

**Technical Manual 2018: Multiple-Choice Online Comprehension Assessment
(MOCCA)**

MOCCA Technical Report MTR-2018-1

Mark L. Davison

University of Minnesota

Gina Biancarosa

University of Oregon

Ben Seipel

California State University at Chico

Sarah E. Carlson

Georgia State University

Bowen Liu

University of Minnesota

and Patrick C. Kennedy

University of Oregon

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.

MOCCA Technical Manual

Contents

Acknowledgements.....	5
MOCCA Theoretical and Measurement Foundations	6
Theoretical Foundation	6
Types of poor comprehenders.....	6
Think alouds and MOCCA.....	7
The Nature of MOCCA	7
Intended Uses.....	9
General Use.....	9
Specific Intended Uses.....	9
Inappropriate Uses of MOCCA and MOCCA Data	9
Administration	11
Administration Qualifications.....	11
Administration System Requirements	11
Development Process.....	16
The Evolution of the MOCCA Item	16
MOCCA Regional Pilot: Spring 2015	18
MOCCA National Field Test: Spring 2016	19
MOCCA National Calibration Study: Spring and Fall 2017	20
Scoring and Interpretation.....	22
Error Propensity Interpretation and Recommendations.....	23
Comprehension Efficiency Interpretation and Recommendations	24
Scaling and Equating	26
Linking Item Design	26
Error Propensity Dimension Calibration	27
Reliability and Precision.....	28
Classical Test Theory: Reliability of Raw Scores	28
2016 National Field Test Data: Descriptive Statistics	28
Internal Consistency Reliability (Alpha) and Standard Errors.	28
2017-18 National Calibration Sample: Descriptive Statistics	29
2017-18 National Calibration Sample: Internal Consistency (Alpha) and Standard	
Errors.....	29
Item Response Theory	32

Marginal reliability	32
Test-retest Reliability over Time	32
Alternate Forms Reliability over Time	32
Validity	35
Construct Validity Study 1.....	35
Predictive Validity Study	36
Construct Validity Study 2.....	39
Smarter Balanced Assessment Consortium English Language and Arts and Mathematics (SBAC ELA and Math	39
TNReady	40
Convergent Validity.....	40
Discriminant Validity.....	41
Fairness	42
Sensitivity Review	42
Average Score Differences	43
Differential Item Functioning (DIF)	44
Norming	46
Sampling	46
References.....	49
Appendix A:.....	51
Teacher Panel Review Details	51
Appendix B:.....	Error! Bookmark not defined.
Norm Table for Comprehension Dimension Scale Scores by Grade	Error! Bookmark not defined.

Acknowledgements

The MOCCA Development Team has many people and organizations to thank and acknowledge for creating, supporting, testing, and implementing MOCCA. The collaborative effort to bring MOCCA to fruition is astounding! The shared theoretical background, technical knowledge, programming expertise, writing skills, and organizational help have made MOCCA the rigorous and useful tool that it is.

Special thanks are extended to our core MOCCA team of undergraduate students, graduate students, and research assistants who aided in item writing, item analysis, recruitment, data collection and other tasks: Joanie Grohman, Sam Hart, Kylie Hiemstra, Patrick Kennedy, Bowen Liu, Lyndsey Park, Sunhi Park, Lina Shanley, and Hyeon-Jin Yoon.

We also have sincere appreciation for our panel of experts who supported our project from the beginning and challenged us to make MOCCA the best it could be: Dr. Paul De Boeck, Dr. Susan Goldman, Dr. Kristen McMaster, and Dr. Paul van den Broek.

We would be remiss if we did not acknowledge the thousands of students and their teachers for participating in our project to help develop and validate MOCCA. We offer special thanks to our teacher panel who helped review our items for appropriateness, content, and readability.

MOCCA would not be possible without the support of our administrative and programming teams. We offer sincere thanks to Susannah Williams, Cindy Sprague, Eugenia Coronado, and Nick Phillips for keeping the MOCCA development on task and organized. We also are indebted to our programming team. Without their skills, MOCCA would not exist in its current form: Jeff Ness, Susan McEvoy, and Scott McCammon. Similarly, we recognize and appreciate the support of Dr. Hank Fien and the Center for Teaching and Learning for hosting MOCCA in their organization.

Finally, we are grateful to our respective institutions and to the Institute of Education Science (IES) for their support of MOCCA: The University of Oregon; California State University, Chico; The University of Minnesota-Twin Cities; and The University of Wisconsin-River Falls. We would also like to acknowledge the guidance and support of our IES program officer: Meredith Larson.

The research reported here was supported by the Institute of Education Science, 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.

MOCCA Theoretical and Measurement Foundations

Designed for students in Grades 3 through 5, the Multiple-choice Online Causal Comprehension Assessment, or MOCCA, identifies students who struggle with comprehension, and helps uncover why they struggle. There are many reasons why students might not comprehend what they read. They may struggle with decoding, or reading words accurately and fluently. They might have limited vocabulary and background knowledge. But there are some students who don't comprehend well and don't fall into these categories.

Researchers have dubbed these readers “poor comprehenders” and have found that they are not making inferences that help them maintain a coherent idea of what a text is about. These poor comprehenders are usually trying to make sense of what they read, but they are relying on strategies that don't fully do the trick. It turns out, they tend to rely on one of two strategies: paraphrasing or making elaborations, which include elaborative inferences, personal associations, and self-explanations. These are great strategies, but neither alone will result in excellent comprehension. What's more, research suggests that students who rely on paraphrasing versus making lateral connections require somewhat different instruction.

Theoretical Foundation

Proficient readers must attend to text characteristics (e.g., letters, sounds, words) as well as their understanding of the content by drawing on explicit text information and background knowledge. Consequently, skill in reading words, although necessary for comprehension, is not sufficient on its own to guarantee comprehension. To comprehend successfully, readers use comprehension processes to build a coherent mental representation of a text. Mental text representations are idiosyncratic, cognitive structures people create to understand the situation. This coherent mental representation is called a situation model. A *situation model* includes events from a text along dimensions of time, space, characters, character goals, and causality.

Research has yielded evidence for poor comprehension among intermediate grade and older readers where no word reading difficulties exist. A preponderance of the evidence for poor comprehenders comes from research with intermediate grades. Although frequency of poor comprehension varies across studies, the occurrence of poor comprehension does not vary.

Types of poor comprehenders. Research has demonstrated that poor comprehenders are not a monolithic group. Typically, they are characterized as failing to engage in necessary comprehension processes. However, the alternative processes in which they engage instead can also distinguish them. When word reading, other component skills, and knowledge are ruled out as causes, research has shown that poor comprehenders make fewer *necessary inferences* than proficient comprehenders do.

During reading, proficient comprehenders engage in a host of comprehension processes, but only some are truly necessary to comprehension. One class of these processes is the *causally coherent inference*. These inferences rely on causal information in the text, and they are necessary for maintaining coherence. To make causally coherent inferences a reader synthesizes events and character goals in a text with relevant background knowledge that is not explicitly

stated in the text. For example, consider this brief text from Thurlow and van den Broek (1997): “Toby wanted to get Chris a present for his birthday. He went to his piggy bank.” Good comprehenders seem to effortlessly infer that Toby goes to his piggy bank to get money to buy Chris a present. Importantly, unless one makes this inference, Toby’s trip to his piggy bank is entirely unmotivated, an apparent non sequitur.

Note that poor comprehenders do make causally coherent inferences, but they tend to make fewer than proficient comprehenders do, *and* they instead tend to rely on one of two reading comprehension processes that are good practices but are neither necessary nor sufficient for maintaining *causal* coherence. The first of these processes is *paraphrasing*. Paraphrases restate or rephrase prior text, which can support coherence, but are not strictly necessary for maintaining causal coherence. Moreover, they do not strictly rely on background knowledge. The other process poor comprehenders tend to overuse is elaboration. Although we prefer the term lateral connection, since this category includes more than just elaboration (e.g. self-explanations, evaluations, and associations), we will use the term “elaboration” because that seems to be the more common term in the literature and because elaboration is the most common response type in this category. Elaborations access background knowledge but are not necessarily *causally coherent* connections.

Research has shown that both good and poor comprehenders can and do use *many* other comprehension processes; however, poor comprehenders can be distinguished by these two processes—paraphrases or elaboration—they rely on when they do not make a causally coherent inference. In other words, what distinguishes poor comprehenders from good comprehenders, holding their word reading and vocabulary constant, is their less consistent and strategic use of causally coherent inferences. And what further distinguishes poor comprehenders *from each other is the comprehension process they tend to overuse instead*: paraphrases or elaboration.

Think alouds and MOCCA. During a think-aloud task a reader reads aloud a unit of text (e.g., a sentence) and verbalizes that which she is thinking about while reading. Think alouds are successful at measuring online comprehension. Think alouds have identified specific comprehension processes (e.g., inferences; paraphrases; associations; metacognitive responses) that take place during reading. In fact, think alouds are the prime source of evidence that readers indeed use different types of comprehension processes. Think alouds are also the prime source of evidence that poor comprehenders can be distinguished diagnostically by the processes they overuse—paraphrases or elaborations. Although think alouds are an online, rich, and reliable source of information on comprehension processes, they are also impractical for schools because of the data collection, coding, and analysis burden they pose. The benefit of think alouds, combined with their limitations, leads to the development of MOCCA as a new assessment tool to identify differences in struggling comprehenders.

The Nature of MOCCA

Each MOCCA item consists of a short narrative text and four corresponding multiple-choice responses to complete a missing sentence (cf. Carlson, Seipel & McMaster, 2014; Davison, Biancarosa, Carlson, Seipel, & Liu, 2018). 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). Each narrative text constitutes a cloze item but, instead of deleting every *n*th 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 2-4 with a mean of 3.0 on the Flesch-Kincaid scale.

For each narrative test item, the student must choose one of three alternative responses 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), the two remaining responses are informative distractors: a *paraphrase* and an *elaboration*.

Intended Uses

General Use

MOCCA is designed to identify and diagnose Grade 3-5 students who struggle with reading comprehension. MOCCA is appropriate for students in Grades 3-5. Use beyond these grades has not been validated and is not supported.

More specifically, MOCCA is designed to identify the cognitive processes that struggling comprehenders overuse while reading (i.e., paraphrasing, repeating text, connections, making irrelevant elaborations or associations).

The information gained from administration can be useful for: measuring general reading comprehension ability (i.e., good or poor comprehender), identifying type of struggling comprehender (i.e., paraphraser, elaborator), determining comprehension efficiency (i.e., fast or slow), informing instruction, bench-marking progress, and making Tier 1 and 2 response-to-intervention (RTI) decisions.

Specific Intended Uses

Each grade level of MOCCA has three validated forms which can be used for progress monitoring or benchmarking. Only benchmarking use has been validated at this time. Nonetheless, multiple forms are available in Grades 3-5 to accommodate those wishing to use MOCCA to monitor progress. Given the slowness with which reading comprehension changes, it is recommended that MOCCA be administered no more than three times per academic year.

MOCCA has concurrent and construct validity evidence to indicate that it provides information similar to other more-traditional reading comprehension assessments (See Validity section below). Therefore, it can be appropriately used as a Tier 1 screening or benchmarking measure for all students.

Additionally, MOCCA has also been validated for use as a cognitive diagnostic tool about individual students who struggle with reading comprehension. That is, data from MOCCA not only identifies those at risk for poor reading comprehension, but also provides instructionally relevant diagnostic information about why a student struggles with reading comprehension. Specifically, it identifies the cognitive reading comprehension processes that a student who struggles with comprehension overuses.

Inappropriate Uses of MOCCA and MOCCA Data

As with any benchmarking or diagnostic measure, MOCCA is best used in combination with other assessments when a complete picture of a child's reading abilities is desired. MOCCA does not provide diagnostic information about decoding or other "low-level" component reading skills.

Although MOCCA provides a *comprehension efficiency* score, this score is not a measure of oral reading fluency. If a teacher or other professional suspects that a student has oral fluency issues, then the student should be evaluated with a more appropriate assessment that is pertinent to oral reading fluency and its component skills (e.g., decoding, phonemic awareness).

