Using Eye-Tracking Measures to Predict Reading Comprehension

Diane C. Mézière[1,2,3], Lili Yu[2,3], Erik D. Reichle[2,3], Titus von der Malsburg[4,5], Genevieve McArthur[2,3]

1. International Doctorate for Experimental Approaches to Language and Brain (IDEALAB), Universities of Groningen (NL), Potsdam (DE), Newcastle (UK), and Macquarie University, Sydney (AU)
2. School of Psychological Sciences, Macquarie University
3. Macquarie Centre for Reading, Macquarie University
4. Institute of Linguistics, University of Stuttgart
5. Department of Brain and Cognitive Science, Massachusetts Institute of Technology

Author note:

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.
Funding: This research was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and two Australian Research Council grants (DP190100719 & DP200100311).
Correspondence: Correspondence concerning this article should be addressed to Diane C. Mézière, Centre for Language and Cognition, Faculty of Arts, University of Groningen, 26 Oude Kijk in ‘t Jatstraat, 9712 EK, Groningen, The Netherlands. Email: d.c.meziere@rug.nl
Prior dissemination: The data and analyses presented here have been partially or fully presented at the Architectures and Mechanisms for Language Processing Conference (Potsdam, September 2020), the CUNY conference on human sentence processing (Philadelphia, March 2021), and the annual meeting of the Society for the Scientific Study of Reading (Lancaster, July 2021).
Acknowledgements: We would like to thank Serje Robidoux for providing statistical and programming advice on this project.

Abstract

Research on reading comprehension assessments suggests that they measure overlapping but not identical cognitive skills. In this paper, we 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 for Reading 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 given sentences with a missing word to read silently, and had to provide the missing word (i.e., a cloze task). Results from linear models predicting comprehension scores from eye-tracking measures yielded different patterns of results between the three tests. Models with eye-tracking measures always explained significantly more variance compared to baseline models with only reading speed, with R-squared 4 times higher for the YARC, 3 times for the GORT, and 1.3 times for the WRAT. Importantly, despite some similarities between the tests, no common good predictor of comprehension could be identified across the tests. Overall, the results suggest that reading comprehension tests do not measure the same cognitive skills to the same extent, and that participants adapted their reading strategies to the tests' varying task demands. Finally, this study suggests that eye-tracking may provide a useful alternative for measuring reading comprehension.

Keywords: reading comprehension; eye-tracking; assessment

The term “reading comprehension” is commonly used by reading researchers to refer to the sum total of processes that support the understanding of text, as well as the mental representations that are the product of those processes (Kintsch, 1998; LaBerge & Samuels, 1974; Perfetti & Stafura, 2014). 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). To understand this sentence, for example, it is necessary to use information about the syntactic categories of its words and their likely semantic roles to generate the units of meaning corresponding to phrases and the sentence as a whole. And similarly, to fully understand this next sentence in relation to the previous sentence, it is necessary to make any number of inferences, such as linking the words “the previous” in the previous phrase back to their referent, the previous sentence.

Reading comprehension is further complicated by the fact is that it does not occur in a vacuum, but instead depends upon the fluid and coordinated operation of a large number of supportive processes. The most obvious of these is the identification of the individual words. Other important examples include “front end” operations involved in visual processing and the focusing and shifting of covert attention, and “back end” processes that include the programming and execution of eye movements. Executive processes are also required to coordinate all of these operations and transfer high-level text representations from working memory to long-term memory. If one then considers the actual *measurement* of reading comprehension using traditional measures, additional processes are introduced that require the capacity to remember (via recognition or recall) the contents of a text. This capacity, in turn, is influenced by motivation and willingness to exert effort to “reconstruct” the meaning of a text.

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." Indeed, the main purpose of this article is to provide evidence of this fact. We will do this by examining the reading comprehension scores of a sample of adults who were tested on three widely used comprehension tests: the *York Assessment for Reading Comprehension* (*YARC*; Snowling et al., 2009), *Gray Oral Reading Test* (*GORT-5*; Wiederholt & Bryant, 2012) , and *Wide Range Achievement Test* (*WRAT-4*; Wilkinson & Robertson, 2006). As will become evident, although these tests were all designed to measure reading comprehension, the reliability of the comprehension scores across tests is modest, reflecting the simple fact that reading comprehension entails a plethora of mental operations, only some of which are captured by any one reading comprehension test. It will also become evident that reading comprehension is dependent upon (or embedded within) basic operations of coordinating both eye movements and attention to efficiently encode and identify the individual words in a text. More specifically, while eye movements measured during the reading of the YARC, GORT, and WRAT stimuli can be used to predict individual comprehension scores, the best set of predictors and their predicting power are subject to a variety of different task demands.

In the remainder of this article, we will first briefly review what is known about reading comprehension and the mental operations that are required to understand text. We will then provide a more detailed overview of the three reading comprehension tests that were used in our eye-movement study, as well as a primer of how eye movements can be used to understand the processes involved in reading. This latter discussion will also include an overview of the relationships between each of the standardly used eye-movement measures and the reading processes that they measure (e.g., individual fixation durations on words

reflect their lexical processing difficulty; Rayner, Ashby, Pollatsek & Reichle, 2004; Reingold, Reichle, Glaholt & Sheridan, 2012; Schilling, Rayner & Chumbley, 1998; for a review, see Rayner, 2009). We will then report the results of an eye-movement study in which the reading comprehension levels of a large sample of participants were measured using the YARC, GORT, and WRAT. Finally, we will conclude with a discussion of the theoretical implications of our findings, as well as their pedagogical ramifications. Although one might anticipate that our findings reflect negatively on previous attempts to measure reading comprehension, we believe that they instead shed light on what is actually being measured and thus illuminate how one might go about designing better tests of reading comprehension in the future.

Models of Reading Comprehension

One of the most prominent models of reading comprehension is the *simple view of reading* (*SVR*; Gough & Tunmer, 1986), which argues that reading comprehension has two main components: word reading and oral language comprehension. While there is a lot of evidence consistent with this theory, for example from studies showing that statistical models including word reading and listening comprehension explain large amounts of variance in reading comprehension (90% Catts, Adlof, & Weismer, 2015; 45-47% Georgiou, Das, & Hayward , 2009; 38-61%, Tilstra, McMaster, Van Den Broek, Kendeou, & Rapp, 2009), it has been argued that it is too simplistic, as both word reading and listening comprehension are themselves complex processes and are both supported by multiple processes and skills. More complex models of reading comprehension have therefore been proposed that try to include the component skills of word reading and listening comprehension to better understand the cognitive skills involved in reading comprehension and the relations between them.

