Hammill Institute on Disabilities

Assessment for Effective Intervention
2018, Vol. 43(3) 169–181
© Hammill Institute on Disabilities 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1534508417728685
aei.sagepub.com
SAGE

Preliminary Findings on the Computer-Administered Multiple-Choice Online Causal Comprehension Assessment, a Diagnostic Reading Comprehension Test

Mark L. Davison, PhD[1], Gina Biancarosa, EdD[2], Sarah E. Carlson, PhD[2], Ben Seipel, PhD[3,4], and Bowen Liu, BA[1]

Abstract

The computer-administered Multiple-Choice Online Causal Comprehension Assessment (MOCCA) for Grades 3 to 5 has an innovative, 40-item multiple-choice structure in which each distractor corresponds to a comprehension process upon which poor comprehenders have been shown to rely. This structure requires revised thinking about measurement issues (e.g., reliability and interpretation of incorrect responses for diagnostic purposes). Using data from a pilot study, the article presents descriptive statistics on correct responses, incorrect responses, and comprehension rate. It also presents reliability data for correct responses and incorrect responses as well as construct validity data on correct responses. Implications for diagnosis and remediation of poor inferential comprehension are discussed.

Keywords

computerized assessment, diagnostic/analytic, reading/literacy

The ability to make causal inferences and track the complexities of causality is fundamental to reading comprehension of narrative text (van den Broek, 1997). Successful causal connections and inferences reflect a reader's ability to monitor character goals, discern physical changes, and recognize cause and effect (Trabasso & van den Broek, 1985; van den Broek, 1997). However, some readers struggle to make causal connections in text (Cain & Oakhill, 1999, 2006; McMaster et al., 2012; Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007).

Making causal connections in text is a higher-level reading skill. Previous research has found a group of struggling readers who do not struggle with lower-level reading skills (e.g., decoding), but rather struggle with higher-level reading skills such as making causal connections during reading (Cain & Oakhill, 2006; Perfetti, 2007). Commonly termed *poor comprehenders*, they exhibit poor comprehension compared with peers with similar word-reading skills and vocabulary (e.g., Cain & Oakhill, 1999, 2006; Carlson, Seipel, & McMaster, 2014; Rapp et al., 2007). Poor comprehenders represent at least 10% of elementary grade readers, have adequate decoding skills, and have below-average comprehension skills (Cain & Oakhill, 1999, 2006; Carlson et al., 2014; Catts, Hogan, & Adlof, 2005; Hulme & Snowling, 2011; Pimperton & Nation, 2010; Rapp et al., 2007). Prior research indicates poor comprehenders are not all the same; they can be differentiated by the types of cognitive processes (e.g., prediction, elaboration, or paraphrasing) on which they rely during reading when trying to make causally coherent inferences (Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). Importantly, poor comprehenders have also been shown to respond differentially to intervention based on their preferred cognitive processes during reading (McMaster et al., 2012; van den Broek et al., 2006).

Existing assessments of reading comprehension distinguish between good and poor comprehenders. However, if a student is identified as a poor comprehender, current assessments often give little information as to *why* the student is a poor comprehender. As a result, teachers receive little diagnostic information that might shape instruction. Until now, identifying the cognitive processes on which poor comprehenders rely, and which could therefore inform intervention efforts, has mainly come from the use of

[1]University of Minnesota, Minneapolis, USA
[2]University of Oregon, Eugene, USA
[3]California State University, Chico, USA
[4]University of Wisconsin–River Falls, USA

Corresponding Author:
Mark L. Davison, Quantitative Methods in Education, Department of Educational Psychology, University of Minnesota, 56 E. River Rd., Minneapolis, MN 55455, USA.
Email: mld@umn.edu

think-aloud protocols (e.g., Ericsson & Simon, 1993; Graesser & Clark, 1985; Graesser, Singer, & Trabasso, 1994; van den Broek, 1990). However, because think alouds require one-to-one recording of responses, transcription, and coding, they are too time-consuming and labor-intensive for regular classroom use.

The goal of the present project was to develop an innovative reading comprehension assessment that is both diagnostic for poor comprehenders and usable in practical educational settings, thereby providing research and practice with a practical assessment that measures why readers struggle with comprehension. The new assessment is a multiple-choice assessment of reading comprehension in which the distractors represent cognitive processes identified in prior think-aloud research as distinguishing between types of poor comprehenders (Carlson et al., 2014). The Multiple-choice Online Causal Comprehension Assessment (i.e., MOCCA) addresses the limitations of prior research by providing reliable scores indicating the types of errors to which students are prone. It also provides information on comprehension rate, a potential indicator of automaticity.

Poor Comprehenders and the Need for Informative Assessments

Prior research has investigated how poor comprehenders differ in the cognitive processes they use to integrate text information with prior knowledge in the creation of a coherent mental model of the text (e.g., Graesser et al., 1994; Kintsch & van Dijk, 1978). This research has revealed different types of poor comprehenders, who differ in the cognitive processes (e.g., predicting, paraphrasing, elaborating) on which they rely, and in some cases, over rely (Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). One group, commonly termed "paraphrasers," is defined by their overuse of paraphrases and related text-based processes during reading. The other group—originally termed as "elaborators"—we call "lateral connectors" because lateral connections include more than just elaborations; they also include personal associations, predictions, explanations, and evaluations of textual content. Paraphrases do not involve making an inference; lateral connections can involve making an inference, but inferences that do not complete the story in a causally coherent way. While poor comprehender research is ongoing and this list of poor comprehender types is likely not exhaustive, research suggests these two groups respond differentially to comprehension interventions, meaning that one size does not fit all when it comes to improving the reading of poor comprehenders (McMaster et al., 2012; van den Broek et al., 2006). While paraphrasers' comprehension improves more than lateral connectors when prompted with questions promoting general connection-making, lateral connectors improve more than paraphrasers when prompted with causal questions focused on the causal structure of a text (McMaster et al., 2012). Thus, distinguishing between types of poor comprehenders holds instructional relevance.

