MOCCA Computerized Adaptive Test Technical Manual

MOCCA Technical Report MTR-2024-1

WORKING DRAFT

Mark L. Davison

David J. Weiss

Joseph N. DeWeese

Yun Leng Wong

University of Minnesota

Gina Biancarosa

Patrick C. Kennedy

Seulbi Lee

University of Oregon

Suggested citation: Davison, M. L., Weiss, D. J., DeWeese, J. N., Wong, Y. L., Biancarosa, G., Kennedy, P. C., & Lee, S. (2024). *MOCCA Computer-adaptive Test Technical Manual: MOCCA Technical Report MTR-2024-1*. Eugene, OR: University of Oregon.

Contents

Acknowledgements

The computerized, adaptive version of MOCCA, the subject of this technical manual, is the result of a tremendous collaborative effort. We extend our thanks to all who contributed. Special thanks to our MOCCA item writing team: Ben Seipel, Ian Egan, John Barnett, Kristen Havens, Nicole Moore, and Rose Gowen. Thanks as well to University of Oregon graduate students who contributed many of the mundane tasks such as performing quality assurance testing of the MOCCA system and administering MOCCA in local schools; we wish to highlight Cayla Lussier, Emily Wilke, John Gallo, Meagan Dorman, and Tasia Brafford. We would be remiss if we did not acknowledge the thousands of students and their teachers for participating in our project.

MOCCA would not be possible without the support of our administrative and programming teams. We offer sincere thanks to Janet Otterstedt, who kept us on time and facilitated all of our considerable recruitment efforts and communications with schools. We also thank Eugenia Coronado and Nick Phillips for their clerical support, Julie Watts for her oversight, and Audrey Desjarlais for her fiscal management. We are indebted to our programming team Scott McCammon and Emberex, who incorporated changes to the MOCCA algorithms with aplomb. Similarly, we recognize and appreciate the support of the Center for Teaching and Learning for continuing to host MOCCA in their organization.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190393 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. We would also like to acknowledge the guidance and support of our IES program officers: Meredith Larson and Vinita Chhabra. Finally, we are grateful to our respective institutions, the University of Oregon and University of Minnesota.

1. Introduction

MOCCA is a screening and diagnostic measure of reading comprehension for Grades 3 to 6. In 2019, MOCCA had 9 40-item forms available for Grades 3 to 5 with 3 forms for each grade. The version of MOCCA reported on in this manual and available for use in schools is now a computerized adaptive test (CAT). CATs operate so that the items that each student experiences depend on how they have performed on prior items, allowing for more precise measurement of student abilities and the delivery of fewer items over less time than the 2019 version of MOCCA. In this manual, we report on the background of MOCCA, how to administer MOCCA and interpret its results, and the results of the development and validation of its CAT form.

Theoretical Foundations

Designed for students in Grades 3 through 6, MOCCA, formerly known as the Multiple-choice Online Causal Comprehension Assessment, 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, which means reading words accurately and fluently. They might have limited vocabulary and background knowledge. Nonetheless, there are some students who don't comprehend and don't fall into these categories. Researchers have dubbed the latter group students with specific reading comprehension difficulties (SRCD).

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.

