings showing slower reading speed and longer fixations when reading aloud compared to silently (e.g., Ashby et al., 2012; Inhoff & Radach, 2014; Vorstius et al., 2014). The reading rates for all three tests are lower than the 248–260 wpm (silent reading, non-fiction and fiction respectively) and 183 wpm (oral reading) from a recent meta-analysis conducted by Brysbaert (2019). A possible explanation for this discrepancy is that reading rates may be slower when reading to answer questions and remember the text (Brysbaert, 2019), as readers may read more carefully in such cases. Notably, the estimates do not differ when only native speakers are included; hence, the lower reading rates cannot be attributed to the multilingual and non-native speakers in our sample. Eye-movement behavior in the YARC and WRAT were more similar to each other, although participants tended to read faster in the YARC, with higher skipping rates and longer go-past times.

Faster reading speed in the YARC, also reflected in higher skipping rates and fewer regressions, may result from the fact that readers could re-read the text while answering the questions. As such, contrary to the GORT, which required them to remember the text in detail, or the WRAT which required them to make predictions about upcoming linguistic material, the YARC did not require readers to

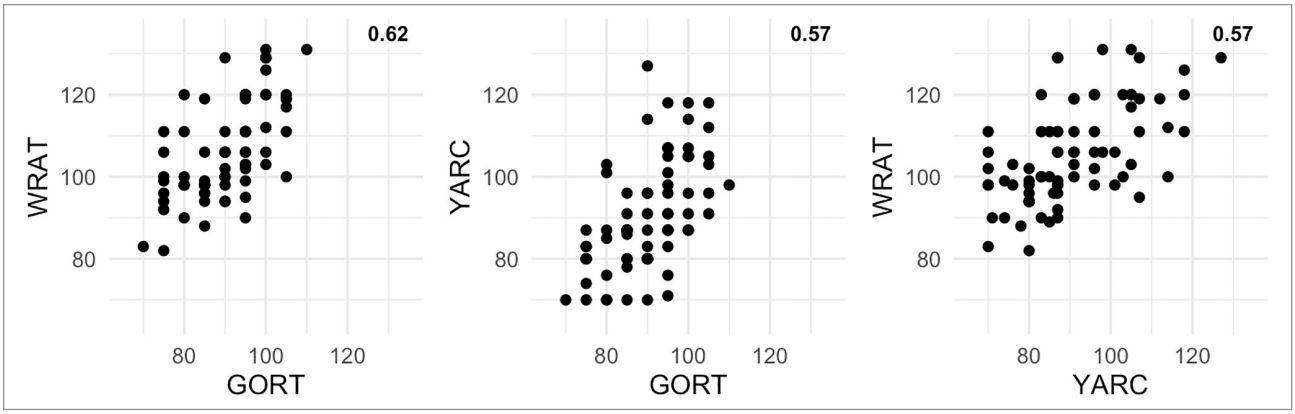
19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 1
Distribution of Comprehension Scores



Note. Figure 1 shows the distribution of comprehension scores for the three tests.

FIGURE 2
Correlations between Comprehension Scores



Note. Figure 2 shows the correlations between comprehension scores.

construct as detailed a representation of the text during the initial reading of the text, which may have led participants to use more shallow processing strategies of the text meaning (e.g., skimming or reading for the gist). The higher regression rates for the GORT and WRAT are also likely due to the specific task demands of these tests. The GORT required the text to be read aloud, leading to more regressions both within and between words in order to keep the eyes apace with the articulation of the text (i.e., to maintain the eye-voice span; Inhoff et al., 2011). On the other hand, the WRAT required readers to make predictions about a missing word in the sentence, for which readers sometimes had to read (part of) the sentence again; for example, to think of an appropriate answer or to check that a word that had been generated fit into the sentence. Hence, the differences that were observed in readers' eye-movement behavior during the three tests can be explained, at least in part, by the differences in task demands between the tests.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1
Summary Statistics of Comprehension Scores

Test	Mean (SD)	Range
YARC	90.25 (13.5)	70–127
GORT	90.35 (9.43)	70–110
WRAT	105.78 (12.07)	82–131

TABLE 2
Average Eye Movement Measures per Test

Measures	YARC	GORT	WRAT
Global			
Reading Speed (wpm)	208	142	157
Average Fixation Duration	230	253	236
Average Forward Saccade Length (chars)	8.9	7.5	8.8
First-Pass			
Skipping Rate	0.62	0.41	0.39
First-Fixation Duration	232	258	230
Gaze Duration	267	340	262
Late			
Regression Rate	0.15	0.20	0.19
Go-Past Time	509	556	468
Total-Reading Time	359	454	424
Linguistic Effects			
Word-frequency Effect on Gaze Duration (log)	0.0007	−0.018	−0.014

Note. wpm = words per minute; chars = characters; values for the word-frequency effect are model estimates.

Correlations Between Test Scores and Eye Movements

We investigated the relationship between test scores and eye movements by running a series of Pearson *r* correlations between participants' test scores and their eye movements while taking the tests. Because the comprehension scores were calculated for the whole test and per participant, all eye-movement measures were aggregated per test and participant. For the word frequency effects, participants' random slopes were extracted from the models to run the correlations with test scores, and these estimates were then aggregated across tests for the correlations with measures averaged across tests. As noted earlier, we did not find wrap-up effects and hence do not report correlations for this measure. All correlation coefficients are shown in Table 3. As these correlations were not used to make inferences but are intended to be descriptive in nature, only the correlation coefficients and confidence intervals are reported. Because the pattern of results in these correlations differed widely between tests, we also calculated the correlations between each participant's averaged score across tests and eye movements averaged across the three tests. In addition, we looked at the correlations between the standard deviation in the test scores and the averaged eye-movement measures to investigate whether reading comprehension ability was associated with consistency in eye-movement behavior.

Predicting Reading Comprehension Scores from Eye Movements

In the second analysis, we investigated whether eye movements could predict reading comprehension scores.[4] In this analysis, we first fitted linear regression models with reading speed and eye-movement measures as predictors of reading comprehension scores. Model assumptions were checked visually with the full model for each of the four datasets (Gelman & Hill, 2007). Then, we evaluated and compared these models to identify the subset of predictors that best predicted comprehension scores. Note that in this approach the usefulness of a predictor is determined by its presence in the best-performing models.