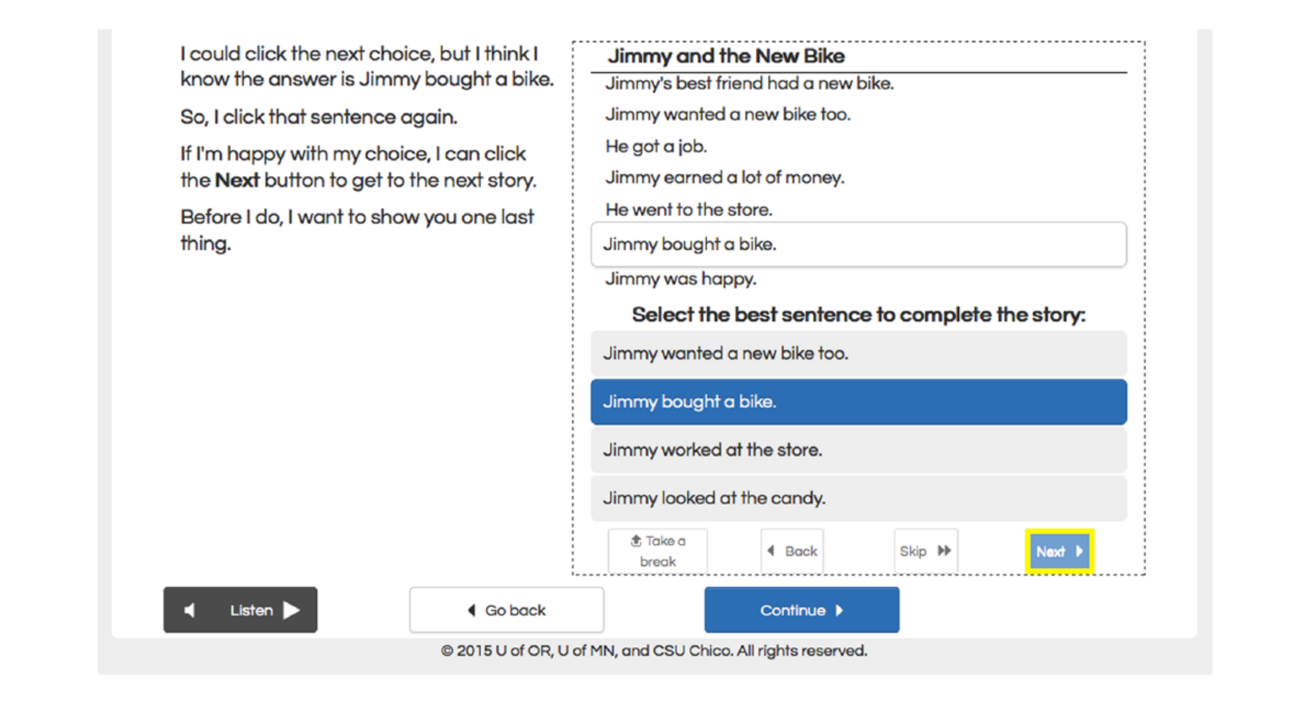
Importantly, prior research distinguishing paraphrasers and lateral connectors (i.e., elaborators) relies on think-aloud protocols. Think alouds ask readers to verbalize what they are thinking as they read (Ericsson & Simon, 1993; Graesser & Clark, 1985; Graesser et al., 1994; van den Broek, 1990). What readers say about their thinking reveals poor comprehenders' predilections for specific types of comprehension processes. However, think alouds are not a practical assessment for classroom teachers. In addition to the time it takes for each child to think aloud individually for two or more texts, each child's statements must also be transcribed, parsed into idea units, and then coded for the cognitive process each unit represents. Such protracted work on the part of a teacher, or even a specialist, is simply not realistic.

At the same time, researchers and educators have long called for innovation in reading comprehension measurement (Klingner, 2004; Pearson & Hamm, 2005; RAND Reading Study Group, 2002; Sarroub & Pearson, 1998; Snyder, Caccamise, & Wise, 2005; Williams, 2006; Wixson, Valencia, & Lipson, 1994). These calls assert our methods for assessing and studying reading comprehension have changed little since standardized measurement of reading comprehension began. Existing measures focus almost exclusively on the end *product* of reading comprehension. That is, comprehension is judged by the understanding a reader can demonstrate *after* reading is finished. Such measures provide little information about the *process* by which a reader constructed meaning *during* reading. The focus on comprehension as a product in traditional reading comprehension measures yields limited insight into how a poor comprehender went astray during reading.

MOCCA

Thus, there is a need for measures of the reading comprehension *process* that yield precisely the sort of information that think alouds yield. At first glance, MOCCA is a straightforward multiple-choice reading comprehension assessment that uses very short stories to gauge intermediate grade children's reading comprehension. However, MOCCA is innovative in that stories and answer choices are crafted to assess a child's ability to maintain causal coherence when reading and any predilection for specific cognitive processes when causal coherence is not maintained, thereby providing information on poor comprehenders similar to that generated by think alouds. MOCCA accomplishes this by using informative distractors.

Informative distractors are not a new idea (e.g., Delmas, Garfield, Ooms, & Chance, 2007; Hermann-Abell & DeBoear, 2011; Hestenes, Wells, & Swackhamer, 1992;



I could click the next choice, but I think I know the answer is Jimmy bought a bike.

So, I click that sentence again.

If I'm happy with my choice, I can click the **Next** button to get to the next story.

Before I do, I want to show you one last thing.

Jimmy and the New Bike

Jimmy's best friend had a new bike.

Jimmy wanted a new bike too.

He got a job.

Jimmy earned a lot of money.

He went to the store.

Jimmy bought a bike.

Jimmy was happy.

Select the best sentence to complete the story:

Jimmy wanted a new bike too.

Jimmy bought a bike.

Jimmy worked at the store.

Jimmy looked at the candy.

© 2015 U of OR, U of MN, and CSU Chico. All rights reserved.

Figure 1. Screen shot of MOCCA item in directions.
Note. MOCCA = Multiple-Choice Online Causal Comprehension Assessment.

Sadler, 1998), but MOCCA implements the idea in a novel fashion. MOCCA focuses on a small number of comprehension processes (e.g., paraphrases, lateral connections), and it more thoroughly evaluates a poor comprehender's tendencies to engage in those processes by including paraphrase and lateral connection options in every item. Prior assessments using informative distractors, called distractor-driven assessments by Hestenes et al. (1992) and concept inventories by Sadler (1998), included a large number of misconceptions as distractors with any one misconception being represented in one or only a very small number of items. While these assessments might yield reliable total correct answer scores, to our knowledge there is little or no evidence that they produce reliable scores reflecting a student's propensity toward a particular incorrect response type. Because each incorrect response type is represented in each item, MOCCA might yield reliable scores reflecting student propensities toward *each particular type of response*.