Administration

Administration Qualifications

MOCCA is considered a *Level A* assessment. This means that there are minimal special qualifications for administration and interpretation of scores. It is recommended that the assessment be administered and interpreted by personnel who have an understanding of MOCCA and of reading comprehension. Specifically, the assessment should be administered by a teacher, paraprofessional, administrator, school psychologists, or other school personal who can maintain data privacy and test security. Scores should only be interpreted by teachers, school psychologists, or administrators who can maintain data privacy and test security.

MOCCA is only validated for computerized administration. Although MOCCA was originally developed in a paper-and-pencil format, no paper-and-pencil versions are available at this time.

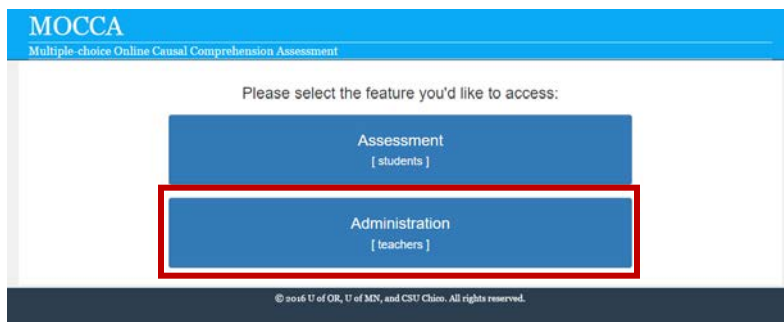
Administration System Requirements

The MOCCA system requires internet connectivity and a modern web browser, such as Chrome, Edge, FireFox, or Safari. Access to the mocca.uoregon.edu website must be allowed over the school network. Headphones are optional, though recommended for students to receive clear introductory instructions. You may also play the instructions over a speaker for all to hear, but must then monitor that individual students are keeping up by clicking Next appropriately.

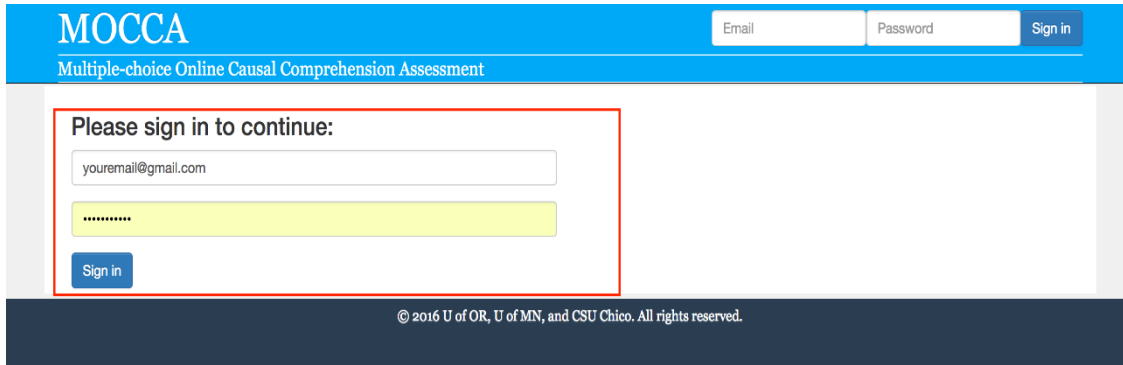
Administration Instructions

Here is the information you will need in order to prepare for MOCCA testing.

1. Click the Administration link at <http://mocca.uoregon.edu>.

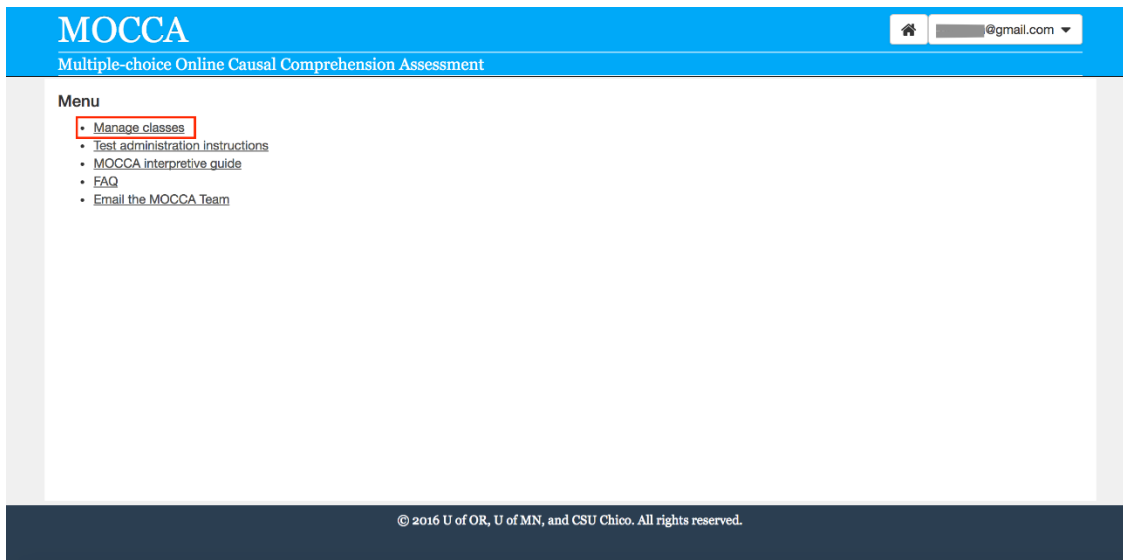


2. Log in using the email address provided to the MOCCA Project (usually your school email) and your password.



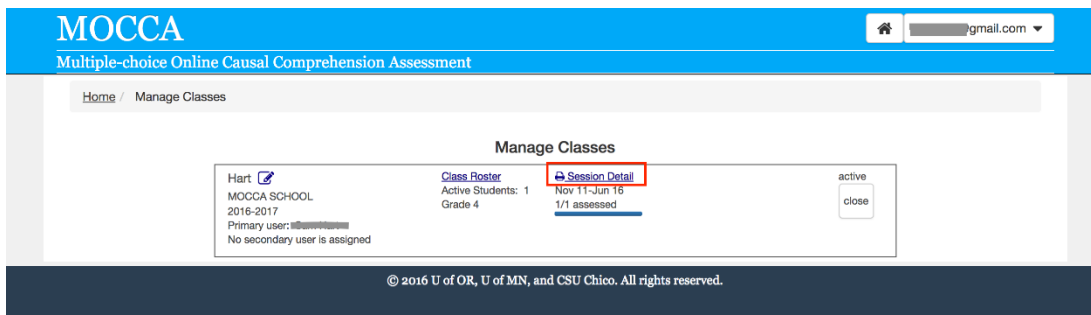
The screenshot shows the MOCCA login page. At the top, there is a blue header with the MOCCA logo and the text "Multiple-choice Online Causal Comprehension Assessment". To the right of the header are input fields for "Email" and "Password", and a "Sign in" button. Below the header, a red box highlights a sign-in form with the heading "Please sign in to continue:". The form contains an email input field with "youremail@gmail.com", a password input field with masked characters, and a "Sign in" button. At the bottom of the page, there is a dark blue footer with the copyright notice: "© 2016 U of OR, U of MN, and CSU Chico. All rights reserved."

3. Once you've signed in, you will see your Home page. Click on "Manage classes."





The screenshot shows the MOCCA Home page. The header is the same as in the previous screenshot. Below the header, there is a "Menu" section with a list of links: "Manage classes", "Test administration instructions", "MOCCA interpretive guide", "FAQ", and "Email the MOCCA Team". The "Manage classes" link is highlighted with a red box. At the bottom of the page, there is a dark blue footer with the copyright notice: "© 2016 U of OR, U of MN, and CSU Chico. All rights reserved."

4. Click on "Session Detail."

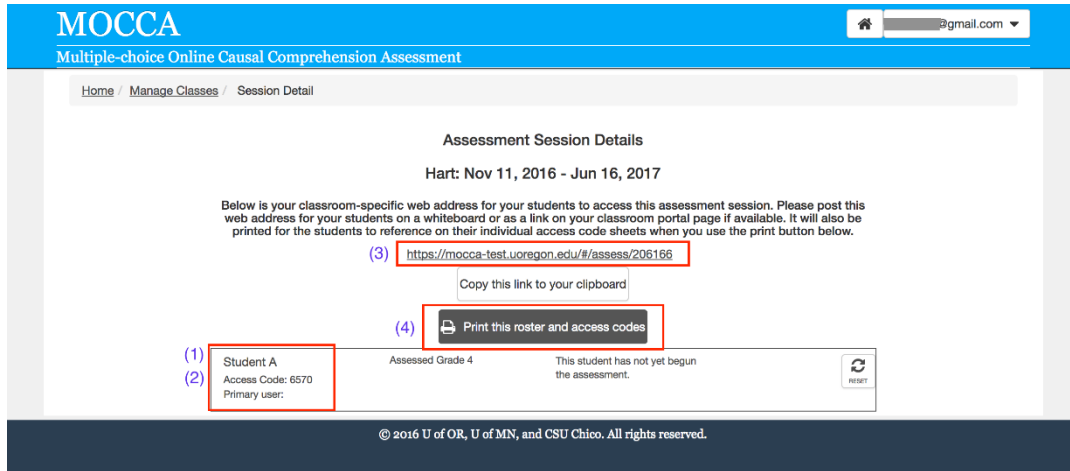


The screenshot shows the MOCCA "Manage Classes" page. The header is the same as in the previous screenshots. Below the header, there is a breadcrumb trail: "Home / Manage Classes". The main content area is titled "Manage Classes" and contains a table with the following information:

Hart 	Class Roster	Session Detail	active
MOCCA SCHOOL	Active Students: 1	Nov 11-Jun 16	close
2016-2017	Grade 4	1/1 assessed	
Primary user: 			
No secondary user is assigned			

At the bottom of the page, there is a dark blue footer with the copyright notice: "© 2016 U of OR, U of MN, and CSU Chico. All rights reserved."

5. The "Session Detail" page provides: (1) the list of your students; (2) each student's unique access code; (3) the classroom specific URL; and (4) an option to print out your class roster.



6. On the day of testing, students will type your classroom specific URL into their web browsers.
7. Students will enter their unique access codes and pick their names under the drop-down menu.



8. Before starting the assessment, students will read or listen to the assent form and instructions.

For more information, please check the Frequently Asked Questions below.

If you have any problems verifying your email or getting into the MOCCA system at any point, please email our team at mocca@uoregon.edu.

Frequently Asked Questions

1. [What are some easy ways that I can show my students the testing URL?](#)
2. [Is there a way to demonstrate the sign-in process before my students begin?](#)
3. [Do my students need headsets?](#)
4. [Can my students change the text size?](#)
5. [How many test items do my students need to complete?](#)
6. [How can I show my students their access codes?](#)
7. [How can I track my students' progress through a MOCCA form?](#)

8. [How can I add a student to my class roster?](#)
9. [When will my students' results be available, and how can I access them?](#)

1. What are some easy ways that I can show my students the testing URL?

You can write your classroom's testing URL on the board or overhead projector. Alternatively, if you have a class webpage, such as Google Classroom, you can post the link there for your students to access.

2. Is there a way to demonstrate the sign-in process before my students begin?

You can sign in to your classroom's testing URL on a classroom projector. Use session code 000000 and access code 0000 to walk through the instructions screens in demo mode.

3. Do my students need headsets?

We suggest providing students with headsets. Students have the option to listen to the assent form and instructions pages. They will not need headsets after the instructions. However, some students may prefer to keep the headphones on for a quieter experience. If headsets are not available, you may want to ask students NOT to listen to the instructions so that the audio instructions will not disturb other students.

4. Can my students change the text size?

Yes. Let students know that there is a text size increaser in the upper right-hand corner of each testing page.

5. How many test items do my students need to complete?

If possible, please ensure that students complete at least 10 test items so that we have enough information to process their data.

6. How can I show my students their access codes?

Each student is assigned a unique access code. After you log in, we encourage you to print your roster from your "Session Detail" page, and then cut out each student's name and access code. You can give these cutouts to your students to use when they log in to the site for testing.