We fitted linear regression models within the Bayesian framework using the "brms" package (Bürkner, 2017, 2018) in *R*, and report predictors' effects based on 95% credible intervals[5]. We considered reading speed and eight eye-movement measures as predictors: mean fixation duration, mean forward saccade length, skipping rate, first-fixation duration, gaze duration, regression rate, go-past time, and total-reading time. We had no expectations about which subsets of predictors were most important for predicting comprehension, and so we ran a linear model for every possible subset of our nine predictors (512 models in total).[6] The number of predictors in the models therefore ranged from no predictors (i.e., the null model) to the full model with all nine predictors. All predictors were scaled such that the estimated effects would be directly comparable across variables (i.e., a model estimate of 4 is half an estimate of 8). We ran this set of models four times: once for each of the three tests, and once with data aggregated across the three tests (giving a total of 2048 models). Given the multicollinearity between measures, estimates in models with multiple predictors should be interpreted with caution and in relation to the other predictors in the model (see Discussion).

Within each set of 512 models individual models were evaluated and compared using *leave-one-out* cross-validation (*LOO*; Gelman et al., 2014; Vehtari et al., 2017). The LOO estimates a model's ability to predict new data by fitting the model as many times as there are data points, leaving out a different data point each time, and then evaluating how well the left-out data point is predicted by the

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3
Correlations between Average and Standard Deviations of Test Scores and Eye Movements

	YARC		GORT		WRAT		Mean score - mean EM		SD scores - mean EM	
Measures	*r*	95% CI	*r*	95% CI	*r*	95% CI	*r*	95% CI	*r*	95% CI
Global										
Reading Speed	0.23	[0.01, 0.43]	0.30	[0.08, 0.50]	0.57	[0.40, 0.71]	0.48	[0.29, 0.64]	0.24	[0.01, 0.44]
Average Fixation Duration	−0.17	[−0.38, 0.05]	−0.11	[−0.34, 0.13]	−0.28	[−0.47, −0.05]	−0.22	[−0.42, 0.01]	−0.11	[−0.33, 0.12]
Average Saccade Length	0.15	[−0.07, 0.36]	0.41	[0.19, 0.59]	0.25	[0.03, 0.45]	0.32	[0.10, 0.50]	−0.02	[−0.24, 0.20]
First-Pass										
Skipping Rate	−0.03	[−0.25, 0.19]	0.09	[−0.15, 0.32]	0.16	[−0.07, 0.37]	0.07	[−0.15, 0.29]	−0.01	[−0.23, 0.22]
First-Fixation Duration	−0.19	[−0.40, 0.04]	−0.07	[−0.30, 0.17]	−0.26	[−0.46, −0.04]	−0.21	[−0.42, 0.02]	−0.09	[−0.31, 0.14]
Gaze Duration	−0.26	[−0.46, 0.04]	−0.35	[−0.55, −0.13]	−0.29	[−0.48, −0.06]	−0.33	[−0.52, −0.11]	−0.08	[−0.30, 0.15]
Late										
Regression Rate	−0.02	[−0.24, 0.21]	0.12	[−0.12, 0.35]	−0.09	[−0.31, 0.14]	−0.03	[−0.25, 0.20]	−0.02	[−0.25, 0.20]
Go-Past Time	−0.09	[−0.30, 0.14]	−0.30	[−0.50, −0.7]	−0.43	[−0.59, −0.22]	−0.34	[−0.52, −0.12]	−0.15	[−0.36, 0.08]
Total-Reading Time	−0.17	[−0.39, 0.05]	−0.36	[−0.55, −0.14]	−0.50	[−0.65, −0.30]	−0.41	[−0.58, −0.21]	−0.17	[−0.38, 0.06]
Linguistic Effects										
Word-Frequency Effect on Gaze Duration	0.40	[0.19, 0.57]	0.49	[0.28, 0.65]	0.32	[0.11, 0.52]	0.48	[0.28, 0.63]	0.04	[−0.18, 0.26]

Notes. *r* = Pearson's correlation coefficient; 95% CI = 95% credible interval.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

model, which is quantified by the *estimated log predictive density* (*elpd*).[7] Importantly, the LOO-elpd is a measure of how well the model predicts new data, and not a measure of how well it explains the data that were used to train the model. To compare how much variance each of the models explained in the data, we calculated the Bayesian R^2 for all models (Gelman et al., 2019). Note that the "best" model according to LOO-elpd is not necessarily the model that explains the most variance in the training data as measured using R^2. Selection of the best model based on R^2 is prone to overfitting. LOO guards against overfitting and is therefore preferable as measure of the goodness of a model.

To investigate which set of eye-movement measures (if any) best predicted reading comprehension scores, we then looked at the output of the 10 best models according to the results of the LOO comparison, and the full model. We chose to look at the best 10 models rather than a single model because we did not have enough data to identify a single best model with sufficient certainty. Although 10 is an arbitrary number, it appears to be sufficient as the general pattern of results is relatively stable across models. Because reading speed was the most robust predictor of comprehension scores in the correlation analysis, we also looked at the output of models with only speed as a predictor, and the models with only eye-movement measures as predictors (i.e., all predictors except reading speed). This allowed us to compare the performance of reading speed and eye-movement measures alone in predicting comprehension. All models are shown in Tables 4–7.[8]

YARC

Results from the YARC (Table 4) suggest that gaze duration ($\widehat{\beta}$: −13.1),[9] go-past time ($\widehat{\beta}$: 9.4), reading speed ($\widehat{\beta}$: 7), and skipping rate ($\widehat{\beta}$: −4.7) are the best predictors of performance on this test, with gaze duration and go-past time explaining the most variance. These results differ from the correlation analysis which showed a relationship only with reading speed and gaze duration. In addition, the elpd values suggest that eye movements coupled with reading speed predict comprehension better than reading speed alone, with higher elpd values for models with both eye-tracking measures and speed (−307.16) than for the speed-only (−309.83) or eye-tracking only (−313.66) models. This shows that having eye-movement measures in addition to reading speed considerably improves predictions of the comprehension scores. The R^2 for the full model is 0.29, indicating that the full model explains 29% of the variance in the test scores.