Each of the 40 MOCCA items is a seven-sentence story that is missing its sixth sentence and therefore has a causal gap that the reader must fill. Multiple-choice responses are alternative sentences that might fit as the sixth sentence. In the original version of MOCCA and the pilot study reported here, the responses represent four types of cognitive processes: (a) a *causally coherent inference*, (b) a *paraphrase*, (c) a *lateral connection*, and (d) a *local bridging inference*. Causally coherent inferences are always the best response to complete the story in a causally coherent manner. Without the type of inference made in this sentence, the story does not completely make sense. Paraphrases are always incorrect to complete the sixth sentence, but mimic what one group of poor comprehenders tend to do when they do not make causally coherent inferences. Paraphrases primarily restate the main goal of the story's main character, but can paraphrase subgoals, updated goals, or the goals of other characters. Lateral connections are also always incorrect, but mimic what the other group of poor comprehenders tend to do when they do not make causally coherent inferences. Lateral connections may represent valid or invalid elaborative inferences, associations, or explanations based on the information found in the fifth sentence of the story; none of which completes the causal chain of events. Note that readers may draw on background knowledge to make either causally coherent inferences or lateral connections, but one fills the causal gap and the other does not. The local bridging distractor type was developed as a fourth response type because they appear in think alouds that all readers generate.

Figure 1 shows a sample item, a story called "Jimmy and the New Bike." Answer 1, "Jimmy wanted a new bike." is the paraphrase and simply states Jimmy's goal explicitly expressed in the second sentence of the story. Answer 2, "Jimmy bought a new bike." is the correct, causally coherent answer. Answer 3 "Jimmy worked at the store" is the local bridging response as it connects to the fifth sentence but does not complete the story. The last answer "Jimmy looks at the candy" is the lateral connect response, an association with candy that does not advance the story. As with all MOCCA stories, each story aligns to a single item avoiding problems associated with a testlet structure that arises when several items refer to a single story.

In addition, because the measure is now computer administered, MOCCA captures data regarding the time children expend on each item and the assessment overall. Of most relevance, pedagogically is the overall time taken to complete the assessment. Divided by the number of items a student gets correct (out of 40 possible items), a reading comprehension rate is derived (minutes per correct response). Although rate is not often considered on reading comprehension assessments, some have argued for its value in measuring children's automaticity (also called efficiency) in reading comprehension (Skinner, Neddenriep, Bradley-Klug, & Ziemann, 2002).

Rate is also a relevant aspect of comprehension because automaticity (i.e., effortless processing; LaBerge & Samuels 1974; Posner & Snyder, 1975) applies to more than just low-level skills like decoding. Retrieval of the meanings of individual words is also ideally automatized (i.e., lexical access; Perfetti, 2010; Perfetti & Lesgold, 1979), as are many of the inferences readers draw as they read (Thurlow & van den Broek, 1997). Thus, readers may be differentiated based not only on how well they comprehend (i.e., how many items they get correct on a test), but also by how efficient they are in executing comprehension processes (i.e., how automatic reading comprehension processing is). As a result, Skinner and colleagues have proposed reading comprehension rate as an analogue to oral reading rate in Grades 4 and beyond, but more importantly as a measure of the efficiency of the reading comprehension process itself that may be more sensitive to change over time than traditional reading comprehension scores (e.g., Neddenriep, Hale, Skinner, Hawkins, & Winn, 2007; Skinner et al., 2002). Rate measures have yielded a number of empirical validations from a range of researchers interested in curriculum-based measurement (e.g., Cianco, Thompson, Schall, Skinner, & Foorman, 2015; Hale et al., 2011; McCane-Bowling, Strait, Guess, Wiedo, & Muncie, 2014; Skinner et al., 2009). There are several ways to measure rate, most of which are variants of minutes per correct or it's inverse, correct per minute.

A limitation of Skinner's approach is that the resulting scale yielded by the suggested formula (i.e., percentage correct per minute) is not easy to interpret and can vary wildly depending on passage and item lengths and the number of questions asked or attempted. For an assessment like MOCCA, where passages are quite short and consistent in length (i.e., seven sentences) with each passage acting as an item, the percentage correct per minute is generally quite small, even for the speediest readers with excellent comprehension. Thus, for MOCCA we opted to calculate rate as the inverse of the way in which oral reading rates are calculated. Instead of dividing number of items correct by total testing time, we divided total testing time by the number of items answered correctly, which yields a minutes-per-correct-item rate.

Challenges in the MOCCA Measurement Model

MOCCA is not the first attempt to create multiple-choice distractors that correspond to errors students make, misconceptions that students hold, or other diagnostic information. Most prior examples come from the sciences or statistics (e.g., Delmas et al., 2007; Hermann-Abell & DeBoear, 2011; Hestenes et al., 1992; Sadler, 1998). In this research, the assessments generally had reliable total scores for the number correct, but did not provide scores indicating the frequency with which a student chose answers representing a particular kind of misconception or error (i.e., the distractors). Nor did they provide much guidance for interpreting student responses in terms of misconceptions or errors.

While distractor-driven assessments seem promising, several problems need to be addressed before their potential can be realized. First, assessments need to provide scores or some other diagnostic indicator that represents information about a student's tendency to select responses corresponding to various processes or errors. Second, the incorrect response scores need to be reliable, and there needs to be an accepted procedure for quantifying the reliability. Third, an appropriate item response theory (IRT) needs to be decided upon and used to address various applied problems, such as equating of forms.

The need for an appropriate IRT is recognized in both the applied test construction literature and the psychometric literature. In analyzing their chemistry items, Hermann-Abell and De Boear (2011) applied the Rasch model to the correct responses. They concluded that, misconceptions and correct responses in their domain may be somewhat hierarchically arranged and that some misconceptions may represent steps in a natural progression toward the development of a correct conception. In a similar study, Sadler (1998) applied Bock's (1972) nominal response model to account for a correct response, misconception responses for each item, and a "don't know" response. The psychometric literature now includes several additional options for IRT models that go beyond the Rasch, 2PL, and 3PL in that they model incorrect as well as the correct response to a multiple-choice item (Adams, Wilson, & Wang, 1997; Bock, 1972; Bradshaw & Templin, 2014; Brown, 2016; Johnson & Bolt, 2010; Samejima, 1979); but, it is still an open question as to whether familiar models, such as the Rasch, 2PL, or 3PL model, can be adapted to incorrect responses or whether one of the newer models (e.g., Adams et al., 1997; Brown, 2016) will prove more useful.

Beyond IRT model choice there are a number of additional unresolved issues, such as determining cut-scores as well as use and interpretation of the comprehension efficiency scores and incorrect response scores. For this pilot study, we present some preliminary descriptive, reliability, and construct validity data on the number correct scores (causally coherent inferences), the incorrect response scores (paraphrases, lateral connects, local bridging inferences), and the comprehension efficiency scores to establish that it is feasible to develop scores based on incorrect responses that are reliable, that MOCCA displays construct (both convergent and discriminant) validity, and that scores based on incorrect responses and comprehension efficiency may be sensitive to instructional/developmental effects across grades.