7. How can I track my students' progress through a MOCCA form?

During testing, you, the teacher/administrator, can look at your "Manage Classes" page to see the progress of your class as a whole. You can also look at your "Session Detail" page to see the progress of each student as he or she is testing. The progress bar you see for each student will indicate how many items out of 40 he or she has completed. The progress bar is color-coded. Gray indicates that the student hasn't completed any items yet. Yellow-orange indicates that the student has begun the assessment but has not met the 10-item minimum. A short green indicates that the student has met the 10-item minimum. A full green bar indicates that the student has completed all of the items on the assessment.

8. How can I add a student to my class roster?

- On the “Manage Classes” page, click "Class Roster."
- Click the “Add Student” button.
- Provide the student’s first and last name. Choose the grade that the student is currently enrolled in under “Grade Enrolled.” Next, choose the grade in which the student will be assessed under “Grade Assessed” (Grade Assessed and Grade Enrolled are usually the same).
- Click the “Save” button to save the student.
- You should now see the student you added, as well as his or her access code, on your “Session Detail” page.

9. When will my students’ results be available, and how can I access them?

Go to the “Manage Classes” page. Click the “close” button on the righthand side of the page to close your session. Once the session is closed, a link to the reports will appear on the “Manage Classes” page underneath the “Session Detail” link. There will be a button to print out reports and a button to download reports.

Contact us at mocca@uoregon.edu if you have any other questions. Please be patient and expect a response within a couple of business days.

Development Process

MOCCA began as a paper-and-pencil assessment with a single form consisting of 40 items. A measurement grant from the Institute of Education Sciences (R305A140185) enabled the refinement of MOCCA into its current form, which is a computer-administered assessment with three forms per grade for Grades 3-5.

The Evolution of the MOCCA Item

The original paper-and-pencil MOCCA included four responses: one correct, two informative incorrect, and one uninformative incorrect response. The final form of MOCCA uses three responses: one correct and two informative incorrect responses. The main reasons for including four responses was to reduce the probability of obtaining the correct answer by random guessing. In theory, this should reduce random error associated with guessing thereby increasing the reliability and the IRT information function throughout the range of ability (i.e., θ) but especially at the low end of the scale where guessing is more common.

Given MOCCA's goal of distinguishing between paraphrasers and elaborators, the test provides the most information about paraphrasing and elaborating when, given a wrong response, either a paraphrase or elaboration. The uninformative incorrect response answer provided no information about whether the person is predominantly a paraphraser or predominantly a later connector. Rodriguez's (2005) meta-analysis concludes that over 80 years of research has consistently supported the use of three answer choices over four (see also Costin, 1970; Grier, 1975; Tversky, 1964). Therefore, we decided to remove the uninformative distractors and use just three response alternatives.

The three response types that MOCCA incorporates are:

- **Causally coherent inferences**, which are the correct responses because they provide necessary information to fill the causal gap between the 5th and 7th sentences, completing the story by stating or implying whether the goal of the main character has been met.
- **Paraphrases**, which are incorrect because they paraphrase either the original or updated goal in a story, but add no new information thereby leaving the causal gap unfilled.
- **Elaborations**, which are also incorrect because they build on the 5th sentence of the story by adding extra-textual information through elaboration, association, or explanation, but do not fill the causal gap.

Item writing began in summer of 2014 and continued through early 2015. The item writing process included five phases:

1. Initial item writing
2. Story review and response writing (paraphrase and elaboration)
3. Response review

4. External review of items
5. Final revisions and review

The item writing team consisted of three of the four MOCCA authors of MOCCA (i.e., Biancarosa, Carlson, and Seipel) and other research personnel for a total of seven item authors. The author team kept a list of story ideas to prevent repeating story ideas unintentionally.

Phase 1 consisted of a single author writing a complete seven sentence story. Flesch-Kincaid grade levels were recorded at the completion of each story, and the story idea list was updated to reflect whether an existing or new idea had been used.

In Phase 2, a team of 2 or more authors, not including the original author, reviewed and revised the item/story. Items were reviewed for (a) causal coherence, (b) accuracy, (c) appropriateness, (d) freedom from bias, and (e) engagingness. As part of the process, the sixth sentence was removed and placed below the story as the causally coherent, or correct, response. Two additional responses were written: one paraphrase and one elaboration. The story review and response writing occurred in a single phase because response writing inevitably led to considerations of whether the story would adequately support the necessary response types. Thus, it was more efficient to compose the response types as part of the story vetting process.

In Phase 3, a different team of 2 or more authors reviewed the item as a whole. This round of review served primarily as a check on the response types but could involve additional revisions to the story as well.

In Phase 4, a panel of six local intermediate grade teachers reviewed the items as intact stories (with the sixth sentence replaced). Teachers reviewed the items for (a) causal coherence, (b) accuracy, (c) appropriateness, (d) freedom from bias, and (e) engagingness. Teachers used a formal scale to rate items on these dimensions. The panel was trained in a single 4-hour session with ample discussion of the purpose of MOCCA, the intent behind the dimensions they were rating, the meaning of the rating scale they used, and training in using the Qualtrics interface for rating. Two items were reviewed as a group with discussion, and then two rounds of rating several items followed. Interrater reliability was not a goal in this training because judgments of the dimensions were necessarily subjective. Instead, the goal was for consistent and appropriate use of the rating scales and Qualtrics system. Teacher reviews were completed in three waves of about 160 items each. Each teacher was randomly assigned to review half of the items in each wave. Thus, each and every MOCCA item was reviewed by three members of the teacher panel.

In Phase 5, the author team reviewed teacher ratings and written feedback in depth. In many cases, revisions were made to stories based on the feedback. Because teacher reviews began before all items had been written, this feedback also helped fine-tune the writing and revision process for the remaining stories. As a result, while teachers flagged over half of the first batch of stories as having marginal or serious problems, each subsequent batch had proportionally fewer stories flagged. For example, teachers consistently flagged stories with explicit references to magic or witches or similar concepts as potentially offensive; some of these stories were revised, while others were retired. This experience led us to fix such occurrences prior to sending stories out to the teacher panel in the second and third review batches.

However, not every flagged story was revised based on teacher feedback. One recurring issue was that teachers often flagged stories as *not needing the sixth sentence* (i.e., being causally coherent without that sentence). Although sometimes they were correct, in most cases it appeared teachers were flagging items for this dimension due to one of two reasons. In most cases, they said the sixth sentence was not necessary, but review revealed that the sentence included an event that *must* occur for the seventh sentence to make sense (i.e., it *was* necessary). However, the event was very easy to infer. The inference may have been so automatic for the teachers that they were unaware of making it. The other times they flagged stories for not needing the sixth sentence was in situations where they explained that they felt any number of different events could have led to the seventh sentence. These stories were not revised because, once again, clearly *something* needed to occur for the seventh sentence to make sense.

In all, over 500 items were written, including revisions of the 40 original paper-and-pencil MOCCA stories. Of these, 480 items made it to the pilot phase with 40 appearing on each of 12 forms (i.e., 4 forms per 3 grades). A handful of items were retired based on feedback from the teacher panel. Others were determined to be too easy or hard in terms of readability (less than second grade level; over sixth grade level). Final exclusions were based on efforts to create forms that were as nearly equal as possible in terms of average Flesch-Kincaid grade levels, and counts of items featuring (a) male, female, and indeterminate gender characters, (b) endings considered to be happy, sad, and neutral emotionally, and (c) whether the emotion at the end was explicitly stated or needed to be inferred. No character names repeated across any items on any form in any grade.

MOCCA Regional Pilot: Spring 2015

The MOCCA pilot occurred in California and Oregon between March and June 2015. A total of 929 students took the assessment. MOCCA personnel oversaw administration of the assessment during the pilot. Four raw scores were computed for each student. The first was the traditional number correct (i.e., the number of items for which the student selected the Causal Coherent response). The remaining three scores were (a) the Paraphrase score, the number of times the student picked the Paraphrase response, (b) the Elaboration score (then called the Lateral Connect score), the number of times the student selected an Elaboration response, and (c) the True Distractor score, the number of times the student selected the uninformative incorrect response.

Results demonstrated that MOCCA was more difficult for third grade students than for fourth and fifth grade students. The average number correct in Grade 3 was barely more than 20 items for three of four forms, but was about 28 items in the higher grades.

Overall, no story/item features, such as various readability formulas and features on which the forms were balanced, predicted item difficulty more than marginally when examined by grade level. More specifically, difficulty was not a function of readability or vocabulary knowledge, suggesting MOCCA successfully isolates difficulty that readers have in making causally coherent inferences rather than tapping lower level component skills.

The internal consistency reliability (i.e., Cronbach's alpha) for the number correct score was excellent across all forms (above .90). The Paraphrase scores had generally high reliabilities, ranging from .71 to .89. The reliabilities for the Elaboration score ranged from .49 - .81.

These results informed revisions for the field test version of MOCCA. The poor reliability of the Elaboration score also informed dropping the true distractor from MOCCA items altogether. Several item statistics were used to inform the item revision process. For the correct response, these included the item difficulty (proportion correct), the item-total correlation, and the two-parameter logistic (2PL) item response theory parameters (difficulty and discrimination parameters.) For the informative, incorrect response alternatives (Paraphrase and Elaboration), the statistics included the "difficulty" (proportion selecting the response) and the item-total correlation, where the total score was the number of items for which the student had selected that incorrect alternative. Items with excellent statistics were examined and compared to those with particularly poor statistics in order to refine item specifications for the informative incorrect responses. Both the paraphrase and elaboration definitions were tightened to offer more concrete and strict guidance regarding good and poor responses of these types. Then, items with poor item-total correlations for any of the three informative scores (i.e., correct, paraphrase, elaboration) were targeted for revision. Items were dropped during revisions if the item was deemed too difficult or complicated to revise. Some items that were deemed inordinately easy in Grades 4 and 5 were also targeted for revision using the refined item specification rules. Finally, items where all item-total correlations, including for the correct response type, were poor were dropped outright.

The final stage of the pilot was to create three new forms per grade level. Forms were again balanced for readability and other features as for the pilot. Some administration features were also refined. Most importantly, the opportunity to skip an item and to review items was removed.

MOCCA National Field Test: Spring 2016

The MOCCA field test occurred nationwide. Participating classrooms were recruited from districts participating in the pilot and from the DIBELS Data System. As a result, 3,721 Grade 3-5 students from over 50 schools in 13 states took the assessment between February and June 2016. This time teachers or building administrators proctored MOCCA without assistance from MOCCA personnel. The average administration time was 35 minutes with a standard deviation of about 15 minutes across grades. Three raw scores were computed for each student. The first was the traditional number correct (i.e., the number of items for which the student selected the Causal Coherent response). The remaining two scores were for the informative distractors: (a) the Paraphrase score or number of times the student picked the Paraphrase response and (b) the Elaboration score or number of times the student selected an Elaboration response.

Results again demonstrated that MOCCA was more difficult for third grade students than for fourth and fifth grade students. However, results also indicated that revision efforts to make the Grade 3 forms somewhat easier and the higher grade forms more difficult were successful

with the mean number correct falling at about 22 items in Grade 3, 25 items in Grade 4, and 27 items in Grade 5.

As before, item statistics informed revisions for the next version of MOCCA. In addition to the item statistics described earlier, Mantel-Haenszel differential item functioning (DIF) statistics were calculated for each item (Dorans & Holland, 1988). Analyses of DIF by gender and Hispanic vs. White ethnicity/race were conducted. Sample sizes for American Indian/Alaskan Natives, Blacks, and Asian/Pacific Islanders were too small for DIF analyses. Eight items with differential item functioning (DIF) for the correct answer were dropped; four of these were for gender-based DIF and four for ethnicity-based DIF (see Fairness Section for more detail). Items with item-total correlations less than or equal to .25 for each of the three scores were also dropped. Table 1 shows the demographic composition of the 2016 National Field Test sample.