GORT

The output of the GORT models (Table 5) suggests that first-fixation duration ($\widehat{\beta}$: 14.5), average fixation duration ($\widehat{\beta}$: −12.4), and average saccade length ($\widehat{\beta}$: 4.1) are the best predictors of performance on this test, followed closely by total-reading time ($\widehat{\beta}$: −6.5). These results differ again from the correlation results, which showed a relationship with gaze duration rather than first-fixation duration, and did not show a relationship with average fixation duration. This indicates that some measures may be predictive of comprehension, but only when evaluated jointly with other variables. Hence, regression analyses may be more revealing of the predictive relationship between eye movements and comprehension than correlations. Interestingly, the effect of reading speed is significant in the speed-only model, and not in models that include eye-movement measures, suggesting that any explanatory power reading speed may have is subsumed by the eye-tracking measures. In line with this, elpd is also much higher for the eye-movements-only model (−246.88) than for the reading-speed-only model (−250.31). The R^2 for the full model is 0.42, indicating that the full model explains 42% of the variance in the test scores.

WRAT

The output of the WRAT models (Table 6) suggests that reading speed ($\widehat{\beta}$: 11.1), skipping rate ($\widehat{\beta}$: −4.4), and regression rate ($\widehat{\beta}$: 4.1) are the best predictors of performance on this test. This pattern is again different from the correlations, although contrary to the other tests, it suggests that fewer predictors are important compared to the correlation analysis which showed most measures as related to comprehension. This suggests that although predictors are correlated to comprehension, they may be redundant and thus not all of them are needed. The eye-movements-only model has only one significant predictor, total-reading time, which is the measure typically most highly correlated to reading speed. Similar to the YARC, models with both reading speed and eye-tracking measures as predictors perform better than models with only reading speed or only eye-tracking measures, with higher elpd values for the top model (−281.01) than for the speed-only (−284.42) or eye-movement-only (−294.02) models. This strongly suggests that eye movements substantially improve predictions over reading speed alone. The R^2 for the full model is 0.46, indicating that the full model explains 46% of the variance in the test scores.

Average Data

The output for the models fit on data aggregated across the three tests (Table 7) shows reading speed ($\widehat{\beta}$: 7.4) and skipping rate ($\widehat{\beta}$: −4.7) as the best predictors of performance on the comprehension tests, closely followed by go-past time ($\widehat{\beta}$: 9.5) and total-reading time ($\widehat{\beta}$: 6.8). The importance of speed as a predictor is particularly clear from the fact that the model with only reading speed as a predictor is also the second-best model according to the LOO. The R^2 for the full model is 0.37, indicating that the full model explains 37% of the variance in the test scores.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4
YARC: Intercepts and Estimates of the Best 10 Models and 3 Comparison Models

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Speed only	EM model	Full model
Global													
Intercept	90.61	90.61	90.61	90.59	90.59	90.59	90.57	90.62	90.63	90.58	90.58	90.58	90.59
Speed (wpm)	5.95	7.50	5.94	7.56	7.00	9.38	6.15	6.83	7.90	6.11	3.13		9.29
Average Fixation Duration			6.20	7.97		9.73	6.40		10.52			2.82	8.55
Saccade Length		−2.73		−2.89		−4.23			−4.96			1.73	−4.29
Frist-Pass													
Skipping	−4.82	−4.66	−4.58	−4.36	−4.27	−3.31	−5.54	−4.96		−5.71		−4.50	−3.30
First-Fixation Duration	7.30	8.97								7.20		5.37	1.55
Gaze Duration	−13.15	−15.64	−11.68	−14.19	−5.19	−18.58	−12.68	−5.91	−20.43	−13.77		−15.22	−19.09
Late													
Regressions							−1.89	−1.71		−1.74		−3.14	0.05
Go-Past	9.34	10.63	8.81	10.06	8.63	7.59	9.97	9.29		10.30		5.89	7.66
Total-Reading Time						6.11			12.36			0.49	6.10
R^2 Bayes	0.23	0.26	0.23	0.26	0.20	0.27	0.25	0.21	0.22	0.25	0.06	0.21	0.29
ELPD-LOO	−307.16	−307.30	−307.31	−307.44	−307.46	−307.76	−307.84	−307.91	−307.95	−308.01	−309.43	−313.66	−310.06

Note. Models 1–10 are ordered based on their ELPD LOO (descending). EM Model = eye-movement measures only model. Green = 95% credibility interval does not include 0; yellow = 90% credibility interval does not include 0; white: 90% credibility interval includes 0; blank: predictor not included in the model.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5
GORT: Intercepts and Estimates of the Best 10 Models and 3 Comparison Models

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Speed only	EM model	Full model
Global													
Intercept	90.96	90.54	90.96	90.93	91.01	90.58	90.56	90.54	90.94	90.58	90.19	90.51	90.93
Speed (wpm)	−5.49		−5.34	−5.53	−6.02				−5.76		3.52		−5.32
Average Fixation Duration	−12.94	−12.87	−12.42	−12.70	−12.50	−11.92	−12.10	−12.74	−12.03	−11.38		−10.86	−11.47
Saccade Length	4.64	3.92	4.29	4.53	4.86	2.74	3.42	4.36	5.32	3.30		4.45	5.31
Frist-Pass													
Skipping									−1.13			−2.32	−1.64
First-Fixation Duration	15.21	13.68	15.27	15.15	14.96	14.95	14.03	13.25	14.54	13.90		13.42	14.85
Gaze Duration			−1.03			−4.77	−1.84			−3.03		−3.17	−2.73
Late													
Regressions				0.19								−0.65	−0.97
Go-Past					−1.44			−2.64		−1.72		2.07	0.14
Total-Reading Time	−8.28	−3.00	−7.74	−8.45	−7.50		−2.28		−8.40			−2.80	−6.24
R^2 Bayes	0.40	0.37	0.40	0.40	0.41	0.36	0.38	0.36	0.41	0.38	0.13	0.39	0.42
ELPD-LOO	−242.29	−242.93	−243.14	−243.20	−243.52	−243.58	−243.77	−243.90	−243.98	−244.05	−250.31	−246.88	−246.85

Note. Models 1–10 are ordered based on their ELPD LOO (descending). EM Model = eye-movement measures only model. Green = 95% credibility interval does not include 0; yellow = 90% credibility interval does not include 0; white: 90% credibility interval includes 0; blank: predictor not included in the model.