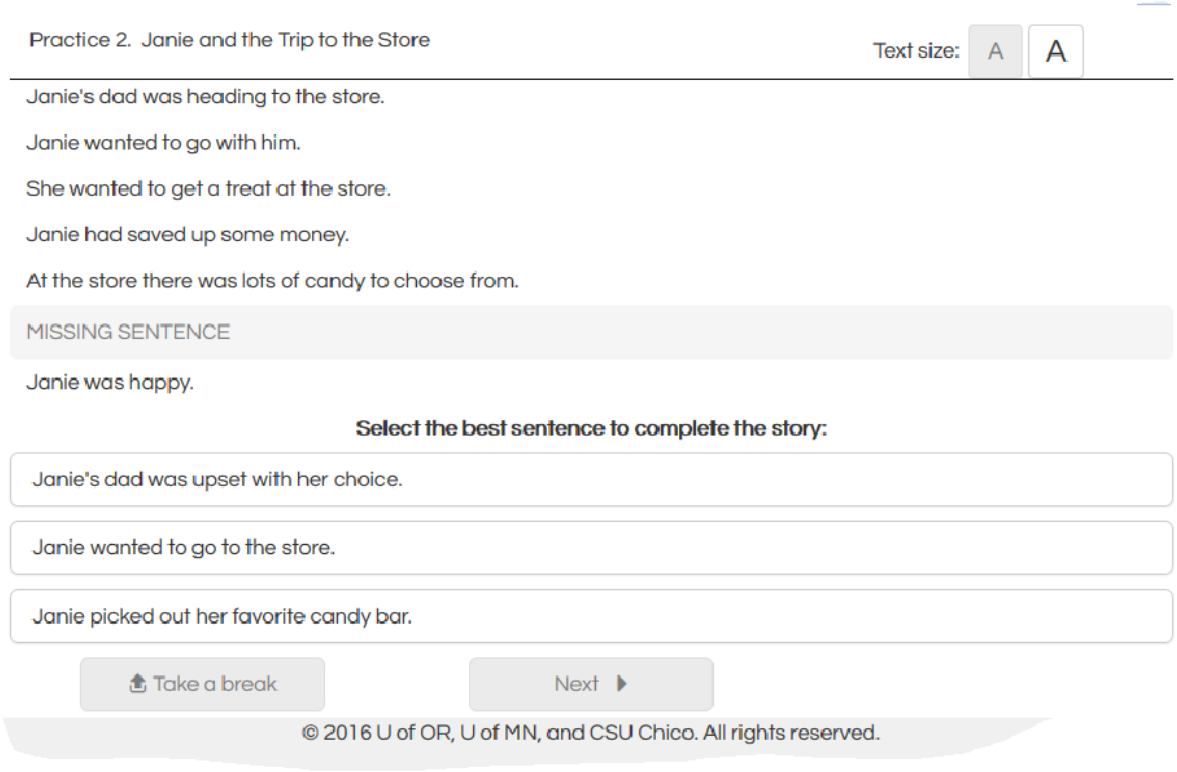
Think aloud research, in which readers report aloud what they are thinking as they read, has demonstrated that students with SRCD tend to rely on one of two cognitive processes or strategies—paraphrasing and making elaborative inferences. These are great strategies, but neither alone will result in excellent comprehension, especially when the reader is not making causally coherent inferences, which are the inferences most necessary for successful comprehension.

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. What's more, research suggests that students who rely on paraphrasing require somewhat different instruction than those making elaborative inferences.

Anatomy of a MOCCA Item

Each MOCCA item contains a seven-sentence paragraph followed by three or five response alternatives. The sixth sentence is missing from the paragraph, and from the presented alternatives, the student must pick the sentence that best completes the item. Whereas multiple-choice response alternatives are usually of two types, correct and incorrect, MOCCA

Figure 1.1 Sample MOCCA Item with Three Response Alternatives



alternatives contain a correct alternative and two types of incorrect alternatives. The correct alternative, the Causal response, is a sentence that best completes the paragraph, usually because it indicates whether the main character(s) achieved their goal. The first type of incorrect alternative, the Paraphrase, is a sentence that repeats information presented in the story. It does not add new information, and so does not move the story to completion. The second type of incorrect, the Elaboration, adds information but does not complete the story usually because it does not indicate whether the main character(s) achieved their goal. A three-alternative item contains one Causal, one Paraphrase, and one Elaboration response. A five-alternative item contains one Causal, two Paraphrase, and two Elaboration responses.

Figure 1 below shows a typical three-alternative item. The sixth sentence is missing. The first alternative is the Elaboration that adds information about her dad's reaction but does not

complete the story by indicating whether Janie achieved her goal. The second alternative is the Paraphrase repeating information in the second and third sentences. The third alternative is the Correct alternative indicating that she achieved her goal of getting a treat.