The Current Study

We report results from a large pilot administration of the new computerized MOCCA. One goal was to assess whether it is possible to obtain reliable scores based on incorrect item responses using either simple total scores or scores based on IRT. A second goal was to assess whether comprehension efficiency scores and scores based on incorrect responses are sensitive to instructional and developmental effects across grades. The third goal was to begin examining convergent and discriminant validity.

Method

Participants

A convenience sample of 920 students in Grades 3 to 5, recruited through emails and personal contacts, took MOCCA online during spring 2015: 341 third-, 327 fourth-, and 252 fifth graders. In all, 48% were male and 52% female. Our sample was predominantly White, but two districts had approximately 20% Hispanic students. Other minority groups were underrepresented relative to their numbers nationally. All of the districts had 48% or more of their students eligible for free or reduced lunch (economically disadvantaged). Although MOCCA is a tool for addressing poor comprehension, the sample was not limited to poor comprehenders.

Instrument

MOCCA was informed by prior think-aloud research (McMaster et al., 2012; Rapp et al., 2007) and traditional curriculum-based measurement research (e.g., Deno, 1985). It is a multiple-choice, online assessment designed to identify comprehension processes used during reading narrative texts (Carlson et al., 2014). Items are narrative texts with a causal structure centered *on a main goal and motivated subgoals and events* (e.g., Trabasso & van den Broek, 1985). Instead of deleting every *nth* word as in traditional cloze or maze tasks, the sixth *sentence* of each seven-sentence text was deleted.

Within a grade, stories are assigned to forms so that the average story reading level and number of words is as nearly equal as possible. Within the reading level and number of words constraint, stories were randomly assigned to forms within a grade. All stories have exactly seven sentences with one missing (i.e., the sixth sentence). For each grade, story reading levels range from one level below grade to one level above grade. For instance, Grade 3 forms contain stories with reading levels from Grades 2 to 4 with a mean of 3.0 on the Flesch–Kincaid scale (Kincaid et al., 1975).

Participants are instructed to choose one of four alternative response types to fill in the deleted sentence. As described in more detail in the literature review, in addition to the correct answer (i.e., a causally coherent inference), two of the three remaining response types are informative distractors: a *paraphrase* and a *lateral connection*. The final response type, a *local bridging inference*, does not correspond to a response type that has differentiated comprehenders in think-aloud research.

The original paper–pencil MOCCA was a single form for Grades 3 to 5 students (Carlson et al., 2014), but for the current pilot we developed 12 computer-administered forms with four at each grade level. Each form had a forward and a backward order, and response types were randomized per item. Participants were randomly assigned to a form at their grade level. Multiple forms facilitate progress monitoring without administering the same form more than once to a student. As with the original MOCCA, each pilot form included 40 items.

MOCCA yields four scores, each representing the number of times the reader selects a response type. In other words, for each response type (causally coherent inference, paraphrase, lateral connection, and local bridging inference) a point is tallied when that response type is chosen. The causally coherent inference is the correct answer, but for identifying patterns in the types of incorrect choices students make, each of the other response types also gets scored.

Procedure

MOCCA was administered on a newly designed online platform. Participants with parental consent took MOCCA as a whole group—either in their school computer lab or in their classrooms on computers or tablets. Each administration was proctored by trained project staff or the classroom teacher, and administration was standardized via the online program. Before beginning the test, students read and/or listened to written instructions on the computer. Students could choose to have the instructions read orally by clicking an option on the screen. The instructions included sample

Table 1. Mean and Standard Deviation (in Parentheses) for Raw Scores of Each Response Type and Comprehension Rate by Grade and Form.

Grade.form	Causal	Paraphrase	Lateral	Local bridging	Comprehension rate
3.1	21.27 (11.20)	4.25 (3.96)	4.05 (3.38)	3.91 (3.48)	2.41 (2.83)
3.2	16.84 (9.99)	5.52 (4.39)	4.44 (3.56)	6.20 (4.00)	2.76 (2.91)
3.3	17.89 (10.05)	4.83 (4.11)	4.13 (3.22)	4.93 (4.04)	3.04 (4.79
3.4	17.03 (9.66)	5.20 (4.56)	3.74 (3.45)	5.49 (4.02)	2.87 (3.86)
4.1	23.55 (11.20)	3.77 (4.18)	3.35 (3.06)	4.66 (4.03)	1.94 (1.85)
4.2	25.37 (11.74)	2.95 (3.67)	3.00 (3.50)	3.15 (4.28)	1.75 (1.38)
4.3	21.35 (9.77)	4.57 (4.31)	4.56 (3.69)	5.05 (3.67)	2.21 (2.26)
4.4	24.08 (10.31)	4.07 (3.60)	3.73 (3.16)	4.01 (4.20)	1.60 (1.07)
5.1	21.87 (10.81)	3.93 (4.06)	3.44 (3.25)	4.28 (3.68)	1.59 (0.97)
5.2	23.27 (11.41)	2.97 (3.25)	3.12 (2.79)	3.56 (3.50)	1.84 (1.54)
5.3	24.19 (10.46)	2.50 (3.48)	3.50 (2.76)	3.80 (3.67)	1.76 (2.02)
5.4	27.89 (11.43)	2.04 (3.56)	2.22 (2.26)	2.12 (2.82)	1.44 (1.23)

Note. Across the rows, means of response types do not sum to 40 because students skipped some items, in which case the item response does not fall into any of our four categories: causal, paraphrase, lateral connect, or local bridging.

items and illustrated how to complete items and navigate through the program. Students were told that they were going to read several short stories, that each story had a missing sentence, and that their job was to pick one out of four sentences below each story that best completed the story. They could click on each of the response types to see the sentences within the context of the story. Once they were happy with their choice, they clicked a "NEXT" button to move to the next item. Participants were able to skip items and were allowed to take a break by pausing the program. The test itself imposes no time limit, but teachers could limit the amount of time to approximately one period (i.e., 30–60 min) with most allowing approximately 45 min.

Results

The mean testing time varied by form from 28.43 to 33.96 min across grades. In Grades 3 to 5, students answered an average of 32.43, 35.31, and 33.68 items respectively of the 40 possible.

Grade Trends

Table 1 shows the mean of the four response type scores by form. The means for the four response types do not add to 40, the number of items on the test, because students could skip items. Especially for Grade 3, the test is fairly difficult. For instance, for Form 3.1, the mean is 21.27 for the Causally Coherent Inference score, indicating that, on average, students correctly answered 53% of the items. For Forms 3.2 to 3.4, students correctly answered slightly less than 50% of the items. If one is to glean diagnostic information from mistakes that students make, students must make some mistakes. Consequently, diagnostic tests based on incorrect answer patterns may need to be somewhat harder than most existing achievement tests.