19362722, 2023, 3, Downloaded from https://ilaonlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6
WRAT: Intercepts and Estimates of the Best 10 Models and 3 Comparison Models

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Speed only	EM model	Full model
Global													
Intercept	105.75	105.79	105.78	105.81	105.79	105.79	105.81	105.78	105.80	105.78	105.84	105.86	105.74
Speed (wpm)	11.84	10.48	11.85	11.35	11.57	10.68	10.41	9.95	10.53	11.87	6.88		10.32
Average Fixation Duration			−3.34	1.16								−1.89	−3.17
Saccade Length									0.03	−0.01		2.73	0.87
First-Pass													
Skipping	−4.66	−4.19	−4.64	−4.48	−4.22	−4.15	−4.49	−3.86	−4.24	−4.69		−1.13	−4.89
First-Fixation Duration	1.82		4.99			2.54	2.41			1.83		−1.66	4.95
Gaze Duration					1.63			2.64				7.62	0.67
Late													
Regressions	4.13	3.71	3.96	4.04	3.79	4.93	4.50	4.10	3.71	4.14		2.55	4.20
Go-Past						−2.19						−3.27	−0.81
Total-Reading Time							−2.14	−2.65				−7.87	−1.63
R^2 Bayes	0.43	0.42	0.45	0.43	0.43	0.44	0.44	0.44	0.42	0.44	0.33	0.35	0.46
ELPD-LOO	−281.01	−281.02	−281.36	−281.50	−281.54	−281.62	−281.63	−281.70	−281.90	−282.00	−284.42	−294.02	−286.00

Note. Models 1–10 are ordered based on their ELPD LOO (descending). EM Model = eye-movement measures only model. Green = 95% credibility interval does not include 0; yellow = 90% credibility interval does not include 0; white: 90% credibility interval includes 0; blank: predictor not included in the model.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7
Average: Intercepts and Estimates of the Best 10 Models and 3 Comparison Models

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	EM model	Full model
Global												
Intercept	95.60	95.61	95.61	95.58	95.58	95.60	95.61	95.61	95.61	95.60	95.66	95.61
Speed (wpm)	7.84	4.72	7.82	6.85	8.37	7.92	8.13	8.85	4.93	8.11		7.47
Average Fixation Duration							−0.57			−5.98	−5.89	−6.86
Saccade Length				2.01							4.67	1.98
First-Pass												
Skipping	−3.94		−4.07	−5.15	−4.26	−3.94	−3.99	−3.28	−0.74	−3.97	−4.09	−5.68
First-Fixation Duration						−0.21				5.72	5.75	8.64
Gaze Duration					−1.40			−2.80			1.58	−3.30
Late												
Regressions			0.83								−0.91	−0.60
Go-Past Time	9.81		9.52	10.02	10.40	9.85	10.12	6.33		10.10	5.83	11.09
Total-Reading Time	−6.90		−6.95	−7.70	−5.88	−6.75	−6.60			−6.93	−10.06	−6.61
R^2 Bayes	0.32	0.23	0.33	0.34	0.33	0.32	0.32	0.30	0.24	0.34	0.30	0.37
ELPD-LOO	−277.16	−277.44	−277.77	−277.92	−278.07	−278.13	−278.21	−278.24	−278.27	−278.32	−285.88	−281.37

Note. Models 1–10 are ordered based on their ELPD LOO (descending). EM Model = eye-movement measures only model. Green = 95% credibility interval does not include 0; yellow = 90% credibility interval does not include 0; white: 90% credibility interval includes 0; blank: predictor not included in the model. For this dataset, because Model 2 only uses reading speed as a predictor, it is not repeated (i.e., Model 2 is equivalent to the Speed-Only models in Tables 4–6).

19362722, 2023, 3, Downloaded from https://ilaonlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Discussion

The Cognitive Skills Measured by Reading Comprehension Assessments

In this study, our first aim was to investigate whether three standardized reading comprehension tests (YARC, GORT, and WRAT) measured the same cognitive skills to the same degree within individuals. The moderate correlations between test scores within individuals are consistent with previous studies reporting modest correlations between reading comprehension tests for both comprehension measures (Andreassen & Bråten, 2010; Keenan et al., 2008; Nation & Snowling, 1997) and diagnosis outcome (Keenan & Meenan, 2014).

We used eye-tracking to investigate the relationship between test scores and the cognitive skills involved in reading. Results yielded different patterns of results between the three tests, suggesting that they do not all measure the same cognitive skills to the same extent. These patterns are summarized in Table 8. Specifically, YARC scores were most strongly associated with early eye-movement measures, which suggests that this test may be more a measure of lexical processing skills (e.g., word identification). This is comparable to findings by Cutting and Scarborough (2006) for the WIAT, which has a similar design to the YARC (i.e., passages read silently followed by open-ended comprehension questions with access to the text), as this test was more strongly associated with word reading ability as opposed to oral language comprehension skills. On the other hand, GORT scores were best predicted by global and early measures. This also suggests that the task demands of this test put more emphasis on lexical processing skills. This appears in contrast with previous findings for comparable tests (e.g., GORT-3, Cutting & Scarborough, 2006; Keenan et al., 2008; NARA, Nation & Snowling, 1997), which suggest that such tests (reading aloud with comprehension questions) tend to be more strongly associated with oral language comprehension skills rather than word reading skills. However, it is in line with Cutting and Scarborough's finding that, within their oral language comprehension component, only lexical skills (e.g., vocabulary) contributed unique variance above sentence processing skills (e.g., syntax) in GORT-3 scores. This suggests that although oral language comprehension explained more variance than word reading (0.093 vs. 0.075), most of this variance is related to lexical skills as opposed to sentence processing skills.

Although this suggests that both the YARC and the GORT are most strongly related to lexical processing skills, results from the linear models yielded opposite patterns between the two tests, as all measures useful for predicting YARC scores were seemingly not useful in predicting GORT scores, and vice versa. Therefore, although there were some similarities between the two patterns in that both included early and late measures as predictors, there were no similarities in terms of which specific measures best predicted comprehension scores. This suggests that differences in task demands between the two tests led to differences in the extent to which they measure the same cognitive skills, which is consistent with previous studies suggesting that reading comprehension tests with similar designs do not measure the same cognitive skills to the same extent (e.g., WIAT vs. GORT-3; Cutting & Scarborough, 2006).