One such theory, the *Direct and Inferential Mediation model* (*DIME*; Cromley & Azevedo, 2007; Cromley, Snyder-Hogan, & Luciw-Dubas, 2010), includes five skills as predictors of reading comprehension: background knowledge, vocabulary, inferences, reading strategies, and word reading, with these skills being assumed to either directly or indirectly influence reading comprehension to different extents. The five skills specified in the DIME model have been found to explain a large portion of the comprehension variance (66% - 100%) in studies with large samples of university students (Cromley et al., 2010), as well as middle school and high school students (Cromley & Azevedo, 2007; Ahmed, Francis, York, Fletcher, Barnes, Kulesz, 2016).

In a recent series of experiments exploring the component skills involved in reading comprehension, Kim (2017, 2020a, 2020b) proposed a theory of reading comprehension, the *Direct and Indirect Effects model of Reading* (*DIER*), which includes a more extensive set of linguistic and cognitive skills on reading comprehension. DIER assumes a hierarchical structure of the predictors of reading comprehension, and includes both direct and mediated effects of these skills. Specifically, it assumes that only word reading, listening comprehension, and reading fluency have a direct effect on reading comprehension, with reading fluency mediating the effects of both word reading and listening comprehension. On the other hand, component skills of word reading (phonology, semantics, and orthography) and listening comprehension (inferences, comprehension monitoring, reasoning, and perspective taking) are argued to have only mediated effects via word reading and listening comprehension. Unlike the SVR, the DIER divides the components of listening comprehension between higher-level cognitive skills (e.g., making inferences) and 'foundational' language skills, namely vocabulary and grammar, and assumes that the latter are necessary but not sufficient for successful listening comprehension. Hence, in this theory, the subcomponents of listening comprehension are themselves assumed to have a hierarchical

structure, whereby vocabulary and grammar have both a direct and indirect effect on listening comprehension via higher-order cognitive skills. Finally, the model assumes that other domain-general cognitive skills such as working memory or attention do not have a direct effect on reading comprehension, but rather a mediated effect via listening comprehension and word reading. The DIER model was tested on large samples of English- and Korean-speaking children and was found to explain between 66% (Kim, 2017, 2020b) and 95% (Kim, 2020a) of the variance in the data. Importantly, the DIER model also assumes dynamic relations between the cognitive skills involved in reading comprehension to allow for mediating effects of both text characteristics and assessment methods. For example, the model assumes that some reading tasks/assessment methods may rely more on word reading skills than listening comprehension skills, hence affecting the relative weight of these skills in supporting successful reading comprehension.

In sum, the DIME and the DIER models similarly assume that successful reading comprehension is supported by multiple processes and cognitive skills, including lower-level (e.g., lexical processing) and higher-level (e.g., making inferences; Kendeou, McMaster, & Christ, 2016; Oakhill & Cain, 2012) cognitive skills, which are supported by language-specific skills (e.g., vocabulary, grammar; Kim, 2017) and domain-general skills (e.g., working memory; see Peng et al., 2018 for a review). Importantly, the relative importance of these processes has been shown to vary with development (e.g., Tilstra et al., 2009), text characteristics (e.g., Kim & Petscher, 2020), and assessment methods (e.g., Collins, Compton, Lindstöm, & Gilbert, 2019). We focus on the latter in the following section.

Reading Comprehension Measures

Previous studies have suggested that different tests of reading comprehension tax different cognitive abilities to different degrees (Keenan, Betjemann, & Olson, 2008; Kendeou, Papadopoulos, Spanoudis, 2012). An early study by Nation and Snowling (1997)

compared the performance of 184 children with typical development on two reading comprehension tests that are widely used in Britain: the *Neale Analysis of Reading Ability* (*NARA*) and the *Suffolk Reading Scale*. In the NARA, children read passages of text out loud followed by comprehension question. 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 accuracy (for single words, nonwords, and text) explained a significant amount of variance for both tests. However, listening 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 listening comprehension could 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 on decoding skills compared to comprehension questions (e.g., Spear-Swerling, 2004).

In a later study, Cutting and Scarborough (2006) investigated the relative contributions of decoding, listening comprehension, as well as other cognitive skills (reading speed, attention, IQ, verbal memory, and rapid serial naming) on three measures of reading comprehension with varying task demands. They tested 97 children on three widely-used reading comprehension tests in the United States: 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: (1) reading modality (aloud: GORT-3; silent: WIAT and GM-R); (2) task (multiple-choice question: GORT-3 and GM-R; open-ended questions: WIAT); and (3) availability of the text during the task (available: GM-R and WIAT; taken away: GORT-3). They found that word reading (measured by word and pseudoword reading accuracy) and oral language accounted for

varying amounts of unique variance, as well as a large and significant amount of shared variance in reading comprehension test scores. Furthermore, the unique contribution of two aspects of oral language skills, lexical skills (i.e., the recognition, naming, and classification of pictures), and sentence-processing skills (i.e., following a set of directions, sentence generation, sentence recall, and an experimental syntactic comprehension measure), varied between the three tests. Lexical skills accounted for unique variance in the GORT-3 (reading aloud with text unavailable for the multiple-choice questions), sentence-processing skills accounted for unique variance in the WIAT (reading silently with text available for the open-ended questions), and both lexical and sentence-processing skills accounted for unique variance in the GM-R (reading silently with text available for the multiple-choice questions). With regards to the other cognitive skills, only reading speed accounted for additional variance. 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 on the relative importance of decoding and oral language comprehension skills in reading comprehension test performance, Keenan, Betjemann, and Olson (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 listening comprehension (cloze procedure, retelling, and comprehension questions tasks) and word and nonword decoding. 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 (reading sentences silently with

picture selection) and the WJPC (reading sentences and passages silently with cloze task), but that listening comprehension was a better predictor for the GORT-3 (reading passages aloud with multiple-choice question) and the QRI-3 (reading passages aloud with open-ended questions and retell). Additionally, they found that differences in the variance explained by decoding skills varied not only between tests but also within test, along with reading development, such that decoding skills accounted for significantly less variance in older reader's scores on the PIAT and the WJPC. These findings further suggest that multiple characteristics of reading assessments, such as text length (e.g., sentence vs. passage), as well as task 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, Colangelo, Foglia, Reguigui, 2017).

Taken together, the vast differences in task characteristics may well explain why correlation coefficients between widely used reading comprehension test vary within and between 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). Studies have shown that some task characteristics are associated with better comprehension performance than others (e.g., Best, Floyd, & McNamara, 2008; Davey & Lasasso, 1984; Ko, 2010; Shohamy, 1984; Wolf, 1993). For example, answering multiple-choice questions tend to result in higher comprehension scores than the open-ended questions for children (Collins et al., 2019) and first- and second-language adult readers (In'nami & Koizumi, 2009). Making the text available when answering comprehension questions also leads to higher

comprehension accuracy compared to when the text was taken away (Andreassen & Bråten, 2010; Davey & Lasasso, 1984).