MOCCA National Calibration Study: Spring 2017 and Fall and Winter 2017-18

In Spring 2017 and Fall and Winter 2017-18, we gathered a national sample of data. Since one purpose of this study was to equate across grades and forms, we first selected ten linking items to be included in all three forms within each grade. Linking items were selected as follows using data from the 2016 field test. Linking items had to have item-total correlations for the correct response that were above .5. Each form provided one linking item, and the tenth linking item was drawn from a fourth grade form. Items within a grade level were chosen so that one was relatively easy (compared to other items at that grade), one was average, and one was relatively difficult in terms of the proportion of students getting the item correct. Easy and difficult were defined as the third and first quartiles for proportion correct. For Grade 4, two items of near-average difficulty were chosen. These guidelines were followed closely for Grades 3 and 4, but items from Grade 5 forms were chosen to be slightly easier in all cases so that they would not be unduly challenging for third grade students. Table 1 shows the demographics for this sample along with totaled demographic representation for the norming sample.

To maintain the test length at 40 items, the common items replaced other items in the various forms. That is, on the various forms some items were dropped and replaced with a common item for purposes of this calibration study. Items with relatively weak item statistics in the 2016 field test were dropped. First, we dropped the items identified in the DIF analysis described above. Then we dropped items that did not seem to distinguish well between Paraphrasers and Elaborators. In this way, we prioritized items that discriminated as well as possible between students' tendencies to choose a paraphrase versus an elaboration.

Table 1

Composition of the Norming Sample by Study Participation and Gender
Race/ethnicity, English Language Learners, Free/Reduced Lunch, and Special
Education

Demographic variable	Category	2016 National Field Test Sample	2017-18 National Calibration Sample	Combined Norming Sample
Gender	Male	51%	50%	50%
	Female	49%	50%	50%
Race	White	64%	56%	61%
	Hispanic	23%	25%	24%
	Black	7%	10%	8%
	Asian	3%	3%	3%
	American Indian	1%	2%	2%
	Native Hawaiian	1%	<1%	<1%
	Two or more	2%	4%	3%
	English Language Learners Status	Yes	10%	10%
	No	90%	90%	90%
Free and Reduced Meal Status	Yes	51%	56%	53%
	No	49%	44%	47%
Special Education Status	Yes	11%	10%	10%
	No	89%	90%	90%

Scoring and Interpretation

MOCCA score reports offer a great deal of information not provided by other reading comprehension assessments. Always be sure to coordinate MOCCA results with other data sources for the best possible decisions about intervention needs.

MOCCA score reports offer several classifications and other data for students. Most unique are the error propensity and comprehension efficiency classifications. These classifications, along with other MOCCA scores are defined as follows.

Grade (Assessment Grade): *Grade* indicates the current grade in which the student is enrolled. *Assessment grade* indicates the grade level of the MOCCA form the student took.

Form: Indicates which grade-level form the student took.

Scaled Score: Indicates student's performance on a scale of 50 to 950, where higher scores are better.

Scaled Score Percentile Rank: Indicates the percentage of students a student would be expected to score as well as or better than in a nationally representative sample of US students.

Risk Status: Indicates whether the degree to which a student is predicted to be at risk of missing end of year learning goals for English language arts.

At risk: Indicates a strong likelihood the student will not meet end of year goals without intervention.

Some risk: Indicates a moderate likelihood the student will not meet end of year goals without intervention.

Minimal risk: Indicates a low likelihood the student will not meet end of year goals without intervention.

Error Propensity: Indicates which type of response (Paraphrase or Elaboration) is dominant in the student's responses if there is a dominant type.

Elaboration: Indicates a poor comprehender who tends to choose more elaborations (i.e., make elaborative inferences and predictive inferences).

Paraphrase: Indicates a poor comprehender who tends to choose more paraphrases (i.e., restate the text, sometimes in their own words, without adding new information).

Indeterminate: Indicates a poor comprehender who could not be classified. The student may show no preference or may not have completed enough items.

Not Applicable: Indicates a student who performs very well on MOCCA (is at minimal risk).

Comprehension Rate: Average minutes per correct item, expressed in minutes:seconds, where lower rates indicate faster readers.

Comprehension Rate Percentile: Indicates the percentage of students a student would be expected to perform as fast as or faster than in a nationally representative sample of US students.

Comprehension Efficiency: Indicates classification based on average minutes per correct answer. A student is labeled Fast if their comprehension efficiency score is 1:24 or less. If this rate is sustained, the student can answer 80% (32) of the items correctly in 45 minutes, the recommended administration time for MOCCA.

Fast and accurate: Comprehension rate is well above average with greater accuracy.

Fast and inaccurate: Comprehension rate is well above average with less accuracy.

Moderate and accurate: Comprehension rate is about average with greater accuracy.

Moderate and inaccurate: Comprehension rate is about average with less accuracy.

Slow and accurate: Comprehension rate is well below average with greater accuracy.

Slow and inaccurate: Comprehension rate is well below average with less accuracy.

Error Propensity Interpretation and Recommendations

Always be sure to coordinate MOCCA results with other data sources. Only students who are deemed at risk or at some risk get an error propensity classification.

Students who receive **indeterminate** as their error propensity may have decoding or fluency problems, may be using guessing as a test-taking strategy, or may have some other issue underlying poor performance on comprehension measures. Teachers should consult the students comprehension efficiency classification, as well as additional data sources, to determine the student's needs.

Students identified with **paraphrase** appear to be relying on paraphrasing and otherwise repeating what they read. They are overly dependent on the text alone for making meaning of what they read. *While these are good strategies for reading, these students need to be encouraged to make inferences to provide missing or implicit information as they read.*

Students identified with **elaboration** appear to be relying on making elaborative inferences and predictive inferences. They are making inferences, but these inferences do not make the most meaning of what they read. *While these are all good strategies for reading, these students need to be encouraged to prioritize maintaining the coherence of the message of what they read. Coherence in narratives often depends on causal relations and how one event or character influences another.*

Beginning in 2018, we plan to classify students in the following way with respect to being a paraphraser or an elaborator using an IRT based error propensity score. See the IRT section below. The dimension will be scaled so that the 0 point is a point of indifference at which a student has a conditional probability of .5 of choosing a paraphrase over an elaborative response on an item of average difficulty given that the student does not select the correct answer. This means that if a student has a positive score on the error propensity dimension, they favor paraphrase responses over elaborative responses, at least by some small amount. Conversely, if the score is negative, they tend to favor elaborative responses over paraphrase

responses by at least a small amount. Let $\hat{\theta}$ be the student's estimated location along the error propensity dimension, and let $s(\hat{\theta})$ be the conditional standard error for $\hat{\theta}$. We will compute a z score for each student as follows: $z = \hat{\theta}/s(\hat{\theta})$. If a person has $\hat{\theta} \geq 1$ and a $z \geq 1$, we will classify the person as a paraphraser. That is, if the student's score is one standard deviation and one standard error or more above the indifference point (0), they will be classified as a paraphraser. If a person has $\hat{\theta} \leq -1$ and a $z \leq -1$, we will classify the person as an elaborator. That is, if the student's score is one standard deviation and one standard error or more below the indifference point (0), they will be classified as an elaborator. If $-1 < \hat{\theta} < 1$, the person will be classified as indeterminate.

Comprehension Efficiency Interpretation and Recommendations

Only students who get at least one item correct (and therefore have a comprehension rate) will receive a comprehension efficiency indicator.

Indicators of comprehension efficiency should not be taken to mean that faster is always better. The main goal is for all students to be *accurate* in their comprehension. A *fast* indication is only good insofar as a student is comprehending (i.e., is *fast and accurate*). Note that accuracy here relates not to decoding, but to a student's ability to resolve causal gaps in a narrative by making a causally coherent inference.

Students who are *fast and accurate* are unlikely to need intervention in making causal inferences. Their comprehension is excellent, and their rate is quite brisk.

Students who are *moderate and accurate* are also unlikely to need intervention in making causal inferences. Their comprehension is excellent, and their rate is average.

Students who are *slow and accurate* comprehend well. They may need to work on fluency or to engage in structured practice to improve their pace. They may also need to work on decoding if they perform better on measures of passage reading than word list reading. Other data is necessary to determine their needs. However, for students with IEPs or receiving English language services where additional time is a recommended accommodation, this designation may be reiterating the need for that accommodation.

Students who are *fast and inaccurate* likely need to slow down. They may be students who rushed through the test either without really reading or without really trying to do well. However, they may also be students who when they read are prioritizing speed over accuracy in decoding or prioritizing fluency over meaning. Additional data is necessary to determine their needs.

Students who are *moderate and inaccurate* may also need to slow down. They may be students who rushed through the test either without really reading or without really trying to do well. However, they may also be students who when they read are prioritizing speed over accuracy in decoding or prioritizing fluency over meaning. Additional data is necessary to determine their needs.

Students who are *slow and inaccurate* do not comprehend well and proceed at a slow pace. Depending on their comprehension rate (also found on MOCCA reports), they may just be a bit slow or very slow. A number of issues may be underlying their performance, including but not limited to poor decoding and/or fluency. Additional data is necessary to determine their needs.

Scaling and Equating

Using the 2016 National Field Test sample and the 2017-18 National Calibration sample, we calibrated items along the the IRT based causal comprehension and incorrect response propensity dimensions. The comprehension dimension was equated across forms and grades using linking items and a combination of concurrent calibration and fixed item parameter equating (FIPE). Capitalizing on the random assignment of forms within a grade, the incorrect propensity dimension was equated within grades using an equivalent groups design, but it has not been equated across grades.

Linking Item Design

In the 2017-18 National Calibration sample, all forms contained ten common items, three third grade items, four fourth grade items and three fifth grade items. These items were used as linking items in the concurrent and FIPE calibration processes. These ten items were placed in the middle of the tests with locations between item 10 and item 30. A given linking item had the same location on all forms.

Comprehension Dimension Calibration

In calibrating items along the comprehension dimension, each item was scored as 1 if correct and 0 if incorrect. A three-parameter logistic model was fitted to all items using the ten common items as anchor items and with the pseudo-guessing parameter constrained equal for all items. The item parameters of each anchor item were held equal across all grades and forms.

The calibration proceeded in two steps, the first involving a concurrent calibration, and the second involving a FIPE.

In step 1, we used the item responses to all items administered in 2017 to perform a concurrent calibration of all items administered in 2017. Linking item parameters were held constant across all grades and forms. The mean and standard deviation of θ were fixed to 0.0 and 1.0 respectively in fourth grade, but were freely estimated in third and fifth grades. Since students were randomly assigned to forms within grades, we assumed the same population mean and standard deviation for all forms within a grade. This process constitutes a concurrent calibration of all items used in 2017.

In step 2, we first augmented the item data file by adding in items from 2016 that had been dropped in 2017. In other words, the data set for this step included items from both 2016 and 2017 for all forms and grades, except for a few items that had been dropped based on DIF or other item statistics. These items were calibrated fixing the parameters for the linking items to the values obtained in step 1 but freely estimating the parameters for all non-linking items. The process constitutes a fixed item parameter equating with the fixed parameters being the item parameters for the linking items fixed at the values obtained in step 1. In our final results, the parameters for the linking items are the values obtained in step 1. For all other items, the parameter values are those obtained in step 2.

Error Propensity Dimension Calibration

In the calibration of items along the error propensity dimension, a response was scored as 2 for a paraphrase response, 1 for a causal coherent response, and 0 for an elaboration responses. For an explanation of this scoring and its rationale, see Liu, Kennedy, Seipel, Carlson, Biancarosa, and Davison (under review). Responses were fitted using the graded response model. All items were calibrated separately by form and grade using the responses of all students who had responded to that item in a particular form and grade either in 2016 or 2017.

Reliability and Precision

Classical Test Theory: Reliability of Raw Scores

2016 National Field Test Data: Descriptive Statistics. Table 2 shows the mean scores for the correct responses (CCI), and the two types of error responses (PAR, ELA) by grade and form for the 2016 Field Test data. For the correct responses, the mean scores increased by grade. For the three third grade forms, the mean scores ranged from 21.72 to 22.96. For 4th grade, they ranged from 24.97 to 25.76. In 5th grade, they ranged from 26.74 to 27.41. Within a grade, the range of means was never more than 1.25 points reflecting the random assignment of students to forms.