In addition, patterns of results for the WRAT differed widely from that of the YARC and GORT. This is not unexpected as previous studies have shown that tests with cloze tasks and comprehension questions can differ in the cognitive skills they measure (e.g., Nation & Snowling, 1997). The results for this test show a clear pattern: faster reading speed, whether it reflects overall reading speed or simply reduced fixation durations, is associated with better performance. Previous studies on the cognitive skills measured by cloze tasks suggest this type of task is typically more strongly related to word reading skills than to oral language comprehension skills (e.g., Suffolk Reading Scale: Nation & Snowling, 1997; WJPC, Keenan et al., 2008). Our results suggest that WRAT scores were best predicted by reading speed and not by any measure of fixation duration. This vast difference between the WRAT and the other two tests likely stems at least in part from the fact that cloze tasks measure comprehension from one's ability to make accurate predictions about linguistic

TABLE 8
Summary Table of Relationship between Eye Movements and Comprehension

Measure	YARC	GORT	WRAT	Average
Global				
Speed (wpm)	+↑		++↑	+↑
Average Fixation Duration		+↓		
Saccades Length		+↑		
Early				
First-Pass Skipping	+↓		+↓	+↓
First Fixation Duration		++↑		
Gaze Duration	++↓			
Late				
Regression rate			+↑	
Go-past Time	+↑			++↑
Total-Reading Time		+↓		+↓

Note. Table 8 summarizes the results of the linear models. "+": significant relationship, "++": strongest relationship, ↓: negative relationship, ↑: positive relationship.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

material, which is rarely the case of comprehension questions. From our data, it is unclear whether the WRAT scores are most strongly associated with lexical or higher-level processes (e.g., sentence processing). However, it suggests that the best predictor of success in this test is the ability to make fast predictions about linguistic material.

Taken together, our results are in line with previous research on the validity of reading comprehension tests and suggest that reading comprehension tests vary not only in the extent to which they measure both word reading and oral language comprehension skills, but also additional skills such as reading aloud or making accurate predictions about linguistic material.

Using Eye Movements to Predict Reading Comprehension Accuracy

The second aim of this study was to investigate the potential of eye movements to predict reading comprehension ability. Our results show that eye movements predict reading comprehension and explained (on average) 39% of the variance in our data. This performance is similar to the models of Southwell et al. (2020) who reported correlations between predicted and observed values around 0.37. Our results therefore provide further evidence that eye-movement measures collected during reading can successfully predict performance on reading comprehension assessments. However, as discussed in the previous section, the results from our statistical modeling analyses yielded significantly different patterns in the relationship between eye-movement measures and reading comprehension tests across the three tests.

The results from the linear models did not yield any predictor common to the three comprehension tests. As such, no single eye-movement measure, or set of measures, could be identified as a good predictor of reading comprehension ability across the three tests. Reading speed and first-pass skipping rate were the most robust predictors, followed by two late measures (go-past time and total-reading time). Although the results differed widely for the YARC, GORT, and WRAT, it is important to note that all three tests included both early and late measures as "best" predictors, suggesting that the comprehension scores were generally best predicted by a combination of measures associated with both early processes of reading (e.g., lexical processing), and higher-level processes of reading (e.g., sentence integration or discourse processing). This is consistent with theories of reading comprehension which suggest that efficient word-processing and higher-level language comprehension skills are necessary for successful text comprehension (Catts et al., 2006; Cromley et al., 2010; Cromley & Azevedo, 2007; Gough & Tunmer, 1986; Kim, 2020a, 2020b). In addition, both global and local measures tended to be useful predictors of performance across tests, suggesting that including local measures as predictors improves predictions over and above the global measures that have typically been used in previous work.

The finding that global measures such as average fixation durations or saccade length are useful predictors of comprehension for the YARC and GORT is in line with previous studies showing that these two measures are useful in predicting performance on reading comprehension measures (D'Mello et al., 2020; Martínez-Gómez & Aizawa, 2014; Southwell et al., 2020). This finding suggests that global measures may be useful in predicting performance on reading comprehension tests using multiple-choice and/or open-ended questions. However, the direction of this relationship differs between studies and tasks. Indeed, previous studies suggest that making more and shorter fixations on average is predictive of better performance on the comprehension tasks. While we did replicate this finding in our correlations and in the models for the GORT, the opposite was found for the YARC with longer fixations associated with better performance (when other variables are included in the model). Similarly, the findings for saccade length are contradictory between studies and tasks, with shorter saccades predicting better comprehension scores in some studies (D'Mello et al., 2020; YARC in this article) while others find the opposite association (Martínez-Gómez & Aizawa, 2014; correlations and GORT in this article). These discrepancies in findings between studies and comprehension measures suggest that while average fixation durations and saccade length may be useful predictors of performance on comprehension questions, the direction of the relationship may be influenced by task demands. We note however that such comparisons between studies should be taken with a grain of salt given the important methodological differences between studies, as correlations between eye-tracking measures limit our interpretation to the predictors included in our models which differ from those used in other studies.

Importantly, the results from the correlations and the linear models yielded different patterns of the relationship between eye movements and comprehension scores. The difference between the results from the correlations and those of the linear models can be explained by the fact that some measures may be correlated with the comprehension scores but not be useful when trying to predict comprehension. In addition, the correlations were run individually for each measure, whereas the linear models included multiple measures. Hence, because eye-movement measures tend to be highly correlated to one another, it is less likely that two highly correlated measures are both important predictors, even though they may both be correlated with comprehension. In addition, the correlations between eye-movement measures underscore the fact that the output of the linear models should be interpreted in relation to the eye-movement measures present in the model.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Reading Speed as a Predictor of Reading Comprehension

Across both analyses, reading speed appeared as one of the most robust correlate of reading comprehension. Reading speed was generally positively correlated to comprehension, such that faster readers tended to perform better in the comprehension tasks. However, while faster readers may often be better comprehenders, reading speed alone may not be a good predictor of comprehension, as fast reading speed does not necessarily entail that comprehension is taking place. This is most clearly shown by the fact that, for all tests, adding eye movements to the models significantly improved predictions, and the amount of variance explained, over a model using reading speed alone. This is in line with Southwell et al.'s (2020) finding that models with only reading speed as a predictor performed at chance level, whereas models with both eye movements and reading times outperformed the reading time only models. Therefore, while reading speed may be a robust correlate of reading comprehension, it is not necessarily a strong predictor of reading comprehension skills.

The Influence of Task Demands on the Relationship Between Eye Movements and Reading Comprehension Accuracy

Finally, the third aim of this study was to examine the influence of task demands on reading behavior and the relationship between eye-movement measures and reading comprehension accuracy. Our results highlight the complexity of the relationship between reading comprehension and eye-movement behavior. The relationship between eye movements and reading comprehension varied with the different task demands of the three tests. This can be attributed, at least in part, to participants adapting their reading strategies to the varying task demands.