What is more important, perhaps, is the fact that the fact that there is colossal task variance in the comprehension tests also means that different cognitive skills and mental processes are engaged by and support reading comprehension in the various tests, as reflected by the variance in the explanation power of the different cognitive skills. For example, listening 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, non-word 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 whilst they read text. In this paper, we investigate the relationship between eye-movement behaviour during reading comprehension tasks and reading comprehension scores, and explore the possibility of using eye movements to predict, and eventually measure, reading comprehension ability.

Eye Movements During Reading

The tracking of eye movements is a widely used, non-invasive method to index online cognitive processing during reading (Rayner, Chace, Slattery, & Ashby, 2006). Eye tracking provides two primary measures of eye-movement behaviour: *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, Myslin, Rayner, & Levy, 2017; Sturt, 2007), and word frequency and predictability (Rayner, Slattery, Drieghe, & Liversedge, 2011; Schilling et al., 1998).

Eye-movement measures can be further 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). *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, Staub & Rayner, 2007; Vasishth, von der Malsburg & Engelmann, 2013). Early measures include *first-fixation duration* (i.e., the duration of the initial fixation on a word conditional upon it occurring during first-pass reading) and *gaze duration* (i.e., the sum of all first-pass fixations). Late measures include *go-past time* (i.e., the sum of all fixations from when the eyes first fixate on the word to when the eyes move to the right off the word, including all regressions to the left of the word) and *total-reading time* (i.e., the sum of all fixations on a word, irrespective of whether the fixations occur after a regression). These word-level measures are typically used to investigate word-level linguistic variables, such as *word frequency* (i.e., words that occur frequently in text tend to be the recipients of fewer, shorter fixations than infrequent words; Schilling et al., 1998). This dichotomy is not strict, however, because word-level measures have been used to study post-lexical integration (e.g., Warren, White, & Reichle, 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).

Predicting comprehension accuracy from eye-movement behaviour

Predicting reading comprehension ability from eye-movement behaviour is no easy task. Indeed, while the relationship between reading comprehension and eye-movement behaviour 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 behaviour 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 suggests that more efficient eye-movement behaviour (e.g., shorter fixations, fewer regressions) tend to be associated with better reading comprehension (e.g., Kim, Petscher, & Vorstius, 2019; Parshina, Sekerina, Lopukhina & von der Malsburg, 2021). 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, Joseph, Adelman, & Nation, 2016). And still other researchers find no relationship (Christianson, Luke, Hussey, & Wochna, 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, Gedeon, and Mendis (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 (correct, half-correct, incorrect response). Although this type of method does not allow for a clear links 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, Gedeon, & Caldwell, 2016; Copeland & Gedeon, 2013; Martínez-Gómez & Aizawa, 2014).

Inhoff, Gregg, and Radach (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, first-pass skipping rate) 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, Gregg, Bixler, and D'Mello (2020) predicted reading comprehension scores in three datasets using seven global eye-movement measures (number of fixations, average fixation duration, mean saccade length, proportion of regressions, proportion of horizontal saccade, fixation dispersion, and reading time). 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. Additionally, 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, Southwell, & Gregg, 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. Additionally, 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 behaviour (Bax & Chan, 2019; Kaakinen & Hyönä, 2010; O'Reilly, Fend, Sabatini, Wang, & Gorin, 2018; Radach, Huestegge, & Reilly, 2008; Schotter, Bicknell, Howard, Levy, & Rayner, 2014), it is important to investigate whether these results can be replicated across comprehension measures (e.g., open-ended questions, providing a text summary).

The present study

In this study, we used eye movements to investigate the cognitive processes that are engaged by three widely used adult reading comprehension tests that have varying task demands: the GORT, the WRAT, and the YARC. The aim of the study was two-fold. Firstly, we investigated the degree to which the three comprehension tests measured the same cognitive skills. Based on previous research, we expected that there would be differences in the extent to which the tests measured cognitive skills involved in reading, and that these

differences would be seen both in participants' test scores, and in their eye-movement behaviour. Secondly, we investigated the potential of eye movements to predict reading comprehension scores. In this second analysis, we aimed to identify a subset of eye-movement measures that best predicted performance in our three comprehension measures.

Method

Participants

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. 60 participants were monolingual native speakers of English, eight 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 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: (1) the York Assessment of Reading Comprehension – Passage Reading Secondary, Australian Edition (*YARC*; Snowling et al., 2009); (2) the Gray Oral Reading Test – 5th edition (*GORT-5*; Wiederholt & Bryant, 2011); and (3) the word

reading and sentence comprehension subtests of the Wide Range Achievement Test – 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 the eye-tracking purposes, the test items were all presented on a computer screen. However, the tests were administered and scored according to the test manual procedures with one exception: for the purpose of obtaining 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 Appendix A.

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) while they were reading. At the end of each passage, participants were asked 13 comprehension questions about the text, and were able to return to the text to answer them. Additionally, 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 comprehension 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[1], 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.

At the start of the test, each participant was instructed that they would be asked to read passages aloud as quickly and accurately as possible and then answer five open-ended questions about the passage content. Each passage was presented on a computer screen over 1 to 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 whilst reading aloud; e.g., incorrect pronunciations, hesitations, skipped words), 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

[1] 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.

scores for time, accuracy, fluency, and comprehension were converted into scaled scores provided by the manual of the GORT-5, which had 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 participant's eye-movements on the same items. Participants thus saw items 20 to 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 colour on a grey background (RGB: 204, 204, 204) on a BenQ Zowie XL2540 screen with a screen

resolution of 1,920 × 1,080 pixels and a refresh rate of 240 Hz. The items from the YARC and GORT (i.e., 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 so as to be able 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 cut-off 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 maximum allowance for the drift correction error was 0.45˚. The eye-tracker collected fixation positions, durations, and saccades. 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 this item was not scored. For these participants, final test comprehension scores could not be calculated accurately and were thus 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 was 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. Additionally, all words around punctuation marks were excluded from analysis (except for the final word in each sentence when calculating the wrap-up effect). For each participant, fixations shorter than 80 ms or longer than 800 ms were excluded. Additionally, forward saccades longer than the perceptual span (20 characters; Rayner, 2009) were excluded from analysis (3% of all forward saccades).