To explore the difficulty of each response type across forms and grades, we tested for mean differences using an alpha of .05. Four univariate analyses of variance (ANOVAs) were performed, one for each score with grade and form nested within grade as the independent variables. The grade effect was significant: $F(2, 908) = 30.03$, 19.98, 7.12, and 14.14, respectively, for the causally coherent inference, paraphrase, lateral connection, and local bridging inference scores. Despite the fact that passage-reading level increased with grade, the causally coherent inference scores generally increased with grade. Correspondingly, the paraphrase, lateral connection, and local bridging inference scores all generally decreased. While the results are cross-sectional, they suggest that MOCCA scores, including the incorrect response type scores, may be sensitive to developmental age effects, instructional grade effects, or both.

Despite randomly assigning items to forms within grades, the form effect within grade was also significant for every score: $F(9, 908) = 3.12$, 2.19, 4.46, and 2.16, respectively, for the causally coherent inference, paraphrase, lateral connection, and local bridging inference scores. For instance, the causally coherent response, Form 1 in Grade 3 had a mean more than three points higher than any other form within that grade.

Reliability

Table 2 shows the internal consistency reliability, alpha, for each form and score. As not-reached items at the end of the test can inflate alpha, it was computed using only students with complete item responses and yielding sample sizes of 55 to 95 per form. The causally coherent inference score,

Table 2. Internal Consistency Reliability (Alpha) for Each Form and Grade.

Grade form	Causal	Paraphrase	Lateral	Local
3.1	.95	.84	.71	.74
3.2	.94	.80	.65	.71
3.3	.92	.71	.49	.69
3.4	.94	.80	.74	.70
4.1	.95	.83	.71	.80
4.2	.96	.82	.81	.86
4.3	.93	.78	.68	.67
4.4	.94	.74	.63	.80
5.1	.95	.82	.70	.75
5.2	.95	.79	.68	.80
5.3	.94	.83	.60	.81
5.4	.95	.89	.68	.83

Table 3. Item Response Theory Marginal Reliabilities for Each Form and Grade, One Parameter Logistic and Two Parameter Logistic in Parentheses.

Grade form	Causal	Paraphrase	Lateral	Local
3.1	.92 (.91)	.79 (.78)	.66 (.68)	.71 (.72)
3.2	.91 (.91)	.74 (.77)	.62 (.74)	.67 (.72)
3.3	.92 (.92)	.77 (.77)	.48 (.73)	.74 (.84)
3.4	.91 (.92)	.71 (.77)	.66 (.68)	.71 (.76)
4.1	.91 (.91)	.75 (.72)	.68 (.90)	.77 (.91)
4.2	.90 (.91)	.71 (.90)	.73 (.90)	.75 (.89)
4.3	.92 (.93)	.79 (.90)	.66 (.89)	.63 (.76)
4.4	.91 (.90)	.71 (.89)	.59 (.90)	.75 (.90)
5.1	.90 (.88)	.73 (.91)	.62 (.91)	.69 (.90)
5.2	.91 (.90)	.72 (.89)	.67 (.79)	.75 (.90)
5.3	.89 (.90)	.69 (.90)	.59 (.90)	.69 (.87)
5.4	.86 (.86)	.63 (.88)	.58 (.88)	.65 (.86)

the correct score, has excellent reliabilities for all forms. For this score, all reliabilities are above .90, an excellent reliability coefficient in the measurement community (Thorndike & Thorndike-Christ, 2010). While not as high as the causally coherent inference reliabilities, the paraphrase score reliabilities were all above .70, ranging from .71 to .89. For the lateral connect and local bridging inference scores, the reliabilities range from .49 to .81 and .67 to .83, respectively. Some of these are below our target of .70, the minimum level for a "good" reliability (Thorndike & Thorndike-Christ, 2010). For the lateral connect score, seven of the 12 are below our target value. In the next iteration of test development, we hope to bring the reliability of every score for every form above .70. Results suggest that incorrect response scores can be reliable in distractor-driven assessments.

Item Response Theory Reliability

Rasch and two-parameter logistic (2PL) models were fitted to the pilot response types, fitting the model separately to each response type. While the multiple-choice format would suggest a three-parameter model, given our sample size averaging about 75 people per form, it may not be possible to accurately estimate a guessing parameter for each item. In the IRT analyses, missing responses were treated as missing, not incorrect.

The marginal reliabilities by form and grade for the Rasch model and 2PL model (2PL in parentheses) are shown in Table 3. The marginal reliability is an internal consistency estimate of IRT true score variance over IRT observed score variance. It can also be considered an estimate of the correlation between IRT scores on the test and a parallel form of itself. There are two notable features in this table. First, for the correct answers, the causally coherent inference responses, there is little difference in the marginal reliability for the Rasch and 2PL models. However, there are some large differences in the Rasch and 2PL marginal reliabilities for the paraphrase, lateral connection, and local bridging inference responses. The marginal reliabilities are generally as high or higher for the 2PL model as compared with the Rasch. While we have not shown the item discrimination parameters here, within each incorrect response type, the discrimination parameters vary widely across items and seem to seriously violate the Rasch assumption of equal discrimination parameters leading to the result that the 2PL model often has a higher marginal reliability for the incorrect response types. Second, for the causally coherent inference responses, the marginal reliabilities are almost all above .90 for both the Rasch and 2PL model, and all are above .85. For the incorrect response types, the reliabilities are more consistently above our target reliability of .70 for the 2PL than for the Rasch. With the exception of Forms 3.1 and 3.4, all of the 2PL marginal reliabilities are above .70.

Table 4 shows the Akaike Information Criterion (AIC) fit statistics for the Rasch and the 2PL models. AIC measures a trade-off between two quantities, the fit of the model to the data as reflected in the deviance statistic (−2 log likelihood), and the number of estimated parameters. Or in other words, these statistics reflect the trade-off between model fit and parsimony. In comparing two models, the model with the smaller AIC is preferred. If the model with more parameters, the 2PL model in our case, has a smaller AIC than does the model with fewer parameters, the Rasch model in our case, then it means the 2PL improvement in fit is large relative to the number of extra parameters in the 2PL model and the 2PL model is preferred for its improved fit. Conversely, if the AIC is smaller for the Rasch, the 2PL improvement in fit is small relative to the number of extra parameters, and the more parsimonious Rasch model is preferred. While the results vary over the grades and forms, the AIC tends to be somewhat smaller for the Rasch than for the 2PL model.

Table 4. Akaike Information Criterion for the Rasch and 2PL (In Parentheses) Models.