As noted earlier, an important feature of the YARC is that readers were able to return to the text to answer the comprehension questions, such that readers were not required to construct a detailed mental representation of the text during initial reading, and may have used more shallow processing strategies of the text meaning. The use of such strategies would be in line with our finding that participants spent less time reading the text on average compared to the GORT and WRAT (i.e., shorter fixations, high skipping rates, few regressions). This may also explain the pattern we see in the predictive relationship between eye-movement measures and comprehension scores. Indeed, useful predictors for the YARC scores include both measures indicative of efficient processing (e.g., short gaze duration indicative of efficient lexical processing) together with features indicating careful reading of the text (e.g., longer go-past times indicative of re-readings necessary for comprehension). This suggests that while efficient reading processes were predictive of good comprehension as might be expected (i.e., efficient readers are good comprehenders), the degree to which readers engaged in behavior necessary for successful comprehension during the initial reading of the text (e.g., re-reading) is also an important predictor of their comprehension, as readers may have read the text less carefully knowing they could return to it later.

On the other hand, a critical feature of the GORT is that readers were required to read the text aloud, hence putting an emphasis on accurate word reading and speech production which was not there in the YARC or WRAT. This emphasis on lexical processing can also be seen in the predictive relationship between eye-movement measures and comprehension scores. Indeed, good performance on the GORT was predicted by longer first-fixation durations and shorter total-reading times. This suggests that good comprehenders spent more cognitive resources on word reading processes (e.g., lexical access) and required less time re-reading the text. Hence, results for both the YARC and GORT suggest that measures associated with ease of lexical access and re-readings of the text are useful predictors of comprehension. However, differences in the direction of the relationship between predictors and comprehension suggest that it is mediated by the differences in task demands between the two comprehension measures as readers used different reading strategies in the two tests.

Lastly, the WRAT primarily assesses a reader's ability to make accurate predictions about linguistic material, as opposed to answering comprehension questions in the YARC and GORT. As noted earlier, the best predictor of performance on the WRAT is how quickly readers were able to do the task, and the best predictors do not include any measure of fixation duration. Instead, the pattern of results suggests that participants who were able to do the task quickly (i.e., short reading speed), while still reading the sentence carefully (i.e., few skips and some regressions) tended to perform better in the task.

Overall, we do find some similarities in both the type of measure that best predict comprehension accuracy (i.e., early and late measures) and the type of reading behavior that are predictive of better reading comprehension scores (e.g., efficient and careful reading). Nevertheless, the results suggest that readers adapt their reading strategies and cognitive processes to the varying task demands, which, in turn, influences both the usefulness of individual predictors and the direction of the relationship between predictors and comprehension accuracy (e.g., "efficient" lexical access predicts better performance in the YARC but poorer performance in the GORT), although the latter can only be interpreted with in the limited scope of the predictors that are included in the model. This has implications for the interpretation of eye-movement behavior in reading studies, whereby longer fixation times are often interpreted as signs of longer processing time and hence higher

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

processing difficulty. Our results suggest that the specific task demands may need to be considered when interpreting standard eye-tracking measures in reading experiments. In the final two sections of this article, we will discuss implications of our results for theories of reading comprehension.

Implications for Reading Theories

This study highlights the complexity of reading comprehension as a theoretical construct, and the difficulties that come with trying to measure it as such. The moderate correlations between the three tests, and the clear differences in reading behavior across comprehension tasks as illustrated by readers' eye movements during reading, provide further evidence for the idea that to understand and define reading comprehension one must also consider characteristics of the reader, the text, and the reading task. The results from this study highlight the fact that readers adapt their reading strategies and cognitive processing to the reading task and goal, such that the processes that are engaged by and support successful reading comprehension can vary not only across reading tasks (e.g., proofreading versus reading for comprehension; Kaakinen & Hyönä, 2010), but also across comprehension tasks. For example, one might consider differences in the processes engaged by everyday reading tasks such as understanding text messages compared to reading a textbook for learning with the delayed goal of passing a test. Although both these reading tasks require successful comprehension of written text, and hence may be assumed to engage a similar core network of cognitive processes necessary for successful comprehension and retention of the text meaning (e.g., lexical processing, making inferences), the goals are very different and likely put different demands on cognitive processes that support comprehension. However, the influence of the reading task such as the reading goal or characteristics of the text are not yet well understood, and are rarely included in theoretical models of reading comprehension (although see Kim, 2020a, 2020b). Further research is therefore necessary to better understand the role that task demands play on the cognitive processes that support reading comprehension, which is critical for the development of reliable measures of reading comprehension ability.

Limitations and Future Directions

One of the limitations of this study relates to the validity of reading comprehension measures. As noted in earlier sections, reading comprehension assessments vary in what they measure, which, in turn, made the identification of eye-movement markers of reading comprehension skills across assessments challenging. An important next step toward the overarching goal of developing a reading comprehension assessment based on eye-movement behavior is to investigate this relationship using reading tasks that more closely resemble natural reading. In addition, this article focused on differences in task demands at the whole test level, and the use of existing reading comprehension assessments that differ in multiple aspects of their design made it impossible to disentangle whether and how individual aspects of the test (e.g., reading modality) affected reading behavior and comprehension. As such factors can influence reading behavior and comprehension, future research should investigate item-level factors when investigating the influence of task demands on eye-movement behavior during reading.

Conclusions

We close by again noting the enormity of reading comprehension, and that its measurement is tantamount to measuring "thinking in general" (LaBerge & Samuels, 1974, p. 320). Our results highlight the influence of the reading task and goal on reading behavior, and the fact that these should be considered when trying to investigate, define, or measure reading comprehension and the cognitive processes that support it. The consideration of task demands is of particular importance for the development and interpretation of reading comprehension assessments used in research, schools, and clinics, since task demands heavily influence the cognitive processes that are engaged, and hence measured, by a particular reading comprehension assessment. We believe progress can be made by more clearly defining what is meant by reading comprehension in broad academic contexts (e.g., reading with the goals of being able to remember and reason about the text content), and we also predict that eye-tracking may ultimately provide a useful way to measure this type of comprehension in a direct, unobtrusive manner.