Data Analysis

We ran two sets of analyses. In the first analysis, we investigated differences between the tests both in terms of the test scores, and participants' eye-movements. In the second analysis, we used eye-movement measures to predict test scores. In both analyses, we included nine variables: (1) reading speed (i.e., number of words read per minute) and eight eye-movement measures: (2) average fixation duration (i.e., mean duration of all fixations in a given text); (3) average forward saccade length (i.e., mean length of all rightward saccades in a text, in character spaces); (4) first-pass skipping rate (i.e., the proportion of words skipped in a text during first-pass); (5) first-fixation duration (i.e., the duration of the initial fixation on a word); (6) gaze duration (i.e., the sum of all first-pass fixations on a word); (7)

regression rate (i.e., the proportion of all regressions made in a text); (8) 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 (9) total-reading time (i.e., the sum of all fixation durations on a word). The latter three measures, contrary to first-pass measures (4, 5, & 6), are posited to reflect higher-level processes such as syntactic processing and the integration of word meanings. Although measures (4) to (9) were calculated based on each word within the tests, these measures were aggregated by test. Additionally, we calculated two linguistic effects on eye movements: (1) *word-frequency effects* on gaze duration (Schilling et al., 1998), and (2) *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 by as the difference between total-reading time of the final word in a sentence compared and the average total-reading time of all 'middle' words in a sentence (excluding the first word of a sentence and all words around punctuation). Because the items 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, all data exclusions, all manipulations, and all measures in the study. The materials of this study are not available, as they are copy-righted. 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, Maechler, Bolker, &Walker, 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

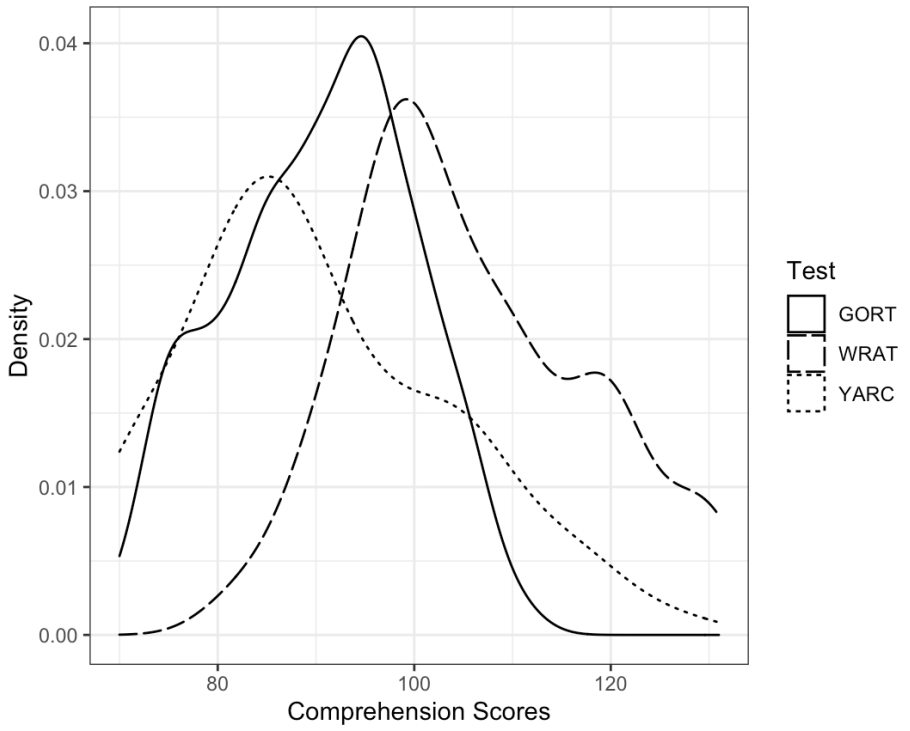
The results showed that participants could receive quite different comprehension scores on the three tests, as shown in Figure 1. Specifically, while most participants received scores within the average range (i.e., within one standard deviation from the mean: 85-115) in each test, many participants performed differently in at least two of the tests. Of our 78[2] participants, only 29 received scores in the same range for all three tests (37%). 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 individual's scores on each test. These correlations are shown in Figure 2. Participant's scores on the GORT were transformed from scaled scored 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: $r = 0.57$).

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

Figure 1

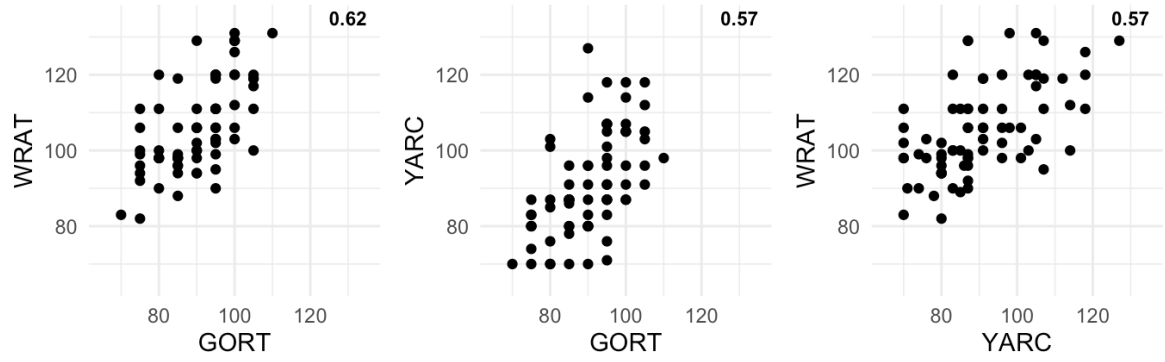
Distribution of Comprehension Scores



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.

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

Differences in Eye Movements Between Tests

Table 2 shows the mean values for reading speed, the various eye-movement measures, and the two linguistic effects. 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, with a random effect of participants. The wrap-up effect is the difference between the average total-reading time on words within the sentence (excluding the first word and any word around punctuation) and the total-reading time on the final word of the sentence. We also investigated differences in eye-movement behaviour 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 linguistic effects (word-frequency and wrap-up effects). Participants tended the read at a slower pace in the GORT, with slower reading speed and longer fixation times, as well as larger effects of word frequency on gaze duration and wrap-up on total-reading time. Eye-movement behaviour 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.

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)	-0.08	-0.11	-0.07
Wrap-up Effect on Total-Reading Time (ms)	30	111	NA

Notes: *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. All correlation coefficients are shown in Table 3. Participants' test scores on the YARC were significantly correlated only to reading speed, gaze duration, and the word-frequency effect

(all *p*s < 0.05). Participants' test scores on the GORT were significantly correlated with reading speed, saccade length, gaze duration, total-reading time, and the word-frequency effect (all *p*s < 0.05). Participants' scores on the WRAT were significantly correlated to all eye-movement measures (all *p*s < 0.05) except for skipping and regression rate.