Grade.form	Causal	Paraphrase	Lateral	Local
3.1	2,831.95 (2,911.22)	1,909.86 (2,061.50)	1,939.71 (2,032.61)	1,916.49 (2,024.84)
3.2	2,833.09 (2,913.16)	2,182.66 (2,288.51)	1,923.65 (2,015.86)	2,331.87 (2,451.12)
3.3	2,378.74 (2,461.11)	1,740.18 (1,846.22)	1,633.74 (1,722.10)	1,845.10 (1,953.46)
3.4	3,043.62 (3,100.36)	2,282.40 (2,385.80)	1,946.34 (2,074.10)	2,380.61 (2,491.39)
4.1	1,842.33 (1,942.87)	1,224.52 (1,328.75)	1,158.23 (1,305.98)	1,399.7 (1,557.63)
4.2	1,607.75 (1,679.48)	1,079.03 (1,242.66)	1,114.93 (1,286.44)	1,117.37 (1,278.98)
4.3	1,870.94 (1,925.31)	1,354.01 (1,472.01)	1,287.93 (1,365.23)	1,484.68 (1,577.15)
4.4	1,744.51 (1,810.30)	1,221.14 (1,369.30)	1,153.98 (1,292.50)	1,167.28 (1,314.61)
5.1	1,846.12 (1,890.23)	1,179.54 (1,318.22)	1,159.55 (1,308.12)	1,268.76 (1,407.45)
5.2	1,842.41 (1,934.72)	1,157.71 (1,361.95)	1,228.07 (1,319.01)	1,278.99 (1,456.70)
5.3	1,736.85 (1,777.02)	878.18 (1,050.75)	1,134.92 (1,249.49)	1,156.73 (1,293.52)
5.4	1,483.40 (1,587.66)	755.23 (933.08)	910.12 (1,049.35)	877.99 (1,020.04)

Note. 2PL = two-parameter logistic.

Comprehension Rate

The last column of Table 1 shows the mean and standard deviation of the Comprehension Rate for each form. Comprehension Rate is defined as the number of minutes spent on MOCCA divided by the number of correct responses. It reflects the average amount of time required to arrive at a correct response; thus, lower times are better. Students may be slow for one of two reasons. They may expend a great deal of time on each item, or they may work rapidly on each item but typically answer one or more items incorrectly before answering one correctly.

To explore grade and form differences in Comprehension Rate, an ANOVA was run with Comprehension Rate as the dependent variable and with grade and form nested within grade as the independent variables. Within grades, Comprehension Rate did not vary significantly across forms, $F(9, 908) = 0.71$, $p = .709$. The grade effect was significant, $F(2, 908) = 16.40$, $p < .001$. As shown in Table 1, the average Comprehension Rate tends to decline from Grades 3 to 5 as one would expect if reading comprehension becomes more automatic as students move up the grades. Results suggest that rate may be sensitive to developmental age or instructional grade effects.

Table 5. Convergent and Discriminant Validity: Correlations of MOCCA Total Correct Score With Reading and Math Scores of Other Tests (Sample Sizes in Parentheses; Cells Contain Corresponding Reading Comprehension and Math Correlations).

State Exam	Third grade	Fourth grade	Fifth grade
Oregon			
Reading CBM	.549**	.612**	.665**
Math CBM	.462**	.430**	.501**
n	36	63	29
ELA state test	NA	.679**	.575**
Math state test	NA	.567**	.467**
n	NA	97	112
California			
ELA state test	NA	NA	.651**
Math state test	NA	NA	.615**
n			72
Reading CBM	NA	NA	.674**
Math state test	NA	NA	.609**
n			73

Note. MOCCA = Multiple-Choice Online Causal Comprehension Assessment; CBM = curriculum-based measure; ELA = English language arts; NA = no data available for that grade, reading test, math test combination. The Oregon state test was the Oregon Assessment of Knowledge and Skills (OAKS). Oregon CBMs are from the easyCBM system.
**$p < .01$.

Construct Validity: Convergent and Discriminant

Table 5 shows the correlation of the MOCCA total correct score with other reading and math tests in seven subsamples. For any grade/test combination, the correlation of MOCCA with the reading test was estimated in the same sample as the correlation of MOCCA with the corresponding math test. Because the subsample sizes within a form were small, correlations were computed aggregating across forms within a grade, so unlike other results above, form differences may have affected these results. Two trends are notable. First, all of the correlations with the criterion reading assessments (Easy curriculum-based measure [CBM] Common Core State Standards [CCSS] comprehension, Oregon Assessment of Knowledge and Skills [OAKS] reading, California state English language arts [ELA], and Star) are significant ($p < .01$), ranging from .549 to .679. These results provide evidence for the convergent validity of MOCCA. Second, for any pair of reading/math tests within a grade, the correlation of MOCCA with the reading test is higher than the correlation with the corresponding math test, although differences can be small. For instance, for the CBM comprehension reading test and the NCTM math test in Grade 3, the MOCCA correlation with CCSS Easy CBM comprehension is .549 whereas the correlation of MOCCA with National Council of Teachers of Mathematics (NCTM) math is .462. The mean difference between MOCCA's correlation with reading and math tests was .107. After applying Fishers r-to-z transformation, we tested the null hypothesis of equal MOCCA correlations with math and reading tests using a paired *t*, $t(6) = 5.819$, $p = .001$. The consistent signs of these verbal/math correlation differences provide evidence across seven samples for the MOCCA discriminant validity.

Discussion

The purpose of this pilot study was to investigate the reliability and validity of a diagnostic measure of reading comprehension in pursuit of providing the field with instructionally relevant information regarding types of poor comprehenders with a test that is feasible for classroom administration. The internal consistency for each response type was assessed. In addition, IRT models were explored. Comprehension rate was examined as a further source of information about comprehension. Finally, data were reported on convergent and discriminant validity of the overall comprehension (number correct) scores. In our analyses, we have taken steps to prevent differences in form means from distorting the results. For instance, in the ANOVA results, we adjusted for form mean differences by including form within grade as a blocking factor. Other analyses (e.g., reliability estimate) were run separately by form.

MOCCA is based on just three types of errors. We are not suggesting that these are the only three types. In practice, a test of limited length could provide reliable scores for only a limited number of error types. Given this limitation, we have chosen to focus on types of errors that have received support from the cognitive research on think alouds and that can potentially occur in most inferential comprehension tasks. Realistically, the applied reliability issue is not so much whether one should focus on a small number of error types, but rather how to select the most important ones. Prior concept inventories (aka distractor-driven assessments; e.g., Delmas et al., 2007; Hermann-Abell & DeBoear, 2011; Hestenes et al., 1992; Sadler, 1998) have covered a much larger number of error types with the result that they cannot yield reliable scores for all of them and may not yield reliable scores for any of them. Rather than providing only a reliable overall score or only a reliable overall score plus error scores of unknown, but probably low, reliability, MOCCA provides a reliable overall score and reliable scores regarding a limited number of error types. In the next edition, we plan to revise the assessment further, thinking of ways we can most efficiently identify comprehension errors.