Ethics and Integrity Statement

We have no known conflict of interest to disclose. The materials of this study are not available. The data and analysis code for this study are available by contacting the corresponding author. This study was pre-registered on the Open Science framework. Pre-registration of the design and analysis plan can be found here: https://osf.io/d7apz. The analysis presented in this paper deviates partly from the pre-registration on the Open Science framework, as the planned analyses were less suited to answer our research questions. This research was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and two Australian Research Council grants (DP190100719 & DP200100311). This research was approved by the Macquarie University ethics committee.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acknowledgments

We would like to thank Serje Robidoux for providing statistical and programming advice on this project. We also thank Joe Magliano, Kiel Christianson, Julie Van Dyke, and two anonymous reviewers for helpful comments on a previous version of the article.

NOTES

[1] In this article, we use the term *task demands* to refer to whole-test task demands including reading modality, comprehension task, and text. Existing reading comprehension assessments typically differ in more than one of these aspects such that it is not possible to investigate the effect of individual factors such as reading modality independently of the others.

[2] The test manual refers to this score as "reading rate." However, because the raw score is calculated as reading time (i.e., how long participants took to read the passage), we use *reading time* instead of reading rate for clarity.

[3] For one participant, only one score out of three was available.

[4] The eye-movement measures used as predictors in the models tend to be highly correlated to each other. However, while multicollinearity of predictors can affect hypothesis testing, this is not the case when models are used to make predictions as is done in this analysis. See McElreath (2020) for a discussion of collinearity in predictive modeling compared with hypothesis testing.

[5] While the credible interval is not the same as the confidence interval, their interpretations are similar. Note that while we report credible intervals, they do not inform our conclusions as to the usefulness of our predictors.

[6] Note that in these models, the order in which the predictors are entered into the model does not affect the results.

[7] This is a Bayesian measure of predictive accuracy that takes uncertainty about the model parameters into account and which naturally penalizes model complexity. When comparing multiple models, the one with the highest elpd score is the best.

[8] This analysis was also run with only monolingual speakers to check for any possible influence of including non-native and multilingual speakers in the sample. As the results of this analysis was highly similar, they are not reported here but can be found in Data S1.

[9] This number can be interpreted as: one standard deviation in the measure (e.g., gaze duration) translates to an average (across models) of −13.1 points on the YARC scale.

REFERENCES

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44–45*, 68–82.

Andreassen, R., & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading*, *33*(3), 263–283.

Ashby, J., Yang, J., Evans, K. H., & Rayner, K. (2012). Eye movements and the perceptual span in silent and oral reading. *Attention, Perception, & Psychophysics*, *74*(4), 634–640.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System*, *83*, 64–78.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137–164.

Blythe, H. I., & Joseph, H. S. S. L. (2012). Children's eye movements during reading. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 644–662). Oxford University Press.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 1–30.

Bürkner, P. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor Comprehenders: A case for the simple view of Reading. *Journal of Speech Language and Hearing Research*, *49*(2), 278–293.

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel (Ed.), *Eye movements: A window on mind and brain* (pp. 341–374). Elsevier Science Ltd.

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*(3), 317–334.

Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2019). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing*, *33*(3), 605–634.

Copeland, L., & Gedeon, T. (2013). Measuring reading comprehension using eye movements. In *4th IEEE international conference on cognitive Infocommunications, CogInfoCom 2013 - proceedings* (pp. 791–796). IEEE.

Copeland, L., Gedeon, T., & Caldwell, S. (2016). Effects of text difficulty and readers on predicting reading comprehension from eye movements. In *6th IEEE conference on cognitive Infocommunications, CogInfoCom 2015 - proceedings* (pp. 407–412). IEEE.

Copeland, L., Gedeon, T., & Mendis, S. (2014). Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, *3*(3), 35–48.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of Reading comprehension. *Journal of Educational Psychology*, *102*(3), 687–700.

Cunnings, I., Patterson, C., & Felser, C. (2014). Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language*, *71*(1), 39–56.

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

Davey, B., & Lasasso, C. (1984). The interaction of reader and task factors in the assessment of reading comprehension. *Journal of Experimental Education*, *52*(4), 199–206.

D'Mello, S. K., Southwell, R., & Gregg, J. (2020). Machine-learned computational models can enhance the study of text and discourse: A case study using eye tracking to model Reading comprehension. *Discourse Processes*, *57*(5–6), 420–440.

Eilers, S., Tiffin-Richards, S. P., & Schroeder, S. (2018). Individual differences in children's pronoun processing during reading: Detection of incongruence is associated with higher reading fluency and more regressions. *Journal of Experimental Child Psychology*, *173*, 250–267.