Because the pattern of results in these correlations differed widely, we also calculated the correlations between each participant's averaged score across tests and eye movements averaged across the three tests. The mean test scores were significantly correlated to reading speed, saccade length, gaze duration, go-past time, total-reading time, and the word-frequency effect (all *p*s < 0.05). Correlations with average fixation duration and first-fixation duration were marginally significant (*p*s = 0.058 and 0.06, respectively). Additionally, 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 behaviour. Correlations between the standard deviations of the test scores and the averaged eye-movement measures did not reach significance (apart from reading speed).

Table 3

Correlations between Average and Standard Deviations of Test Scores and Eye Movements.

Measures	YARC	GORT	WRAT	Mean Score - Mean EM	SD Scores - Mean EM
Global					
Reading Speed	**0.23***	**0.30***	**0.57***	**0.48***	**0.24***
Average Fixation Duration	-0.17	-0.11	**-0.28***	-0.22**	-0.11
Average Saccade Length	0.15	**0.41***	**0.26***	**0.32***	-0.02
First-Pass					
Skipping Rate	-0.03	0.09	0.16	0.07	-0.01
First-Fixation Duration	-0.19	-0.07	**-0.26***	-0.21$^{p=0.06}$	-0.09
Gaze Duration	**-0.26***	**-0.35***	**-0.29***	**-0.33***	-0.08
Late					
Regression Rate	-0.02	0.12	-0.09	-0.03	-0.02
Go-Past Time	-0.09	-0.30	**-0.43***	**-0.34***	-0.15
Total-Reading Time	-0.17	**-0.36***	**-0.50***	**-0.41***	-0.17
Linguistic Effects					
Word-Frequency Effect on Gaze Duration	**0.35***	**0.52***	**0.24***	**0.45***	0.05
Wrap-Up Effect on Total-Reading Time	-0.21	-0.16	NA	-0.17	-0.06

Note: * $p < 0.05$; ** $p = 0.058$.

Predicting Reading Comprehension Scores from Eye Movements

In a second analysis, we investigated whether eye movements could predict reading comprehension scores. In this analysis, we first fitted linear regression models with reading speed and eye-movement measures as predictors of reading comprehension scores. Then, we evaluated and compared these models to identify the subset of predictors that best predicted comprehension scores.

We fitted linear regression models within the Bayesian framework using the 'brms' package (Bürkner, 2017; 2018) in *R*, and made inference about predictors' effects based on 95% credible intervals[3]. 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)[4]. 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 *z*-transformed 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 2,048 models).

Within each set of 512 models per test, individual models were evaluated and compared using *leave-one-out* cross-validation (*LOO*; Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 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,

[3] While the credible interval is not the same as the confidence interval, their interpretations are similar.
[4] Note that in these models, the order in which the predictors are put into the model does not affect the results.

and then evaluating how well the left-out data point is predicted by the model, which is quantified by the *estimated log predictive density* (*elpd*[5]). 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, Goodrich, Gabry, & Vehtari, 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 preferrable 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 ten best models according to the results of the LOO comparison, as well as the full model. We chose to look at the best ten models rather than a single model because we did not have enough data to identify a single best model with sufficient certainty. Although ten is an arbitrary number, it appears to be enough as the general pattern of result 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 to 7.

[5]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.

YARC

Results from the YARC (Table 4) suggest that gaze duration ($\hat{\beta}$: -13.1)[6], go-past time ($\hat{\beta}$: 9.4), reading speed ($\hat{\beta}$: 7), and skipping rate ($\hat{\beta}$: -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. Additionally, 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) 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 ($\hat{\beta}$: 14.5), average fixation duration ($\hat{\beta}$: -12.4), and average saccade length ($\hat{\beta}$: 4.1) are the best predictors of performance on this test, followed closely by total-reading time ($\hat{\beta}$: -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, although the effect of reading speed is significant in the speed-only model, and is not in models that include eye-movement measures, suggesting that any

[6] 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.

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 ($\hat{\beta}$: 11.1), skipping rate ($\hat{\beta}$: -4.4), and regression rate ($\hat{\beta}$: 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 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. Similarly 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) 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) show reading speed ($\hat{\beta}$: 7.4) and skipping rate ($\hat{\beta}$: -4.7) as the best predictors of performance on the comprehension tests, closely followed by go-past time ($\hat{\beta}$: 9.5) and total-reading time (mean estimate: 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.

Table 4

YARC: Intercepts and Estimates of the Best Ten Models and Three 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
	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
Global	Average Fixation Duration			6.20	7.97		9.73	6.40		10.52			2.82	8.55
Global	Saccade Length		-2.73		-2.89		-4.23			-4.96			1.73	-4.29
Global	Skipping	-4.82	-4.66	-4.58	-4.36	-4.27	-3.31	-5.54	-4.96		-5.71		-4.50	-3.30
Frist-Pass	First-Fixation Duration	7.30	8.97								7.20		5.37	1.55
Frist-Pass	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
Frist-Pass	Regressions							-1.89	-1.71		-1.74		-3.14	0.05
Late	Go-Past	9.34	10.63	8.81	10.06	8.63	7.59	9.97	9.29		10.30		5.89	7.66
Late	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

Notes: Models 1 to 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.

Table 5

GORT: Intercepts and Estimates of the Best Ten Models and Three 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
	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
Global	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
Global	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
	Skipping									-1.13			-2.32	-1.64
Frist-Pass	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
Frist-Pass	Gaze Duration			-1.03			-4.77	-1.84			-3.03		-3.17	-2.73
	Regressions				0.19								-0.65	-0.97
	Go-Past					-1.44			-2.64		-1.72		2.07	0.14
Late	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

Notes: Models 1 to 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.

Table 6

WRAT: Intercepts and Estimates of the Best Ten Models and Three 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
	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
Global	Average Fixation Duration			-3.34	1.16								-1.89	-3.17
Global	Saccade Length									0.03	-0.01		2.73	0.87
	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-Pass	First-Fixation Duration	1.82		4.99			2.54	2.41			1.83		-1.66	4.95
First-Pass	Gaze Duration					1.63			2.64				7.62	0.67
	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
Late	Go-Past						-2.19						-3.27	-0.81
Late	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

Notes: Models 1 to 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.

Table 7

Average: Intercepts and Estimates of the Best Ten Models and Three 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
	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
Global	Average Fixation Duration							-0.57			-5.98	-5.89	-6.86
Global	Saccade Length				2.01							4.67	1.98
Global	Skipping	-3.94		-4.07	-5.15	-4.26	-3.94	-3.99	-3.28	-0.74	-3.97	-4.09	-5.68
First-Pass	First-Fixation Duration						-0.21				5.72	5.75	8.64
First-Pass	Gaze Duration					-1.40			-2.80			1.58	-3.30
First-Pass	Regressions			0.83								-0.91	-0.60
Late	Go-Past Time	9.81		9.52	10.02	10.40	9.85	10.12	6.33		10.10	5.83	11.09
Late	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 to 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).