Scoring of a MOCCA Item

For each item, students receive a score on a dichotomous response variable used to derive an item response theory score θ_1 based on a three-parameter model. The response variable is

variable X_{1ij}, for person *i* (*i* = 1, …, *I*) and item *j* (*j* = 1, …, *25*):

$X_{1ij} = 1$ if the response of person *i* to item *j* is correct (1)

= 0 if the response of person *i* to item *j* is incorrect

The score θ_1 was calibrated on a common metric across all grades and forms, a metric that has mean 0 and variance 1.0 in the calibration/equating sample composed of 3rd, 4th, and 5th graders. The mean of 0.0 is the approximate mean of the 4th graders. The calibration sample is a prior sample, not the sample of the current study.

For each item, students also receive a score on a second response variable X_{2ij}:

$X_{2ij} = 1$ if person *i* chose a Paraphrase incorrect response for item *j* (2)

= 0 if person *i* chose an Elaboration incorrect response for item *j*

= missing if person *i* chose the correct answer

Using this response variable and a two-parameter logistic model, each person receives a second IRT score θ_2 on the Process Propensity dimension. θ_2 is a bi-polar dimension such that students who predominantly choose an Elaboration response when making a mistake will fall at the negative end of the dimension. Students who predominantly choose a Paraphrase response when they make a mistake will fall at the positive end. The 0 point on the scale is an indifference

points such that students with a score of 0 have a .5 probability of choosing the Paraphrase (or the Elaboration) response when they make a mistake.

The model for this second variable differs from the usual 2-PL model in that it is a conditional probability:

$$\pi_{2ij}(X_{2ij} = 1 | X_{1ij} = 0) = \frac{\exp[\alpha_2(\theta_2 - \beta_2)]}{1 + \exp[\alpha_2(\theta_2 - \beta_2)]} \quad (3)$$

a probability conditional on $X_{1ij} = 0$.

Students do not receive a score as such based on the Process Propensity Dimension. Rather they receive a classification based on θ_2 and a likelihood ratio statistic. The process begins by defining an indifference region around the indifference point, a region defined by an upper bound *UB* and a lower bound *LB*. In our case, *LB* = -0.5, *UB* = 0.5, and the indifference point is $\theta_2 = 0$. Points below the indifference region are all in the Elaboration region; points above the indifference region are all in the Paraphrase region. Next, we compute two likelihoods. Then we find the point in the region $\theta_2 > UB$ that maximizes the likelihood of the student's response vector for the second response variable $\mathbf{X}_2$. The likelihood at that point is $L(\text{UB}|\boldsymbol{X})$. Next, we find the point in the region $\theta_2 < LB$ that maximizes the likelihood of the student's response vector variable $\mathbf{X}_2$. The likelihood at that point is $L(\text{LB}|\boldsymbol{X})$. The likelihood ratio is defined as follows:

$$LR = \frac{L(\text{UB}|\boldsymbol{X})}{L(\text{LB}|\boldsymbol{X})} \quad (4)$$

We then select two cut-offs *A* and *B,* such that *0* < *A* < *B.* If LR < *A*, the person is classified as having an Elaboration Propensity Process. If LR > *B*, the person is classified as having a Paraphrase Propensity Process. We set *A* = 1/9 and *B* =9. This means that, if the likelihood of the response vector below the indifference region is at least nine times the likelihood above the indifference region, the person will be classified as having an Elaboration

propensity. If the likelihood above the indifference region is at least nine times the likelihood below the indifference, then the person is classified as having a Paraphrase propensity. . If *A* <*LR* < B, the student receives an Inconclusive classification for the process propensity. This system does not classify persons per se, but rather classifies their process propensity when making a mistake. As described above, the person's θ_2 is used to classify their process propensity.