For the paraphrase responses, the means declined over grades. For 3rd – 5th grades respectively, the means ranged from 6.74 – 7.92, 5.53 – 5.71, and 5.23 – 5.31. For the Elaboration responses, the means fell from 3rd – 4th grades, but there was overlap of means in 4th and 5th grades. For 3rd – 5th grades respectively, the Elaboration means ranges dfrom 5.43 – 6.77, 5.02 – 5.43, and 4.55 – 5.86.

Table 2
Means and Standard Deviations (in Parentheses) of Raw Scores for MOCCA by Form and Grade in the 2016 National Field Test Sample

	CCI	PAR	ELA
Form 3.1	21.93 (10.81)	7.92 (6.81)	5.90 (4.46)
Form 3.2	21.72 (10.24)	7.59 (6.33)	6.77 (4.50)
Form3.3	22.96 (10.68)	6.74 (5.93)	5.43 (4.49)
Form 4.1	25.76 (10.62)	5.56 (5.61)	5.07 (4.41)
Form 4.2	25.75 (10.60)	5.71 (5.69)	5.02 (4.36)
Form 4.3	24.97 (10.65)	5.53 (5.40)	5.26 (4.58)
Form 5.1	26.93 (9.65)	5.23 (5.15)	5.86 (4.48)
Form 5.2	27.41 (9.81)	5.28 (5.13)	4.55 (4.15)
Form 5.3	26.74 (10.36)	5.34 (5.44)	5.24 (4.64)

Note: CCI = Causal Coherent Inference, PAR = Paraphrases, ELA = Elaboration

2016 National Field Test Data: Internal Consistency Reliability (Alpha) and Standard Errors. Internal consistency estimates of reliability were computed from the 2016 National Field Test data. These reliabilities are shown in Table 3. The reliabilities for the correct responses (CCI) are excellent, all above .90. Those for the paraphrase (PAR) responses are also excellent, above. 85. Those for the elaboration responses (ELA) are somewhat lower but still good; all are above .70 and all but one are above .75. Classical test theory estimates of

standard errors computed from the internal consistency estimates of reliability are shown by grade and form in Table 4.

2017-18 National Calibration Sample: Descriptive Statistics. Tables 5 – 7 show data from the 2017 Calibration sample similar to that in Tables 2 – 4 for the 2016 National Field Test sample: raw score means and standard deviations, internal consistency reliability coefficients (alpha), and standard errors of measurement. Table 5 shows the means and standard deviations of the raw scores.

The means in Table 5 show many of the same trends as do those in Table 2. The CCI means generally increase with grade, although there is overlap between the means in 3rd and 4th grades. The PAR and ELA means generally decline with grade, although there is a small overlap for the ELA means in 3rd and 4th grades as well as 4th and 5th grades.

2017-18 National Calibration Sample: Internal Consistency (Alpha) and Standard Errors. Table 6 shows the internal consistency reliabilities for raw scores from the 2017-18 National Calibration sample. They are slightly higher than those in Table 3, especially for the ELA score. The reliabilities for the CCI are excellent, all above .94. Those for PAR are also excellent, all above .87. Those for ELA are very good, all above .80. Table 7 shows the standard error of measurement for each score estimated from the internal consistency reliability in Table 6 and the standard deviation in Table 5.

Table 3

Internal Consistency Estimates of Reliability (Coefficient Alpha) by Form and Grade for Raw Scores for the 2016 National Field Test Sample

	CCI	PAR	ELA
Form 3.1	0.939	0.886	0.763
Form 3.2	0.925	0.865	0.721
Form 3.3	0.933	0.863	0.779
Form 4.1	0.937	0.876	0.796
Form 4.2	0.933	0.869	0.793
Form 4.3	0.934	0.861	0.804
Form 5.1	0.927	0.859	0.766
Form 5.2	0.931	0.86	0.792
Form 5.3	0.940	0.872	0.815

Note. CCI = Causal Coherent Index, PAR = Paraphrase, ELA = Elaboration

Table 4

Classical Test Theory Estimates of Standard Errors Computed from the Internal Consistency Estimates of Reliability in Table 3 and the Standard Deviations in Table 2

	CCI	PAR	LCN
Form 3.1	2.671	2.298	2.169
Form 3.2	2.801	2.328	2.373
Form 3.3	2.761	2.195	2.114
Form 4.1	2.664	1.983	1.990
Form 4.2	2.744	2.058	1.984
Form 4.3	2.735	2.014	2.028
Form 5.1	2.612	1.976	2.157
Form 5.2	2.603	1.931	1.895
Form 5.3	2.531	1.945	1.980

Note: CCI = Causal Coherent Index, PAR = Paraphrases ELA = Elaboration

Table 5
Means and Standard Deviations (in Parentheses) of Raw Scores for MOCCA by Form and Grade
in the 2017-18 National Calibration Sample

	CCI	PAR	ELA
Form 3.1	23.43 (11.60)	7.75 (7.08)	5.24 (4.68)
Form 3.2	24.97 (11.29)	6.30 (6.59)	5.47 (4.83)
Form3.3	23.76 (11.40)	7.73 (6.99)	4.96 (4.55)
Form 4.1	23.09 (11.90)	5.49 (5.90)	4.97 (4.75)
Form 4.2	24.78 (11.68)	5.68 (6.19)	4.46 (4.52)
Form 4.3	25.43 (11.07)	5.22 (5.86)	4.60 (4.56)
Form 5.1	28.75 (10.60)	4.47 (5.36)	4.13 (4.21)
Form 5.2	28.88 (10.98)	4.56 (5.13)	3.98 (4.47)
Form 5.3	28.28 (10.41)	4.53 (5.17)	4.56(4.58)

Note: CCI = Causal Coherent Inference, PAR = Paraphrases, ELA = Elaboration

Table 6
Internal Consistency Estimates of Reliability (Coefficient Alpha) by Form and Grade for Raw Scores in the 2017-18 National Calibration Sample

	CCI	PAR	ELA
Form 3.1	0.947	0.906	0.820
Form 3.2	0.950	0.900	0.813
Form 3.3	0.947	0.895	0.804
Form 4.1	0.950	0.893	0.830
Form 4.2	0.949	0.900	0.824
Form 4.3	0.945	0.891	0.816
Form 5.1	0.943	0.892	0.821
Form 5.2	0.947	0.878	0.848
Form 5.3	0.942	0.878	0.833

Note. CCI = Causal Coherent Index, PAR = Paraphrase, ELA = Elaboration

Table 7
Classical Test Theory Estimates of Standard Errors Computed from the Internal Consistency Estimates of Reliability in Table 6 and the Standard Deviations in Table 5

	CCI	PAR	LCN
Form 3.1	2.671	2.171	1.986
Form 3.2	2.525	2.084	2.089
Form 3.3	2.624	2.265	2.014
Form 4.1	2.661	1.930	1.958
Form 4.2	2.638	1.957	1.896
Form 4.3	2.596	1.935	1.956
Form 5.1	2.531	1.761	1.781
Form 5.2	2.528	1.792	1.743
Form 5.3	2.507	1.806	1.872

Note. CCI = Causal Coherent Index, PAR = Paraphrases ELA = Elaboration

Item Response Theory

Internal consistency estimates of reliability for IRT based comprehension and incorrect response propensity were examined by calculating marginal reliabilities in the 2016 National Field Test sample and the 2017-18 National Calibration sample (see Tables 8 and 9).

Marginal reliability. When examined using IRT, the first dimension (comprehension) had excellent reliability, ranging from .88 to .91, while the second, or incorrect response propensity dimension, had less satisfactory reliabilities ranging from .55 to .70.

Similar reliabilities were found in the 2017-18 National Calibration Sample. The reliabilities for the Comprehension Dimension were excellent ranging from .86 to .90. Those for the Incorrect Response Propensity Dimension were less satisfactory, ranging from .58 - .77. The marginal reliabilities were lower in Grade 5 than either Grade 3 or 4. As the number of incorrect responses by students decreases, the reliability of their incorrect response propensity scores may decrease as there are fewer incorrect responses from which to estimate the propensity.

Test-retest Reliability over Time

A subset of students in the 2017-18 National Calibration Sample took the same form a second time. The sample included 71 3rd graders, 64 4th graders, and 61 5th graders. Table 10 shows the test-retest correlation for each raw score and the two IRT scores, comprehension and incorrect response propensity. At each grade, students took Form 1 twice: that is, Form 3.1 in 3rd grade, Form 4.1 in 4th grade, and Form 5.1 in 5th grade.

The raw CCI score and its IRT counterpart, the Comprehension Dimension, had excellent test-retest reliabilities over time, all over .85 except for the CCI score in Grade 3 where the reliability was just under .85 (.847). Reliabilities for the Paraphrase score were all over .7. Those for the Elaboration score were all over .6. The Incorrect Propensity Reliabilities were very low: ranging from .254 - .270. In 3rd - 5th grades, student reading ability is in flux. It is quite possible, that a student's propensities to choose paraphrase over elaboration responses may not be stable during this period of reading development. This score would seem to reflect a student's instructional needs at a given moment, needs that change during the instructional process.

Alternate Form Reliability over Time

A subset of the 2017-18 National Calibration sample took a test and an alternate form. The sample sizes were small, 25 in 3rd grade, 33 in 4th grade, and 32 in 5th grade. At each grade, the students took Form 2 and Form 3 with approximately half taking Form 2 first and half taking Form 3 first.

Table 11 shows the alternate form reliabilities. The reliabilities for the CCI and Comprehension dimensions were all over .80 except that for the CCI in Grade 3 where the reliability was just under .80 (.797). The reliabilities for the PAR and ELA were all over .70, except for ELA in Grade 5 (.601). Those for the IRP dimension were low and seemed to decrease by grade. Like the Test-retest reliabilities of Table 10, the Alternate Form reliabilities

in Table 11 suggest that a student's Incorrect Response Propensity may characterize a student at a point in time and may suggest a needed instruction at that time, but the student's propensity and instructional needs may change over time.

Table 8

Marginal Reliabilities for the Comprehension and Incorrect Response Propensity Dimensions in the 2016 Field Test Sample

Form	Comprehension	Incorrect Response
3.1	0.91	0.70
3.2	0.91	0.62
3.3	0.91	0.59
4.1	0.90	0.55
4.2	0.89	0.62
4.3	0.90	0.60
5.1	0.88	0.55
5.2	0.88	0.60
5.3	0.88	0.59

Table 9

Marginal Reliabilities for the Comprehension and Incorrect Response Propensity Dimensions in the 2017-18 National Calibration Sample

Form	Comprehension	Incorrect Response
3.1	.90	.77
3.2	.88	.62
3.3	.88	.68
4.1	.88	.65
4.2	.88	.67
4.3	.88	.63
5.1	.86	.62
5.2	.86	.62
5.3	.87	.58

Table 10

Test-retest Reliability Over Time by Grade for Raw and IRT-based MOCCA Scores

Grade	CCI	PAR	ELA	Comprehension Dimension	IRP Dimension
3	.847	.779	.613	.860	.216
4	.877	.717	.696	.876	.220
5	.895	.791	.782	.890	.070

Note. CCI = Number of Causal Coherent Inferences, PAR = number of paraphrase inferences, ELA = number of elaboration inferences, and IRP = Incorrect Response Propensity

Table 11

Alternate Forms Reliability Over Time by Grade for Raw and IRT-based MOCCA Scores

Grade	CCI	PAR	ELA	Comprehension Dimension	IRP Dimension
3	.797	.727	.781	.805	.632
4	.890	.812	.757	.827	.398
5	.887	.854	.609	.883	NS

Note. NS = not significant at the .05 level. CCI = Number of Causal Coherent Inferences, PAR = number of paraphrase inferences, ELA = number of elaboration inferences, and IRP = Incorrect Response Propensity

Validity

Our goal in the validity studies reported below was to evaluate the criterion and construct related validity of MOCCA. In evaluating construct validity, we studied the convergent validity of MOCCA by correlating scores with other reading tests or English language arts tests that contain a reading component. We predicted that MOCCA would demonstrate discriminant validity by correlating more highly with reading/language arts tests than with mathematics tests. We were particularly interested in the relationship between MOCCA and performance on statewide assessments, given the importance of those statewide assessments in the educational process.