An assumption of MOCCA is that some students have stable error preferences across items. However, error choices vary within students as a function of text and task characteristics. Such within-person variation may preclude the possibility that a student chooses the same type of incorrect response every time they make a mistake, but it does not preclude a probabilistic consistency in which the probability of choosing one type of response is more likely than others. All test responses, including correct responses, vary within individuals across items as a function of task and text features. Just as this does not preclude reliable differences in correct responses, it need not preclude reliable differences in error responses.

Internal Consistency

The reliability of each MOCCA response type score was calculated as a way to fill the gap in measurement research by assessing the reliability of both correct and incorrect responses of a reading comprehension assessment. Across forms, internal consistency for the correct causal coherent inference responses was excellent, but this finding is not at all surprising or innovative in measurement research. The more interesting finding is that incorrect response tendencies were measured reliably. Across forms, the reliabilities for the incorrect paraphrase response were all above .70, suggesting that this incorrect response type yields internally consistent scores. However, not all reliability coefficients were above .70 for both the lateral connect and local bridging inference responses.

Comprehension Rate

Even though the readability levels of passages increased from Grades 3 to 5, the causally correct inference mean score increased and the mean comprehension rate decreased. That is, across the grades, students became more accurate and more efficient as one would expect if comprehension is becoming both more accurate and more automatic as students move up the grades. When students apply reading to learn new content, automaticity in comprehension processes allows them to devote most of their attention to content, rather than reading process (Chall & Jacobs, 2003). Given that these data are cross-sectional, it is difficult to explain the comprehension rate grade effects causally in that they can be attributed to maturation, practice, education, or simply a cohort effect. However, these results suggest that the assessment may be sensitive to two key aspects of automaticity in comprehension: accuracy and rate.

The current findings are consistent with what others exploring rate of reading comprehension have found (Hale et al., 2011; Neddenriep et al., 2007; Skinner et al., 2002). That is, a developmental trend of greater efficiency at higher grade levels has been observed over multiple reading comprehension measures where rate was examined. The consistency across studies and measures, along with the higher correlations for comprehension rate with criterion reading comprehension measures found in other research (Hale et al., 2011; McCane et al., 2014; Neddenriep et al., 2007; Skinner et al., 2002) suggests that comprehension rate might yield important information about efficiency in the reading comprehension process.

Implications for Future Research and Test Revision

Within a grade, the forms varied in their causally correct inference mean scores, indicating that forms vary in difficulty within grade. By rearranging items across forms so that some easier items are moved to currently harder forms and some harder items are moved to easier forms, the planned revisions will make the forms more nearly equal in difficulty. Last, statistical equating will be used to adjust for any remaining differences in difficulty. In future research with larger sample sizes for parameter estimation, we will be able to use item response theory–based equating to adjust for whatever form differences remain after reassigning items to forms.

With a combination of item and response type revisions, we hope to bring all internal consistency reliability coefficients to .70 or higher for all response type categories for our first national field test. To accomplish this, we first plan to clarify the definitions of response types so that the different types are more clearly distinguishable with the result that some students will choose (or not choose) a type of response more consistently across items because the types are more clearly recognizable. Second, we plan to drop the local bridging response so that there will be three response types per item. It has proven difficult to write local bridging responses that are clearly distinguishable from both the paraphrase and lateral connect types. In addition, the local bridging inference response type did not provide any diagnostic information between comprehender types in think alouds. Whereas, we originally favored four responses to reduce guessing effects, Rodriguez's (2005) meta-analysis concludes that over 80 years of research has consistently supported the use of three options. Finally, we propose to eliminate the "SKIP" button for items, forcing students to choose an option before proceeding to the next item, so that every item a student reaches but does not answer correctly will provide information about the type of error favored by the student.

Implications for Instruction

Our reporting of comprehension rate raises the question of whether students should be encouraged to read faster. We do not think it wise to encourage speed reading as such, but rather education should be designed to develop reading automaticity which would result in an appropriately fast speed. By an appropriately fast speed, we mean an efficient rate of *accurate* comprehension. Students should first be encouraged to develop accurate comprehension, and in this phase of learning, some students may need to slow down to improve their comprehension. However, after a sufficient level of accuracy has been achieved, the goal would be automaticity as evidenced by efficient accuracy. The automaticity literature emphasizes guided practice as the primary means of achieving automaticity once comprehension has been attained (LaBerge & Samuels, 1974; Logan, 1997; Samuels, Ediger, Willcutt, & Palumbo, 2008; Samuels & Flor, 1997). Even at the stage of developing automaticity, it needs to be recognized that faster is not always better because, for every reading task, there is a minimum time below which even the best readers cannot go without sacrificing comprehension.

One challenge for the future is to develop interpretive materials to assist teachers in understanding the test results so they can use the test to identify poor comprehenders for additional instruction and differentiate types of poor comprehenders for the purposes of differentiating the additional instruction, possibly along the lines suggested by McMaster et al. (2012). We plan to have an analog to a "passing" score on the number correct score or the IRT dimension that corresponds to it. For those who do not reach the passing score, diagnostic information will be provided. Of those who do not reach the passing score, some may be classified as slow, meaning that the rate at which they produce correct responses is not sufficient to reach the passing score in a fairly typical 45-min testing session, the recommended testing time. Those who do not reach the passing score will also receive feedback about their predominant type of incorrect response (paraphrase or lateral connection) if there is one. Students who either do not make many errors or do not demonstrate a consistent preference for paraphrase or lateral connection responses (i.e., show relatively even proportions of choosing each response type) will be reported as *indeterminate* in their reading comprehension processing. The incremental validity and diagnostic utility of the incorrect response scores, over and above the causally correct score, will be of particular interest in future research. At this early stage in the research, however, we do not know how much incorrect scores can improve instructional decisions or add to our understanding of how students perform in class over and above the information provided by a single number correct score.

The utility of comprehension rate information for instruction is ambiguous. Given the side effects of a strong focus on oral reading fluency (Samuels, 2007), additional work is needed to determine whether an instructional focus on rate is of practical utility. It is an open question as to whether comprehension fluency can be improved simply with guided reading practice or whether something more is required.