Discussion

Do reading comprehension tests measure the same cognitive skills?

In this study, our first aim was to investigate whether three standardised reading comprehension tests (the YARC, GORT, and WRAT) measured the same cognitive skills and 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 (Keenan et al, 2008; Andreassen & Bråten, 2010, Nation & Snowling, 1997) and diagnosis assignment (Keenan & Meenan, 2014).

We used eye-tracking to investigate the relationship between test scores and the cognitive skills involved in reading. Results from two analyses 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 summarised 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. 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 to word reading ability as opposed to listening comprehension skills. On the other hand, the GORT scores were equally associated to both early and late measures, but best predicted by global and early measures. This also suggests that the task demands of this test may 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 suggests that such tests (reading aloud with comprehension questions) tend to be more strongly associated with listening comprehension skills rather than word reading skills.

However, it is in line with Cutting and Scarborough's finding that, within their listening comprehension component, only lexical skills contributed unique variance above sentence comprehension skills in GORT-3 scores, suggesting that although listening comprehension explained more variance than decoding (0.093 versus 0.075), most of this variance 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 exact opposite patterns between the two tests, as all measures indicated as good predictors of 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 good predictors, there were no similarities in terms of which specific measures best predicted comprehension scores. This suggests that despite some similarities, differences in task demands between the two tests led to differences in the extent to which they measure the same cognitive skills, and this conclusion 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 versus GORT-3; Cutting & Scarborough, 2006).

Additionally, 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 and Snowling, 1997). The results for this test show a clear pattern: faster reading speed, whether it be overall reading speed or fixation durations, is associated with better performance. Previous studies on the cognitive skills measured by cloze tasks suggests this type of task is typically more strongly related to word reading skills than to listening comprehension skills (e.g., Suffolk Reading Scale: Nation & Snowling; WJPC, Kennan et al., 2008). Our results suggests that WRAT scores were more strongly related to late eye-movement measures,

which are associated with later processes of reading such as sentence integration, but ultimately were best predicted by reading speed and not by any measure of fixation durations. 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 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. 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 decoding skills and listening comprehension skills, but also additional skills such as reading aloud or making accurate predictions about linguistic material.

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	+			++
	Toral Reading Time		+		+

Note: Table 8 summarizes the results of the linear models. '+': significant relationship, '++': strongest relationship.

Do eye-movement measures predict reading comprehension?

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. However, as discussed in the previous section, the results from both the correlations and our statistical modelling analyses yielded significantly different patterns in the relationships 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 linear models appeared to differ 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 (such as lexical processing), as well as higher-level processes of reading (such as sentence integration or discourse processing). This is consistent with theories of reading comprehension which suggests that both good word-processing skills and good higher-level language comprehension skills are necessary for successful text comprehension (Catts et al., 2006; Cromley & Azevedo, 2007; Cromley et al., 2010; Gough & Tunmer, 1986; Kim, 2020a, 2020b).

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 yet not be very useful when trying to predict comprehension. Additionally, 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, although they may both be correlated with comprehension.

Across both analyses, reading speed appeared as one of the most robust predictors 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, as well as the amount of variance explained, over a model using reading speed alone. Therefore, while reading speed may be a robust correlate of reading comprehension, it is not necessarily a strong predictor of reading comprehension skills.

Finally, our results highlight the complexity of the relationship between reading comprehension and eye-movement behaviour. 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. Interestingly, the relationship between eye-movement measures and comprehension was not always in the same direction, such that longer fixation times (i.e., longer processing times) were not always associated with poorer comprehension ability (e.g., longer go-past time predicted higher scores in the YARC). Similarly, higher skipping rates, previously found to be associated with higher spelling ability (e.g., Veldre & Andrews, 2016; Slattery & Yates, 2018), predicted lower performance on comprehension. This has implications for the interpretation of eye-movement behaviour in reading studies, whereby longer fixation times are often interpreted as signs of longer processing time and hence higher 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 and reading instruction.

Implications for reading theories

The most obvious implication of our results is that, although reading "comprehension" has been almost universally adopted as a theoretical construct along with the accompanying supposition that it can be measured as such, the fact that comprehension scores correlated only modestly within individual readers suggests that it instead refers to a loose set of interrelated skills that allow the contents of a text to be used in a variety of task-specific ways. For example, in our study, although the YARC probably corresponds most closely to what might be described as reading comprehension (i.e., the capacity to convert written text into mental representations that afford the capacity to later answer questions about the text), the GORT additionally required participants to read aloud while the WRAT required participants to actively make predictions about the identities of missing words. These task-demand differences attenuated the reliability of the comprehension measures within individuals (i.e., between-test *r*s ≈ 0.6). These task-demand differences also manifest in the patterns of eye-movement behaviours that were evident as participants engaged with the texts used in the comprehension measures.

If our conjecture is correct, then reading "comprehension" as a theoretical construct may have a different ontological status than other core processes that support reading and that are standardly described or conceptualized as being unitary procedures or mechanisms (for a review of these theories and models, see Reichle, 2021). For example, despite marked differences in their assumptions, word-identification models share the assumption that a single set of procedures or mechanisms are used to access the pronunciations and meanings of word from their orthographic forms (e.g., Ans, Carbonnel, & Valdois, 1998; Coltheart,

Rastle, Perry, Langdon, & Ziegler, 2001; Plaut, McClelland, Seidenberg, & Patterson, 1996). And similarly, although sentence-processing models make different assumptions about the roles that syntactic and semantic information play in constructing the larger units of meaning corresponding to phrases and sentences, all of these models assume that the procedures or mechanisms that are responsible for constructing these representations are largely invariant across individuals (e.g., Elman, 1990; Frazier & Rayner, 1982; Spivey & Tanenhaus, 1998; Van Dyke & Lewis, 2003). Thus, among researchers who attempt to understand word identification and/or sentence processing, there seems to be a consensus belief that the processes that support these core reading activities can be described and simulated using a relatively small set of theoretical assumptions that are fixed across individuals, largely automatic in nature, and thus less subject to strategic control.