Construct Validity Study 1

Using a subset of the 2015 Pilot Study for which data on other reading and math data were available, we correlated the CCI scores with reading and math scores from other assessments (Davison, Biancarosa, Carlson, & Seipel, 2018). Table 12 is taken from that study. Results must be interpreted with some caution since the test in the 2015 Pilot Study included a fourth response option for each item that is not found on currently operational forms. Table 12 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 form differences may have affected these results. Two trends are notable. First, all of the correlations with the criterion reading assessments (Easy CBM CCSS comprehension, OAKS reading, California state ELA, and STAR) are significant ($p < .01$), ranging from .549 - .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 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. For a more detailed description of this study, readers can consult Davison et al, (2018).

Table 12
Convergent and Discriminant Validity: Correlations of MOCCA Total Correct Score with Reading and Math Scores of Other Tests (Sample Sizes in Parentheses)

	3 rd Grade	4 th Grade	5 th Grade
Easy CBM CCSS	.549**	.612**	.665**
Comprehension	(36)	(63)	(29)
NCTM Math	.462**	.430**	.501**
	(36)	(63)	(29)
Oaks Reading	NA	.679**	.575**
		(97)	(112)
Oaks Math	NA	.567**	.467**
		(97)	(112)
CA State Test: ELA	NA	NA	.651**
			(72)
CA State Test: Math	NA	NA	.615**
			(72)
Star Reading	NA	NA	.674**
			(73)
CA State Test Math	NA	NA	.609**
			(73)

Note. NA = no data available for that grade, reading test, math test combination.

** $p < .01$. Taken from Davison et al., 2018, p. 176.

Predictive Validity Study

Using a subset of the 2016 National Field Test for which data on the Smarter Balanced Assessment Consortium English Language Arts (ELA) test were available, we evaluated MOCCA’s usefulness as a tool for identifying students who will subsequently fail to reach the proficient level on the SBAC assessment. (Biancarosa, et al., in press). For a description of the SBAC Language Arts test, see below. In this study, we were particularly interested in whether information about the pattern of errors made by a student added over and above the total score (CCI score) in predicting whether a student would fail to reach the proficient level. Therefore,

we studied six MOCCA scores: CCI = number correct, NI = number incorrect, PAR = number of paraphrase responses, ELA = number of elaborations, and NR = number of not-reached items. There are three ways that one can fail to give a correct answer on MOCCA: giving a paraphrase response (PAR), giving an elaboration (ELA), or not answering the item (NR). The comprehension efficiency variable (CE) was also included in this study, although we give it less attention here. In this analysis, we were interested in the question of whether students with different patterns of errors would have unequal probabilities of failing to reach proficiency? For instance, would a student who primarily omitted an item when they responded incorrectly have a higher or lower probability of not reaching proficiency on the SBAC than would a student who primarily chose an elaborative response when they responded incorrectly?

Table 13 compares the mean score for SBAC Proficient and Not Proficient students on each MOCCA score by grade. SBAC Proficient and Not Proficient students differed significantly ($p < .05$), not just in their total scores CCI, but on all of the MOCCA scores, at every grade including their paraphrase (PAR), elaboration (ELA), Not Reached (NR), and Comprehension Efficiency (CE) scores. However, the effect sizes were largest for the total score.

We then fitted three logistic regression models to the data using a logistic version of a linear model for evaluating total scores and subscores proposed by Davison & Davenport (2002, Davison, Davenport, . The first model contained only a single predictor, CCI. The second included three predictors corresponding to ways of incorrectly answering an item: PAR, ELA, and NR. The third added a fourth predictor to Model 2, comprehension efficiency CE. Using the likelihood ratio test to compare the models, Model 2 with three error predictors fit the data significantly better ($p < .05$) than did the model with a single, total score predictor. This leads to the conclusion that people with the same number of MOCCA incorrect responses, but different patterns of incorrect responses, did NOT necessarily have the same probability of not reaching proficiency on the SBAC. Especially at grade 3, students with a large proportion of elaborative incorrect responses were at greater risk of not reaching proficient. Those with a large proportion of not reached incorrect responses were at lower risk of not reaching proficient.

Table 14 shows results for 3rd graders, including the proportion of correct predictions, sensitivity, and specificity for the model that included only the number correct and the model that included the three error pattern variables as predictors.

The model that included the three incorrect response vectors fit the data better and modestly improved prediction. Readers interested in more detail concerning this study are referred to Biancarosa et al. (in press).

Table 13

Means, Standard Deviations, and Effect Sizes for MOCCA Scores by Grade and Proficiency

Score	<i>M</i>		<i>SD</i>		<i>g</i>
	Proficient	Not Proficient	Proficient	Not Proficient	
Third grade (<i>n</i> proficient = 120; <i>n</i> not proficient = 125)					
CCI	26.87	14.96	8.57	7.87	1.44**
NI	13.13	25.04	8.57	7.87	-1.44**
PAR	3.28	8.17	4.23	5.48	-0.99**
ELA	3.10	6.94	2.54	4.57	-1.03**
NR	6.75	9.93	8.84	10.37	-0.33*
CE	1.46	2.63	1.09	1.70	-0.81**
Fourth grade (<i>n</i> proficient = 119; <i>n</i> not proficient = 91)					
CCI	31.08	16.65	8.84	9.69	1.56**
NI	8.92	23.35	8.84	9.69	-1.56**
PAR	1.97	6.37	3.11	5.54	-1.01**
ELA	2.23	5.10	2.53	4.21	-0.85**
NR	4.72	11.88	8.38	11.45	-0.73**
CE	1.06	2.71	0.56	3.85	-0.64**
Fifth grade (<i>n</i> proficient = 101; <i>n</i> not proficient = 69)					
CCI	31.89	20.49	8.15	9.56	1.30**
NI	8.11	19.51	8.15	9.56	-1.30**
PAR	1.88	4.17	2.63	3.94	-0.71**
ELA	1.91	4.55	2.32	3.96	-0.85**
NR	4.32	10.78	7.87	9.81	-0.74**
CE	0.96	1.84	0.40	1.25	-1.02**

* $p < .05$. ** $p < .01$. *Note.* CCI = Number correct; NI = Number incorrect; PAR= Number of paraphrases; ELA = Number of elaborations; NR = Number not reached; CE = Comprehension efficiency (i.e., minutes per correct response). Adapted from Biancarosa et al. (in press).

Table 14

Comparison of Logistic Regression Models with One (Causal Coherent) and Three (Paraphrase, Elagorate, and Not Reached) Predictors for Identifying Below Proficient students on the SBAC

	Number Correct	Paraphrase, Elaboration, and Not Reached
Predictive Power	76%	78%
Specficity	73%	76%
Sensitivity	79%	79%
-2 Log Likelihood	234.99	253.34
df	1	3
Nagelkerke pseudo-R ²	.437	.538

Note: -2 Log Likelihood, df, and Nagelkerke pseudo-R² results taken from Biancarosa et al. (in press).

Construct Validity Study 2.

In a further analysis of relations between MOCCA and statewide tests, we correlated SBAC ELA and math scores with MOCCA scores for students in the 2017-18 National Calibration sample for whom SBAC scores were available. The analysis was limited to students with both SBAC ELA and mathematics scores. A similar analysis was performed for the Tennessee Readiness (TNReady) assessment English language arts and mathematics assessments. However, the analysis of TNReady was conducted only for 3rd grade since we limited these analyses to samples with at least 50 people. As evidence of convergent validity, it was predicted that MOCCA would correlate significantly ($p < .05$) with the ELA assessments. As evidence of discriminant validity, it was predicted that correlations with the ELA assessments would be higher than those for the mathematics assessments.

Smarter Balanced Assessment Consortium English Language and Arts and Mathematics (SBAC ELA and Math). SBAC is one of the Common Core States Standards aligned measures for Grades 3-8 developed by a consortium of 15 states including States of Oregon, California, Michigan, and Connecticut. SBAC ELA and Mathematics are summative, individually-administered, computer-adaptive tests. Across subjects, SBAC is comprised of two components: (a) The Computer-Adaptive Test (CAT) using traditional assessment questions, such as multiple choice, matching tables, and drag and drop; (b) The Performance Tasks (PT) use interactive activities to assess students' ability to apply their critical-thinking and problem solving skills for a set of complex real-world problems that are coherently connected to a common theme. In PT, tasks are provided in a variety of information forms (i.e., readings, video clips, data), to which students respond either in writing or speaking. SBAC is not timed, but the estimated testing time for SBAC ELA is about 3.5 hours, and for SBAC math is about 2.5 hours.

In Grades 3-5, SBAC ELA is comprised of 43-47 items on the four domain-specific claims: reading, writing, speaking/listening, and research. Each claim consists of assessments

targets that represent more detailed information of content. Students' scores are reported as IRT scale scores ranging between 2,000 and 3,000, with 2,432 in Grade 3, 2,473 in Grade 4, and 2,502 in Grade 5 representing the scores for meeting the state ELA achievement standards. Reported reliability measured by the standard error of measurement (SEM) ranged from 24.7 – 26 (Smarter Balanced Assessment Consortium, 2016). No validity information is available yet.

In Grades 3-5, SBAC Math is comprised of 33-40 items on the three domain-specific claims: concepts and procedures, problem solving/modeling and data analysis, and communicating and reasoning. As with SBAC ELA, each claim consists of assessment targets. Students' scores are reported as IRT scale scores ranging between 2,000 and 3,000, with 2,436 in Grade 3, 2,485 in Grade 4, 2,528 in Grade 5 representing the score for meeting the state math achievement standards. In Grades 3-5, reported reliability measured by SEM ranged from 19.5 – 23.7 (Smarter Balanced Assessment Consortium, 2016). No validity information is available yet.

TNReady. The TNReady English language arts (4 subparts) exam assesses the Tennessee Academic Standards through literary and informational texts requiring students to demonstrate the ability to read closely, analyze text, answer text-dependent questions, provide a written response to a prompt, and demonstrate command of the English language. Additionally, in grades 3 and 4, fluency, comprehension, and listening skills are measured. The subparts are administered in separate sessions that take approximately three hours total with the exact testing time depending on grade.

The TNReady mathematics assessment Mathematics (3 subparts) contains calculator permitted and calculator prohibited subparts. It assesses the Tennessee Academic Standards requiring students to demonstrate a deep conceptual understanding of mathematics, number sense, fluency, problem solving and an understanding of the grade-level horizontal coherence embedded within the standards. The mathematics test will focus approximately 70 percent of the assessment items on major work of the grade and approximately 30 percent of the items on supporting work. The assessment requires approximately an hour and a half.

Convergent Validity. Table 15 shows the correlations of MOCCA with English language arts, reading and mathematics scores of the SBAC and TNReady assessments. Correlations with both the MOCCA CCI raw scores (CCI) and IRT-based, equated comprehension dimension scaled scores (Comp.) are shown. At all three grades, both the MOCCA CCI and Comp. scores were correlated significantly ($p < .01$) with the SBAC ELA scores. The SBAC correlations range from .570 for the MOCCA CCI and SBAC language arts scores in 4th grade to .785 for the MOCCA Comp. and SBAC language arts scores in 3rd grade. Similarly, the MOCCA scores are both significantly correlated with the TNReady readings scores ($p < .01$). These correlations support the predictive and convergent validity of MOCCA comprehension scores.