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. New York, NY.
Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.
Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10.
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*(6), 1430–1440.
Inhoff, A. W., Gregg, J., & Radach, R. (2018). Eye movement programming and reading accuracy. *Quarterly Journal of Experimental Psychology*, *71*(1 Special Issue), 3–10.
Inhoff, A. W., & Radach, R. (2014). Parafoveal preview benefits during silent and oral reading: Testing the parafoveal information extraction hypothesis. *Visual Cognition*, *22*(3–4), 354–376.
Inhoff, A. W., Solomon, M., Radach, R., & Seymour, B. A. (2011). Temporal dynamics of the eye–voice span and eye movement control during oral reading. *Journal of Cognitive Psychology*, *23*(5), 543–558.
In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244.
Jones, M. W., Kelly, M. L., & Corley, M. (2007). Adult dyslexic readers do not demonstrate regularity effects in sentence processing: Evidence from eye-movements. *Reading and Writing*, *20*(9), 933–943.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.
Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during Reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(6), 1561–1566.
Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281–300.
Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing Reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125–135.
Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction*, *22*(5), 354–367.
Kim, Y.-S. (2017). Why the simple view of Reading is not simplistic: Unpacking component skills of Reading using a direct and indirect effect model of Reading (DIER). *Scientific Studies of Reading*, *21*(4), 310–333.
Kim, Y.-S. (2020a). Hierarchical and dynamic relations of language and cognitive skills to Reading comprehension: Testing the direct and indirect effects model of Reading (DIER). *Journal of Educational Psychology*, *112*(4), 667–684.
Kim, Y.-S. (2020b). Toward integrative Reading science: The direct and indirect effects model of Reading. *Journal of Learning Disabilities*, *53*(6), 469–491.
Kim, Y. S. G., & Petscher, Y. (2021). Influences of individual, text, and assessment factors on text/discourse comprehension in oral language (listening comprehension). *Annals of Dyslexia*, *71*(2), 218–237.
Kim, Y. S. G., Petscher, Y., & Vorstius, C. (2019). Unpacking eye movements during oral and silent reading and their relations to reading proficiency in beginning readers. *Contemporary Educational Psychology*, *58*, 102–120.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.
Ko, M. H. (2010). A comparison of Reading comprehension tests: Multiple-choice vs. Open-ended. *English Teaching*, *65*(1), 137–159.
Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, *58*(2), 239–261.
Kuperman, V., Matsuki, K., & Van Dyke, J. A. (2018). Contributions of reader-and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1687–1713.
Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73.
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323.
Leinenger, M., Myslín, M., Rayner, K., & Levy, R. (2017). Do resource constraints affect lexical processing? Evidence from eye movements. *Journal of Memory and Language*, *93*, 82–103.
Mancheva, L., Reichle, E. D., Lemaire, B., Valdois, S., Ecalle, J., & Guérin-Dugué, A. (2015). An analysis of reading skill development using EZ reader. *Journal of Cognitive Psychology*, *27*(5), 657–676.
Martínez-Gómez, P., & Aizawa, A. (2014). Recognition of understanding level and language skill using measurements of reading behavior. In *International conference on intelligent user interfaces, proceedings IUI* (pp. 95–104). ACM.
McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.
Murray, W. S., & Kennedy, A. (1988). Spatial coding in the processing of anaphor by good and poor readers: Evidence from eye movement analyses. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 693–718.
Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility. *British Journal of Educational Psychology*, *67*(3), 359–370.
O'Reilly, T., Feng, D. G., Sabatini, D. J., Wang, D. Z., & Gorin, D. J. (2018). How do people read the passages during a reading comprehension test? The effect of reading purpose on text processing behavior. *Educational Assessment*, *23*(4), 277–295.
Parshina, O., Sekerina, I., Lopukhina, A., & von der Malsburg, T. (2022). Monolingual and bilingual reading strategies in Russian: An exploratory scanpath analysis. *Reading Research Quarterly*, *57*(2), 469–492.
Pedersen, T. L. (2020). Patchwork: The composer of plots. *R package version 1.1.2*. https://CRAN.R-project.org/package=patchwork
Perfetti, C., & Stafura, J. (2014). Work knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37.
Powers, D. E., & Wilson Leung, S. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement*, *32*(2), 105–129.
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research*, *72*(6), 675–688.
Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.
Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720–732.

19362722, 2023, 3, Downloaded from https://ilaonlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in Reading. *Scientific Studies of Reading*, *10*(3), 241–255.
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*(3), 448–465.
Rayner, K., & Reingold, E. M. (2015). Evidence for direct cognitive control of fixation durations during reading. *Current Opinion in Behavioral Sciences*, *1*, 107–112.
Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514–528.
Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford University Press.
Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S., White, S. J., & Rayner, K. (2013). Using EZ reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, *33*(2), 110–149.
Roy-Charland, A., Colangelo, G., Foglia, V., & Reguigui, L. (2017). Passage independence within standardized reading comprehension tests. *Reading and Writing*, *30*(7), 1431–1446.
Sabatini, J., O'Reilly, T., & Deane, P. (2013). Preliminary Reading literacy assessment framework: Foundation and rationale for assessment and system design. *ETS Research Report Series*, *2013*(2), i-50.
Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2019). Engineering a twenty-first century Reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, *20*(4), 1–23.
Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*(6), 1270–1281.
Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*(1), 1–27.
Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226.
Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147–170.
Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., Nation, K., & Hulme, C. (2009). *YARC York assessment of Reading for comprehension passage Reading*. GL Publishers.
Southwell, R., Gregg, J., Bixler, R., & D'Mello, S. K. (2020). What eye movements reveal about later comprehension of long connected texts. *Cognitive Science*, *44*(10), 1–24.
Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology*, *25*(2), 121–148.
Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.
Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language*, *116*, 1–18.
Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562.
Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*(2), 477–488.
Tilstra, J., McMaster, K., Van Den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading*, *32*(4), 383–401.
Van Dyke, J. A. (2021). Introduction to the special issue: Mechanisms of variation in Reading comprehension: Processes and products. *Scientific Studies of Reading*, *25*(2), 93–103.
Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.
Vorstius, C., Radach, R., & Lonigan, C. J. (2014). Eye movements in developing readers: A comparison of silent and oral sentence reading. *Visual Cognition*, *22*(3–4), 458–485.
Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775.
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*(1), 132–137.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.
Wiederholt, J. L., & Bryant, B. R. (2012). *Gray oral reading tests—Fifth edition (GORT-5)*. Pro-Ed.
Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test (WRAT4)*. Psychological Assessment Resources.
Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL Reading comprehension. *The Modern Language Journal*, *77*(4), 473–489.
Wonnacott, E., Joseph, H. S. S. L., Adelman, J. S., & Nation, K. (2016). Is children's reading "good enough"? Links between online processing and comprehension as children read syntactically ambiguous sentences. *Quarterly Journal of Experimental Psychology*, *69*(5), 855–879.

Submitted June 13, 2022
Final revision received February 23, 2023
Accepted March 6, 2023

DIANE C. MÉZIÈRE (corresponding author) is a postdoctoral researcher in the department of Psychology at the University of Turku, Turku, Finland; email: diane.meziere@utu.fi

LILI YU is a lecturer at the School of Psychological Sciences at Macquarie University, Sydney, NSW, Australia; email: lili.yu@mq.edu.au

ERIK D. REICHLE is a professor of Cognitive Psychology in the School of Psychological Sciences at Macquarie University, Sydney, NSW, Australia; email: erik.reichle@mq.edu.au

TITUS VON DER MALSBURG is a junior professor in the Institute of Linguistics at the University of Stuttgart, Stuttgart, Germany; email: titus.von-der-malsburg@ling.uni-stuttgart.de

GENEVIEVE MCARTHUR is a professor of Cognitive Science affiliated with the Macquarie Centre for Reading (MQCR), Sydney, NSW, Australia; email: genevieve.mcarthur@mq.edu.au

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Supporting Information

Additional supporting information may be found in the online version of this article on the publisher's website:10.1002/rrq.498/suppinfo

Data S1. This file contains additional information about the scoring procedures for the YARC, GORT, and WRAT, as well as reliability information about the tests. In addition, it contains tables of results for models run with only the monolingual English speakers in our sample (Tables B1–B4).

19362722, 2023, 3, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.498 by Australian Catholic University Library - Electronic Resources, Wiley Online Library on [25/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License