In contrast to both word identification and sentence processing, the collection of processes that support reading "comprehension" appear to be broader in scope and to vary both between and within individuals (e.g., as a function of skill, motivation to understand, the nature of the material being read, etc.). The procedures include whatever operations are needed to convert the meanings of individual phrases or sentences into propositions (McKoon & Ratcliff, 1992, 1998), but also the variety of possible inferences that are necessary to construct an accurate situation model of the text that is being read (Graesser, Singer, & Trabbaso, 1994; Graesser & Zwaan, 1995). Additionally, these procedures include the executive routines that are required to coordinate these high-level "comprehension" processes with the systems that identify words and process sentences, and that control the allocation of attention, the encoding of new information into memory, and the movement of the eyes (Reichle, 2021). It is perhaps for that reason that comprehension has been largely ignored even within models that attempt to, for example, explain how the process of identifying words affects (and affected by) eye-movement behaviour (e.g., Reilly & Radach, 2006; Snell,

Leipsig, Grainger, & Meeter, 2018) and/or sentence processing (e.g., Just & Carpenter, 1987; Van Dyke & Lewis, 2003). Indeed, models of text comprehension tend to be fairly limited in their explanatory scope, with the models most often being limited to simulating the encoding and recall of text propositions from memory (e.g., Kintsch, 1988; Kintsch & van Dijk, 1978) or the making of certain types of automatic (e.g., Goldman & Varma, 1995; Myers & O'Brien, 1998) or controlled (e.g., Langston & Trabasso, 1999; Schmalhofer, McDaniel, & Keefe, 2002) inferences.

In closing this section, one might contrast text comprehension with other core reading processes by way of analogy to the sport of football. For example, if one compares the processes of word identification, sentence processing, and the control of attention and eye movements to individual football players, then text comprehension might be likened to the overall performance of the team; although performance of individual team members is certainly predictive of overall team performance, the relationship between the two will not be perfect because the team's performance also depends upon a number of variables that simply cannot be predicted using information about individual players (e.g., how well they cooperate on the field). We therefore suspect that attempts to accurately measure comprehension will, like measures of a team's overall performance, remain imperfect to the extent that those measures fail to capture variables that reflect the coordinated activities of the whole ensemble—activities that are not captured by measuring the behaviours of individuals even though their performance contributes to that that of the overall group. Bearing this football analogy in mind, we now turn to the implications of our findings for reading instruction.

Implications for reading pedagogy

Continuing with the football analogy, if there is no ideal way to measure "reading comprehension," what would a person do if they were, say, working as a reading support professional in the Premier League Reading Stars program, and needed to determine if a

young footballer was falling behind in their reading comprehension? At this stage of our imperfect understanding, we would suggest that a good starting point would be to be fully aware that there is no perfect assessment of reading comprehension. A reading comprehension test may have the most up-to-date norms, may be most popular with colleagues, may be most affordable, or most accessible, or simply an old favourite, but none of those factors produce an ideal reading comprehension test for two reasons. First, as the outcomes of the current study help demonstrate, different reading comprehension tests tax different cognitive processes and different task demands. Second, the nature of reading comprehension varies across reading development and across situations. For example, testing if a child can read everyday texts, such as school menus or birthday party invitations, is a far cry from testing if an adult can understand paragraphs from 18th century historical texts. Thus, at this point in time, the most sensible way to assess reading comprehension is to find a test that uses a combination of texts and task demands that most closely matches the type of reading comprehension under scrutiny. For example, if our young footballer was failing in class, it would be worth consulting with the teacher to understand what type of reading comprehension was of concern. However, if the footballer's parents were concerned that the child was struggling to understand everyday texts, such as text messages, then an assessment that uses everyday texts might be more useful.

Of course, this approach requires extra effort from reading professionals, who are often already over-stretched. This is why we—like many others—have not given up on the idea of finding a simpler way to measure reading comprehension in all readers in a relevant and ecologically-valid way. This is one of the long-terms goals of the research in this study. However, we have some way to go. At this stage, we believe that eye movements from a reading comprehension test could allow us to infer that someone is struggling with, for example, basic word identification. This inference would be based on common markers that

are indicative of, for example, a less-skilled reader or someone reading text that is too difficult (e.g., more, longer fixations, shorter saccades, more regressions, etc.; see Reichle et al., 2013). However, to definitively know that these eye-movement patterns (which can reflect difficulty at many different levels of text processing) actually stem from difficulty with *lexical* processing, the sentences and discourse would ideally be very simple so that any irregularities in the eye movements cannot be attributed to, for example, difficulty with syntactic parsing or the making of inferences. Further, the reading comprehension test would need to devoid of any secondary task demands (e.g., reading aloud or guessing specific target words). Using the latter criterion, we would probably not use the GORT or WRAT, but would instead opt to use the YARC. But if possible, we would also update the YARC passages for easier-to-read (simple) declarative texts, such as any of the short stories by Rad Bradbury. And we would inform the person being tested that they must understand the text during the first pass well enough to answer comprehension questions. (In other words, they cannot refer back to the text when answering the questions.) Finally, the texts would ideally begin with common words and very simple sentences, but then progress to very rare words, more difficult sentence structures, etc. This would allow a within-person contrast of eye-movement markers to identify where in the text the reading starts to become too difficult. By adopting this basic approach, one might hope to develop texts that contain a variety of embedded diagnostics that would allow the clinician to very rapidly infer the specific types of difficulties a struggling reader is experiencing. This would then also allow the clinician to recommend specific types of remedial training to address those difficulties in a targeted manner, somewhat analogous to how individualized medicine (e.g., gene therapy) is being used to treat various types of disease.

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). The history of psychology is filled with attempts to measure such high-level abilities and traits (e.g., general intelligence; Wechsler, 1958), so it should not be surprising that attempts to measure comprehension, like these other efforts, are inherently limited in terms of both their reliability and validity. As our study demonstrates, these limitations reflect the multitude of ways in which the contents of a text might be used to support behaviour that is indicative of a reader having understood a text, and because the behaviours that are required to extract meaning from text are also highly adaptive and thus quite varied in nature. Attempts to measure text comprehension, therefore, should not be made without careful consideration of a reader's background knowledge and goals because, without such consideration, inferences about a reader's capacity to comprehend text will be as misdirected as, for example, inferences about someone's problem-solving ability without knowing if the person knew the answers in advance or was even attempting to solve the problems. Despite these admonishments, however, we believe that 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.

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.

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review*, *105*, 678-723.

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.

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.

Catts, H. W., Herrera, S., Nielsen, D. C., & Bridges, M. S. (2015). Early prediction of reading comprehension within the simple view framework. *Reading and Writing, 28*(9), 1407–1425.

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 (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 341–374). Amsterdam, Netherlands: Elsevier Science Ltd.

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*, 605–634.

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

Copeland, L., & Gedeon, T. (2013). Measuring reading comprehension using eye movements. *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings*, 791–796.

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

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).