Table 15

Convergent and Discriminant Validity: Correlations of MOCCA CCI Raw Score and Comprehension Dimension with SBAC and TNReady English Language Arts, Reading, and Mathematics Scores (Sample Sizes in Parentheses)

		3 rd Grade	4 th Grade	5 th Grade
SBAC	ELA & CCI	.643** (247)	.570** (287)	.671** (147)
	ELA & Comp.	.785** (247)	.745** (287)	.725** (147)
	Math & CCI	.560** (192)	.464** (274)	.545** (139)
	Math & Comp.	.608** (192)	.593** (274)	.543** (139)
TNReady	Reading & CCI	.644** (106)	NA	NA
	Reading & Comp.	.681** (106)	.NA	NA
	Math & CCI	.525** (107)	NA	NA
	Math & Comp	.548** (107)	NA	NA

Note. NA = no data available for that grade, SBAC = Smarter Balanced Assessment Consortium, TNReady = Tennessee Readiness Assessment, ELA = English Language Arts, CCI = Causal Coherent Inference (Number Correct Score on MOCCA, Comp. = Scaled Score on MOCCA Comprehension Dimension

** $p < .01$

Discriminant Validity. MOCCA scores are also significantly correlated with SBAC and TNReady mathematics scores, but less highly correlated with the mathematics scores than the reading/language arts scores (see Table 15). For instance, in 3rd grade, the MOCCA Comp score correlates .785 with SBAC language arts but .608 with SBAC mathematics. As another example, in 4th grade MOCCA CCI correlates .570 with SBAC ELA but only .464 with SBAC mathematics. A similar pattern exists in the TNReady data. For instance, the correlation of MOCCA Comp with TNReady reading is .681 whereas the correlation of MOCCA Comp with mathematics is .548. The lower correlations of MOCCA scores with reading than with mathematics scores on the SBAC and TNReady provide support for the discriminant validity of MOCCA. These results support the conclusion that MOCCA measures reading, a language art, rather than some more general cognitive ability.

Fairness

The fairness of MOCCA was examined in two ways. First, all items were subjected to a sensitivity review by teachers as part of the content review described earlier. In this step, items were evaluated by a panel of teachers as to whether the content would be offensive or would provide an advantage to one demographic group over others. Second, items were examined for differential item functioning by gender and by ethnicity (Hispanics vs. Whites). For other ethnicities, sample sizes were not sufficiently large for a DIF analysis. In addition, means of various scores are shown below by gender and ethnicity.

Sensitivity Review

As described in the Development Section, a panel of local intermediate grade teachers was engaged to review all MOCCA items before they were used. Teachers were paid for their time and asked to render their professional opinion as to the fairness of MOCCA items, among other things. Six teachers made up the panel. At least one teacher served each of the grade levels targeted by MOCCA (i.e., Grades 3-5). In addition, one teacher had expertise serving students in special education and another with serving students with limited English proficiency in the targeted grade levels. The latter worked in a Title 1 room in a Spanish-English dual language school and also had extensive English learner experience, including personal experience.

The panel reviewed MOCCA items as intact stories (with the sixth sentence replaced in the story). Teachers reviewed the items for (a) causal coherence, (b) accuracy, (c) appropriateness, (d) freedom from bias, and (e) engagingness. Bias was defined for the teachers as bias with respect to gender, ethnicity, national origin, disability status, or sexual orientation. We asked teachers to flag a story if the content would be offensive to or if it would disadvantage members of a particular group, including ones not listed.

Teachers used a four-point Likert scale to rate items on these dimensions. Specifically, teachers rated items for freedom from bias as: *Not at all*, *Marginally*, *Adequately*, and *Completely*. We define these terms below.

- **Not at all.** A story was rated as *Not at all* free from bias if the teacher had major concerns about the item with regard to offensiveness or unfair advantage. The item was considered unacceptable as-is and revision or omission was strongly recommended.
- **Marginally.** A story was rated as *Marginally* free from bias if the teacher had real concerns about the item with regard to bias. The item might work as-is, but revisions would likely improve it enough to make a difference in student performance.
- **Adequately.** A story was rated as *Adequately* free from bias if the teacher had some (mild) concerns about the item with regard to the offensiveness or unfair advantage. Although revisions might improve it, they would not be likely to make a difference in student performance. The item could be better but is ultimately okay as-is.

- **Completely.** A story is rated as *Completely* free from bias if the teacher had no concerns about the item with regard to offensiveness or unfair advantage. The item was considered perfectly acceptable as-is.

When teachers selected *Not at all* or *Marginally*, they added a clarifying comment as to why they rated the story that way. Teachers also had a space to add any general comments, concerns, or questions they had regarding each story they reviewed.

The panel was trained in a single 4-hour session with ample discussion of the purpose of MOCCA, the intent behind the dimensions they were rating, the meaning of the rating scale they used, and training in using the Qualtrics interface for rating. Two items were reviewed as a group with discussion, and then two rounds of rating several items followed. Interrater reliability was not a goal in this training because judgments of the dimensions were necessarily subjective. Instead, the goal was for consistent and appropriate use of the rating scales and Qualtrics system. Teacher reviews were completed in three waves of about 160 items each. Each teacher was randomly assigned to review half of the items in each wave. Thus, each and every MOCCA item was reviewed by three members of the teacher panel.

The author team reviewed teacher ratings and written feedback in depth. When an item was flagged for bias, it was revised if possible but retired if revision was not possible. For example, teachers consistently flagged stories with explicit references to magic or witches or similar concepts as potentially offensive to some religions; some of these stories were revised, while others were retired. Stories that inadvertently reinforced negative stereotypes were likewise revised or retired. In short, any item flagged for bias was either substantially changed or was omitted from MOCCA.

Average Score Differences

To test for demographic differences, we performed analyses of variance that included gender, race/ethnicity, English as second language, special education, free/reduced lunch, grade, and test form within grade as factors. Type III sums of squares were used to test effects at the .05 level of significance. Table 16 shows the means of MOCCA raw scores by grade and gender. While females have the higher mean number correct at every grade, the differences were not significant after taking the other factors in the design into account. Table 17 shows the MOCCA scores by grade and ethnicity. On the CCI score, Asian and White students outperformed Latino and Black students; students not on free or reduced meals outperformed those receiving free and reduced meals; those not served by special education outperformed those served by special education; and those not identified as English learners outperformed those identified as such. Results reported below are for the 2016 National Field Test sample because the item content of the 2016 forms most closely matches that for the operational forms starting in 2018. Results in Tables 16 and 17 have been aggregated across forms within a grade, and forms were randomly assigned within grade.

Table 16

Means and Standard Deviations (in parentheses) of Raw Scores by Gender in the 2016 National Field Test Sample

	Grade 3		Grade 4		Grade 5	
	Male	Female	Male	Female	Male	Female
CCI	21.19(10.48)	23.10(10.68)	24.24(10.43)	25.40(10.97)	26.48(10.08)	28.18(9.92)
PAR	7.89(6.29)	6.73(6.49)	6.05(5.55)	5.24(5.58)	5.63(5.39)	4.41(4.83)
LCN	6.34(4.70)	5.49(4.23)	5.54(4.60)	4.71(4.32)	5.50(4.60)	4.44(4.14)

Table 17

Mean and Standard Deviations (in Parentheses) of MOCCA Raw Scores by Grade and Ethnicity in the 2016 National Field Test Sample

		Grade 3	Grade 4	Grade 5
		CCI	White	23.57(10.71)
	Af. American	18.73(9.54)	20.08(9.24)	23.43(9.73)
	Hispanic	18.05(9.22)	21.14(9.87)	23.67(9.91)
	Asian	27.52(10.11)	26.31(11.47)	28.00(10.35)
PAR	White	9.46(6.31)	8.92(6.55)	7.33(5.72)
	Af. American	8.73(6.36)	6.59(5.61)	6.42(5.69)
	Hispanic	4.00(3.85)	3.33(3.88)	4.38(4.58)
	Asian	9.46(6.31)	8.92(6.55)	7.33(5.72)
ELA	White	7.63(4.89)	7.71(4.81)	6.57(4.72)
	Af. American	7.08(4.58)	5.58(4.46)	6.12(4.65)
	Hispanic	7.63(4.89)	7.71(4.81)	6.57(4.72)
	Asian	7.08(4.58)	5.58(4.46)	6.12(4.65)

Differential Item Functioning (DIF)

DIF analyses were run for all items in a form. The analysis was run form by form using all items in the form as anchors. For each item, the analysis yielded several statistics including the following: the Mantel-Haenszel chi square statistic with significance level, the odds ratio, the log odds ratio, the ETS delta statistic, and a classification of items based on the delta statistic. The three classes are A = negligible DIF, B = Moderate DIF, and C = Large DIF. Analyses were run for gender and for White vs. Hispanic ethnicity. The sample sizes for other ethnicities were deemed to be small for analysis. Overall, the evidence for item DIF was negligible, both for gender and White vs. Hispanic ethnicity.

In the gender analysis, the 360 items (40 items per form, 3 forms per grade, and 3 grades) were analyzed, and only four items displayed significant DIF at the .05: none in 3rd grade, two in 4th grade, and two in 5th grade. The proportion of items displaying significant DIF, .01, was less than the expected number given an alpha level of .05.

Results were extremely similar for the White vs. Hispanic analysis. Only four of the 360 items displayed significant DIF. The number of items displaying DIF could readily be explained by sampling error if the null hypothesis of no DIF was true for every item.

Nevertheless, items identified as having significant DIF were subjected to an author review of items and response alternatives. Based on this review, the eight unique items displaying DIF were dropped and replaced for the forms that became operational in Fall 2018.

Norming

A major goal for MOCCA was to ensure the national representativeness of MOCCA norming data. In service to this goal, procedures were used to recruit students from across the country with the goal of matching national Census proportions for geographical region (i.e., South, Midwest, Northeast, and West), school locality (i.e., city, suburb, town, rural), and individual student demographics (i.e., gender, race/ethnicity).

Sampling

Participants were recruited through local connections and the DIBELS Data System (DDS) and word of mouth. The DDS is operated by the Center on Teaching and Learning (CTL) at the University of Oregon. The DDS itself exists as a protected research project at the UO, and historically university researchers have offered structured opportunities for school partners to participate in research. Email announcements via DDS were sent intermittently during the 2015-16 and 2016-17 school years and in the fall of 2017. The DDS website also included a news announcement about the study during active recruitment periods during the study. In addition, local connections with district and school personnel and word-of-mouth through other researchers, as well as limited cold calling, were used to round out the sample in areas where recruitment was initially weak (e.g., the South).

The number of students by grade in the 2016 National Field Test and 2017-18 National Calibration samples is depicted in Table 18, and their breakdown by state and region are reported in Tables 19 and 20 by year. Note that students tested off grade level were not included in the final norming sample, and data from students who participated in testing in more than one year were only used for the first year in which they participated. Thus, *ns* do not sum to those in Table 18. As is evident from Tables 20 and 21, the West was over-represented, while other regions were under-represented.

Table 18

Sample Sizes by Grade and Academic Year for the Norming Sample

	Grade 3	Grade 4	Grade 5	Total
2015-2016	1577	1498	1215	4290
2016-2017	887	781	647	2315
2017-2018	558	452	399	1576
Total	3022	2731	2261	8181

Similarly, as noted in Table 1, although the norming sample was diverse, it did not perfectly reflect proportions of student groups (i.e., gender and race/ethnicity) in the population. As a result, student scores were weighted using the latest Condition of Education report for proportions of students in the United States of different genders and racial and ethnic backgrounds.

Appendix B contains a table of norms for the IRT-based Comprehension score. This score has been equated across grades and forms. Thus, a given score corresponds to the same absolute level of ability in all grades, but a given scale score will have a lower percentile rank in 5th grade than in 4th or 3rd. Thus, for each possible percentile, the table in Appendix B shows the score or range of scores associated with each percentile rank from 1 – 99 for each grade separately.