Using think-aloud responses to classify poor comprehenders, McMaster et al. (2012) found that in whole class instruction paraphrasers benefited more from a general questioning strategy whereas lateral connectors (labeled "elaborators" in their study) benefited more from a causal questioning strategy. Results suggest that MOCCA may be useful in differentiating instruction. In more recent research, however, McMaster and colleagues failed to replicate these results in small group, individualized instruction (McMaster, Espin, & van den Broek, 2014). These findings lead to further questions as to whether or not differentiating questioning strategies for different types of struggling comprehenders is equally effective in groups of different sizes.

Conclusion

We presented evidence supporting the convergent and discriminant validity of MOCCA, results that begin to address the construct validity of MOCCA. Results provide cross-sectional evidence for grade trends that may indicate sensitivity to instruction, but instructional validity has yet to be investigated. Results also show that reliable incorrect response scores are possible, but not assured. As with most inventories with diagnostic subscores, the utility of the error propensity and comprehension rate in differentiating instruction needs further research.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140185 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403–425. doi:10.1007/s11336-013-9350-4

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*, 135–160. doi:10.1007/s11336-014-9434-9

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*, 489–503.

Cain, K., & Oakhill, J. V. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*, 683–696.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40–53.

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 23–36). Mahwah, NJ: Lawrence Erlbaum.

Chall, J. S., & Jacobs, V. A. (2003). Poor children's fourth-grade slump. *American Educator*, *27*, 14–17.

Cianco, D., Thompson, K., Schall, M., Skinner, C., & Foorman, B. (2015). Accurate reading comprehension rate as an indicator of broad reading in students in first, second, and third grades. *Journal of School Psychology*, *53*, 393–407. doi:10.1016/j.jsp.2015.07.003

Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*, 28–58.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232. doi:10.1177/001440298505200303

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Graesser, A. C., & Clark, L. F. (1985). The generation of knowledge-based inferences during narrative comprehension. *Advances in Psychology*, *29*, 53–94.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395. doi:10.1037/0033-295x.101.3.371

Hale, A. D., Henning, J. B., Hawkins, R. O., Sheeley, W., Shoemaker, L., Reynolds, J. R., & Moch, C. (2011). Reading assessment methods for middle-school students: An investigation of reading comprehension rate and Maze accurate response rate. *Psychology in the Schools*, *48*, 28–36. doi:10.1002/pits.20544

Hermann-Abell, C. F., & DeBoear, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*, 184–192. doi:10.1039/C1RP90023D

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, *30*, 141–158. doi:10.1119/1.2343497

Hulme, C., & Snowling, M. J. (2011). Children's reading comprehension difficulties nature, causes, and treatments. *Current Directions in Psychological Science*, *20*, 139–142.

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*, 92–114. doi:10.3102/1076998609340529

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394. doi:10.1037/0033-295X.85.5.363

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention*, *29*, 59–70. doi:10.1177/073724770402900408

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323. doi:10.1016/0010-0285(74)90015-2

Logan, G. D. (1997). Automaticity of reading: Perspectives from the instance theory of automatization. *Reading and Writing Quarterly*, *13*, 123–146.

McCane-Bowling, S. J., Strait, A. D., Guess, P. E., Wiedo, J. R., & Muncie, E. (2014). The utility of maze accurate response rate in assessing reading comprehension in upper elementary and middle school students. *Psychology in the Schools*, *51*, 789–800.

McMaster, K. L., Espin, C. A., & van den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice*, *29*, 17–24.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., . . .Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*, 100–111.

Neddenriep, C. E., Hale, A. D., Skinner, C. H., Hawkins, R. O., & Winn, B. D. (2007). A preliminary investigation of the concurrent validity of reading comprehension rate: A direct, dynamic measure of reading comprehension. *Psychology in the Schools*, *44*, 373–388. doi:10.1002/pits.20228

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Lawrence Erlbaum.

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

Perfetti, C. (2010). Decoding, vocabulary, and comprehension. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 291–302). New York, NY: Guilford Press.

Perfetti, C. A., & Lesgold, A. M. (1979). Coding and comprehension in skilled reading and implications for reading instruction. *Theory and Practice of Early Reading*, *1*, 57–84.

Pimperton, H., & Nation, K. (2010). Suppressing irrelevant information from working memory: Evidence for domain-specific deficits in poor comprehenders. *Journal of Memory and Language*, *62*, 380–391. doi:10.1016/j.jml.2010.02.005

Posner, M. I., & Snyder, C. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Lawrence Erlbaum.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, *11*, 289–312.

Rodriguez, M. C. (2005, Summer). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*, 3–13.

Sadler, P. M. (1998). Psychometric models of student misconceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 165–396.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report No. 79-4). Knoxville: Department of Psychology, University of Tennessee.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly*, *42*, 563–566.

Samuels, S. J., Ediger, K.-A. M., Willcutt, J. R., & Palumbo, T. J. (2008). Role of automaticity in metacognition and literacy instruction. In S. E. Israel, C. C. Block, K. L. Bauserman & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 41–59). New York, NY: Taylor & Francis.

Samuels, S. J., & Flor, R. F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly*, *13*, 107–121.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House*, *72*, 97–105.

Skinner, C. H., Neddenriep, C. E., Bradley-Klug, K. L., & Ziemann, J. M. (2002). Advances in curriculum-based measurement: Alternative rate measures for assessing reading skills in pre-and advanced readers. *The Behavior Analyst Today*, *3*, 270–281.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C. E., & Hawkins, R. O. (2009). The validity of reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools*, *46*, 1036–1047.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension. *Topics in Language Disorders*, *25*, 33–50.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). New York, NY: Pearson.

Thurlow, R., & van den Broek, P. (1997). Automaticity and inference generation during reading comprehension. *Reading & Writing Quarterly*, *13*, 165–181.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*, 612–630.

van den Broek, P. W. (1990). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 423–446). Hillsdale, NJ: Lawrence Erlbaum.

van den Broek, P. W. (1997). Discovering the cement of the universe: The development of event comprehension from childhood to adulthood. In P. W. van den Broek, P. Bauer & T. Bourg (Eds.), *Developmental spans in event comprehension and representation: Bridging fictional and actual events* (pp. 321–342). Hillsdale, NJ: Lawrence Erlbaum.

van den Broek, P. W., McMaster, K., Rapp, D. N., Kendeou, P., Espin, C., & Deno, S. (2006, June). *Connecting cognitive science and educational practice to improve reading comprehension*. Paper presented at the Institute of Education Sciences Research Conference, Washington, DC.

Williams, J. P. (2006). Stories, studies, and suggestions about reading. *Scientific Studies of Reading*, *10*, 121–142. doi:10.1207/s1532799xssr1002_1

Wixson, K. K., Valencia, S. W., & Lipson, M. Y. (1994). Issues in literacy assessment: Facing the realities of internal and external assessment. *Journal of Reading Behavior*, *26*, 315–337.