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.

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178-210.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.

Georgiou, G. K., Das, J. P., & Hayward, D. (2009). Revisiting the "simple view of reading" in a group of children with poor reading comprehension. *Journal of Learning Disabilities, 42*(1), 76–84.

Goldman, S. R. & Varma, S. (1995). CAPping the construction-integration model of discourse representation. In C. Weaver, S. Mannes, & C. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 337-358). Hillsdale, NJ, USA: Erlbaum.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6–10.

Graesser, A. C., Singer, M., & Trabbaso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395.

Graesser, A. C. & Zwaan, R. (1995). Inference generation and the construction of situation models. In C. A. Weaver, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in homor of Walter Kintsch* (pp. 117-139). Hillsdale, NJ, USA: Erlbaum.

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.

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.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329-354.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Newton, MA, USA: Allyn and Bacon.

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., McMaster, K. L., & Christ, T. J. (2016). Reading Comprehension: Core Components and Processes. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 62–69.

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., & Petscher, Y. (2020). Influences of individual, text, and assessment factors on text/discourse comprehension in oral language (listening comprehension). Annals of Dyslexia.

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.

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

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Ko, M., H.,. (2010). A Comparison of Reading Comprehension Tests: Multiple-Choice vs. Open-Ended. *English Teaching, 65*(1), 137–159.

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

Langston, M. C. & Trabasso, T. (1999). Modeling causal integration and availability of information during comprehension of narrative texts. In H. van Oostendrop & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 29-69). Mahwah, NJ, USA: Erlbaum.

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.

Martínez-Gómez, P., & Aizawa, A. (2014). Recognition of understanding level and language skill using measurements of reading behavior. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 95–104.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*, 440-466.

McKoon, G. & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistics research in the 1990s. *Annual Review of Psychology*, *49*, 25-42.

Myers, J. L. & O'Brien, E. O. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131-157.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: the validity and utility. *British Journal of Educational Psychology, 67*(3), 359–370.

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91–121.

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, Titus (2021). Monolingual and bilingual reading strategies in Russian: An exploratory scanpath analysis. Reading Research Quarterly, 57(2). http://dx.doi.org/10.1002/rrq.414

Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork

Peng, P., Barnes, M., Wang, C. C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). Meta-analysis on the relation between reading and working memory. *Psychological Bulletin, 144*(1), 48–76.

Perfetti, C. & Stafura, J. (2014). Work knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22-37.

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

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, Vienna, Austria.

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research*, *72*(6), 675–688. https://doi.org/10.1007/s00426-008-0173-3

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*, 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*, 720-732.

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., & 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.

Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford, UK: 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.

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

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from survival analysis of fixation durations. *Cognitive Psychology*, *65*, 177-206.

Roy-Charland, A., Colangelo, G., Foglia, V., & Reguigui, L. (2017). Passage independence within standardized reading comprehension tests. *Reading and Writing, 30*(7), 1431–1446.

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*, 1270-1281.

Schmalhofer, F., McDaniel, M. A., & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes*, *33*, 105-132.

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. https://doi.org/10.1177/0956797614531148

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147–170.

Slattery, T. J., & Yates, M. (2018). Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology, 71*(1 Special Issue), 250–259.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-Reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*, 969-984.

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).

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.

Spivey, M. J. & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1521-1543.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116*(1), 71–86.

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. & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285-316.

Vasishth, S., von der Malsburg, Titus, & Engelmann,F. (2013). What eye movements can tell us about sentence comprehension. Wiley Interdisciplinary Reviews: Cognitive Science, 4(2), 125–134. http://dx.doi.org/10.1002/wcs.1209

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.

Veldre, A., & Andrews, S. (2016). Semantic preview benefit in English: Individual differences in the extraction and use of parafoveal semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 837.

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.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence (4th edition)*. Baltimore, MD: Williams & Witkins.

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., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & 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)*. Austin, TX: Pro-Ed.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test 4 (WRAT4).* Psychological Assessment Resources, Lutz.

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.

Wolf, D., F. (1993). A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension. *The Modern Language Journal, 77*(4), 473–489.

Appendix A

This Appendix provides additional information about the scoring and reliability of our three reading comprehension tests: the YARC, GORT, and WRAT.

YARC

Scoring: The raw scores for comprehension, reading rate, and summary were transformed into ability scores. The ability scores for comprehension and reading rate were then converted into standard scores, with a mean of 100 and a standard deviation of 15. No standard score was calculated for the summary task. All scoring was done with the YARC online conversion tool (https://rgt.testwise.net/YARC_Aust_Pri_index.htm).

Test reliability: The test reliability of the YARC was calculated using Cronbach's alpha. The overall reliability coefficients for the passages used in this study was 0.71 for comprehension, and 0.77 for the summary, which indicates good reliability. The validity of the comprehension questions was assessed by calculating the percentage of participants who could answer the questions correctly without reading the text. For the items used in this study, the percentages of participants answering correctly ranged from 0 to 18.6%, with a mean of 2.5%. The validity of the summary questions was assessed by correlations with the comprehension scores. Because the correlations were low for all items, the (original test) authors note that the summary scores may not be entirely reliable and should be used with caution.

GORT

Scoring: The maximum score was 5 for reading rate, accuracy, and comprehension, and 10 for fluency. For reading rate and accuracy, each score (0-5) corresponds to a range of reading times and number of errors indicated in the manual, thus participants are assigned a score based on these cut-offs. For example, a participant who reads item 6 in 45 seconds and makes 3 reading errors would receive a score of 5 for reading rate, 4 for accuracy, and 9 for

fluency. The comprehension score is the sum of all questions answered correctly. These scores were then transformed into scaled scores as per the manual instructions. These scores have a mean of 10 and a standard deviation of 3. Participants who scored 8-12 were classified as ‘average’, 7-6 as ‘below average’, and 5-4 as ‘poor’. The sum of the comprehension and fluency scaled scores was then calculated to obtain a participant’s reading oral index scores from the manual. This index is a standard score with mean of 100 and standard deviation of 15.

Test reliability: Test reliability was calculated with Cronbach’s alpha for reading rate, accuracy, comprehension, and fluency, and with Guilford’s alpha for the oral reading index. The coefficient alphas range from 0.91 to 0.97, indicating good reliability. Test validity was informed by the percentage of questions that could be answered correctly without reading the passages. The majority of the questions (88%) were only answered correctly by 10% or less of the sample, indicating that most of the questions were passage-dependent.

WRAT

Test Reliability: Test reliability was measured with Cronbach’s alpha. The reliability coefficients for the word reading and sentence comprehension subtests are respectively 0.92 and 0.93 for the green form, indicating good reliability.