Table 19
Number of Participants in Norming Sample by State and Grade for the 2016 Field Test Sample

	Region	District	Schools	Grade 3 N (%)	Grade 4 N (%)	Grade 5 N (%)	Tested Off Grade N (%)	Total N (%)
AL	South	1	2	122 (7.6)	108 (6.9)			230 (5.2)
AZ	West	1	4	171 (10.6)	117 (7.5)	168 (13.7)	52 (77.6)	508 (11.4)
CA	West	5	15	397 (24.6)	363 (23.3)	261 (21.2)		1,021 (22.9)
CO	West	1	1	19 (1.2)	17 (1.1)	22 (1.8)		58 (1.3)
MA	Northeast	1	1	16 (1.0)	14 (1)	14 (1.1)		45 (1.0)
MI	Midwest	2	2	108 (6.7)	106 (6.8)	62 (5.0)		276 (6.2)
MN	Midwest	2	2	70 (4.3)	76 (4.9)	74 (6)	15 (22.4)	235 (5.3)
MO	Midwest	1	1	22 (1.4)	20 (1.3)	20 (1.6)		62 (1.4)
OH	Midwest	2	5	133 (8.3)	118 (7.6)	16 (1.3)		267 (6)
OR	West	7	16	261 (16.2)	330 (21.2)	342 (27.8)		933 (20.9)
PA	Northeast	4	11	213 (13.2)	218 (14)	228 (18.5)		659 (14.8)
SC	South	1	1	37 (2.3)	36 (2.3)			73 (1.6)
TX	South	2	2	24 (1.5)	13 (0.8)	23 (1.9)		60 (1.3)
WA	West	2	2	19 (1.2)	19 (1.2)			38 (0.9)
Total				1,611	1,557	1,230	67	4,465
N (%)		32	65	(36.1)	(34.9)	(27.5)	(1.5)	(100)

Table 20
Number of Participants in Norming Sample by State and Grade for the 2017-18 National Calibration Sample

State	Region	Districts	Schools	Grade 3 <i>N (%)</i>	Grade 4 <i>N (%)</i>	Grade 5 <i>N (%)</i>	Total <i>N (%)</i>
AZ	West	3	6	282 (19.3)	289 (22.2)	243 (20.8)	814 (20.7)
CA	West	4	10	160 (11.0)	115 (8.8)	88 (7.5)	363 (9.2)
CO	West	1	1	11 (0.8)	21 (1.6)	0 (0.0)	32 (0.8)
DC	South	2	2	65 (4.4)	17 (1.3)	27 (2.3)	109 (2.8)
FL	South	1	1	5 (0.3)	2 (0.2)	5 (0.4)	12 (0.3)
GA	South	3	3	199 (13.6)	179 (13.7)	204 (17.5)	582 (14.8)
IL	Midwest	2	2	90 (6.2)	58 (4.4)	75 (6.4)	223 (5.7)
KS	Midwest	1	1	0 (0.0)	25 (1.9)	0 (0.0)	25 (0.6)
MD	South	1	1	56 (3.8)	38 (2.9)	43 (3.7)	137 (3.5)
MI	Midwest	1	1	0 (0.0)	0 (0.0)	23 (2.0)	23 (0.6)
MN	Midwest	1	4	161 (11.0)	92 (7.1)	50 (4.3)	303 (7.7)
NJ	Northeast	1	1	25 (1.7)	23 (1.8)	27 (2.3)	75 (1.9)
OH	Midwest	1	1	0 (0.0)	39 (3.0)	20 (1.7)	59 (1.5)
OR	West	5	12	246 (16.8)	302 (23.2)	265 (22.7)	813 (20.7)
PA	Northeast	2	2	36 (2.5)	27 (2.1)	60 (5.1)	123 (3.1)
TN	South	1	1	114 (7.8)	34 (2.6)	0 (0.0)	148 (3.8)
UT	West	1	1	11 (0.8)	14 (1.1)	11 (0.9)	36 (0.9)
VT	Northeast	1	1	0 (0.0)	29 (2.2)	26 (2.2)	55 (1.4)
Total		32	51	1461 (37.2)	1304 (33.2)	1167 (29.7)	3932 (100)

References

- August, D., Francis, D. J., Hsu, H. Y. A., & Snow, C. E. (2006). Assessing reading comprehension in bilinguals. *The Elementary School Journal*, *107*(2), 221-238.
- Biancarosa, G., Kennedy, P. C., Yoon, H.-J., Seipel, B., Carlson, S. E., Liu, B., & Davison, M. L. (in press). Constructing subscores that add validity: A case study identifying students at risk. *Educational and Psychological Measurement*.
- 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.
- Davis, B. J., Johnston, A. M., Barnes, M. A., & Desrochers, A. (2007). *Bridging inferences in children from grades three to eight*. Poster presented at the Annual Convention of Canadian Psychological Association. Halifax, Nova Scotia, Canada.
- Davison, M. L., Biancarosa, G., Carlson, S. E., & Seipel, B. (2018). Preliminary findings on the computer-administered Multiple-choice Online Causal Comprehension Assessment, a diagnostic reading comprehension test. *Assessment for Effective Intervention*, *43*(3), 169 – 181.
- Davison, M. L., Davenport, E. C. Jr., Chang, Y.-F., Vue, K., & Su, S. (2015). Criterion-related Validity: Assessing the Value of Subscores. *Journal of Educational Measurement*, *52*, 263 – 279.
- Davison, M. L. & Davenport, E. C. Jr. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods*, *7*(4), 468-484.
- Dorans, N. & Holland, P. W. (1988). Differential item performance and the Mantel-Haenszel procedure. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35 – 66). Hillsdale NJ, Erlbaum.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.
- Good, R. H. III, Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A. & Latimer, R. J. (2011). *DIBELS Next Technical Manual*. Eugene, OR: Dynamic Measurement Group.
- Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*(2), 627-635.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- Liu, B., Kennedy, P. C., Seipel, B., Carlson, S. E., Biancarosa, G., & Davison, M. L. (under review). Can we learn from student mistakes in a formative, reading comprehension assessment.
- Pike, M. M., Barnes, M. A., & Barron, R. W. (2010). The role of illustrations in children's inferential comprehension. *Journal of experimental child psychology*, *105*(3), 243-255.

- Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*(4), 304-325.
- Smarter Balanced Assessment Consortium (2015). *English language arts & literacy computer adaptive test (CAT) and performance task (PT) stimulus specifications*. Retrieved from <https://www.smarterbalanced.org/wp-content/uploads/2015/08/ELA-Stimulus-Specifications.pdf>.
- Smarter Balanced Assessment Consortium (October 6, 2016). *Smarter Balanced Assessment Consortium: 2014-2015 Technical Report*. Retrieved from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better DECISIONS through science. *Scientific American, 283*, 82-87.
- Tennessee Department of Education (2018). *Overview of testing in Tennessee*. Retrieved from <https://www.tn.gov/education/assessment/testing-overview.html>, August 30, 2018.
- Thorndike, R. M., & Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education* (8th ed.). New York, NY: Pearson

Appendix A: Teacher Panel Review Details

The teacher review panel reviewed all MOCCA items along for the following characteristics.

- **Causal coherence.** Causal coherence means the story is causally coherent. More specifically, we wanted to be sure that the sixth sentence is coherent with the story and includes *necessary* information without which the story would not make sense. We asked teachers to flag stories if the sixth sentence was not necessary to the story, especially its conclusion (the seventh sentence).
- **Accuracy.** Accuracy means that the item is free of factual errors. We asked teachers to flag a story if they found any factual errors or illogical content in a story.
- **Appropriateness.** An item is appropriate if the content falls within the domain of reading material that teachers expect students to comprehend at the given grade level. The vocabulary, syntax, sentence length, and content should be appropriate for the grade. We asked teachers to flag a story if they deemed any one of these characteristics as inappropriate for the grade(s) in which it might be used.
- **Bias.** Bias primarily means bias with respect to gender ethnicity, national origin, disability status, or sexual orientation. We asked teachers to flag a story if the content would be offensive to or if it would disadvantage members of a particular group.
- **Engagement.** Engagement is the extent to which the content of the passage will engage readers' attention. We asked teachers to rate stories based on how engaging they would be for the grade(s) in which it might be used.

Except for engagement, which was rated on an unanchored 9-point Likert scale, teachers rated each of these characteristics (sometimes multiple features per characteristic) using a four-point Likert scale: *Not at all*, *Marginally*, *Adequately*, and *Completely*. We define these terms below.

- **Not at all.** A story is rated as *Not at all* if the teacher had major concerns about the item with regard to the characteristic rated. The item is considered unacceptable as-is and revision or omission is strongly recommended.
- **Marginally.** A story is rated as *Marginally* if the teacher had real concerns about the item with regard to the characteristic rated. The item might work as-is, but revisions would likely improve it enough to make a difference in student performance.
- **Adequately.** A story is rated as *Adequately* if the teacher had some (mild) concerns about the item with regard to the characteristic rated. Although revisions might improve it, they would not be likely to make a difference in student performance. The item could be better but is ultimately okay as-is.

- **Completely.** A story is rated as *Completely* if the teacher had no concerns about the item with regard to the characteristic rated. The item is considered perfectly acceptable as-is.

When teachers selected *Not at all* or *Marginally* for any criterion, they added a clarifying comment as to why they rated the story that way. Teachers also had a space to add any general comments, concerns, or questions they had regarding each story they reviewed.

Appendix B: Norm Table for Comprehension Dimension Scale Scores by Grade

Percentile Rank	3 rd Grade		4 th Grade		5 th Grade	
	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit
1	50	134	50	134	50	157
2	135	147	135	149	158	174
3	148	157	150	162	175	189
4	158	166	163	176	190	203
5	167	173	177	186	204	217
6	174	179	187	195	218	232
7	180	188	196	204	233	245
8	189	197	205	214	246	255
9	198	205	215	224	256	264
10	206	211	225	231	265	272
11	212	216	232	238	273	278
12	217	223	239	244	279	288
13	224	232	245	249	289	304
14	233	241	250	259	305	314
15	242	249	260	269	315	322
16	250	255	270	276	323	331
17	256	261	277	283	332	338
18	262	269	284	290	339	345
19	270	275	291	298	346	352
20	276	281	299	304	353	358
21	282	286	305	310	359	362
22	287	291	311	316	363	368
23	292	297	317	323	369	374
24	298	303	324	329	375	382
25	304	308	330	336	383	390
26	309	313	337	343	391	397
27	314	319	344	351	398	406
28	320	326	352	359	407	413
29	327	333	360	365	414	420
30	334	338	366	373	421	426
31	339	342	374	380	427	432
32	343	346	381	385	433	437
33	347	351	386	391	438	441
34	352	356	392	395	442	446
35	357	361	396	400	447	452
36	362	365	401	405	453	460
37	366	370	406	410	461	465
38	371	376	411	416	466	471
39	377	382	417	420	472	476
40	383	389	421	424	477	481

41	390	396	425	428	482	487
42	397	404	429	434	488	495
43	405	409	435	440	496	501
44	410	415	441	447	502	506
45	416	421	448	454	507	511
46	422	426	455	460	512	516
47	427	430	461	465	517	522
48	431	434	466	470	523	527
49	435	440	471	476	528	531
50	441	447	477	484	532	536
51	448	452	485	489	537	541
52	453	457	490	497	542	545
53	458	463	498	507	546	549
54	464	468	508	514	550	554
55	469	473	515	519	555	559
56	474	477	520	523	560	564
57	478	482	524	529	565	569
58	483	487	530	535	570	575
59	488	492	536	542	576	580
60	493	498	543	549	581	584
61	499	502	550	556	585	589
62	503	506	557	561	590	594
63	507	511	562	567	595	599
64	512	517	568	571	600	605
65	518	522	572	576	606	610
66	523	526	577	580	611	615
67	527	532	581	585	616	621
68	533	539	586	591	622	626
69	540	546	592	597	627	630
70	547	553	598	602	631	636
71	554	559	603	607	637	642
72	560	566	608	612	643	645
73	567	572	613	617	646	648
74	573	578	618	621	649	655
75	579	585	622	626	656	661
76	586	592	627	632	662	667
77	593	599	633	638	668	674
78	600	606	639	646	675	679
79	607	612	647	652	680	683
80	613	618	653	658	684	688
81	619	625	659	665	689	693
82	626	632	666	673	694	700
83	633	639	674	678	701	707
84	640	645	679	685	708	716
85	646	651	686	693	717	723

86	652	658	694	702	724	731
87	659	667	703	709	732	738
88	668	677	710	716	739	744
89	678	687	717	722	745	752
90	688	697	723	726	753	762
91	698	705	727	731	763	781
92	706	713	732	736	782	796
93	714	728	737	743	797	799
94	729	747	744	758	800	806
95	748	769	759	780	807	812
96	770	783	781	799	813	813
97	784	785	800	811	814	821
98	786	794	812	818	822	833
99	795	950	819	950	834	950