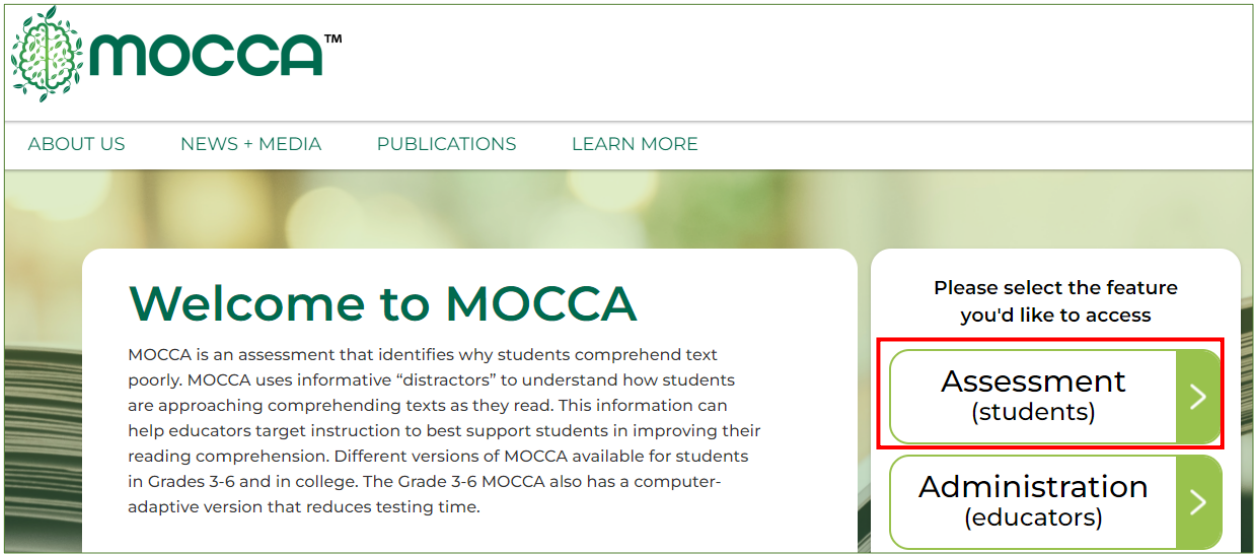
For people with estimated $\theta_1 < 0$, their process propensity received a classification of Paraphrase, Elaboration, or Inconclusive as described above. If $\theta_1 \geq 0$, the student received a classification of Causal. This was done for two reasons. First, students for whom $\theta_1 \geq 0$ committed few mistakes, not enough to classify their process propensity with confidence. Secondly, they were generally good readers who would not need additional support and therefore did not need diagnostic information to individualize extra support. To be considered as having taken MOCCA, a student had to complete at least 10 items in either the FIT or VCAT conditions.

Demonstration of the Student Experience

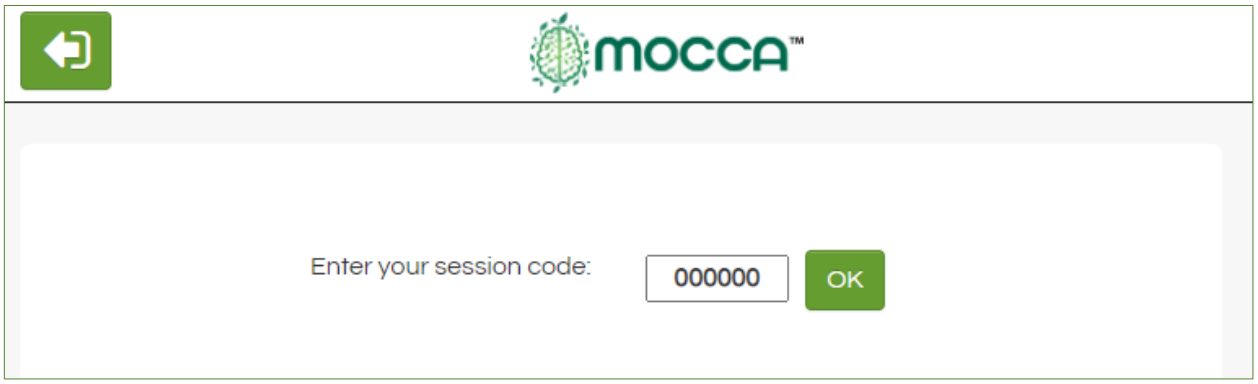
To see several sample items and experience MOCCA from the student perspective, follow the steps below.

1. Navigate to https://mocca.uoregon.edu

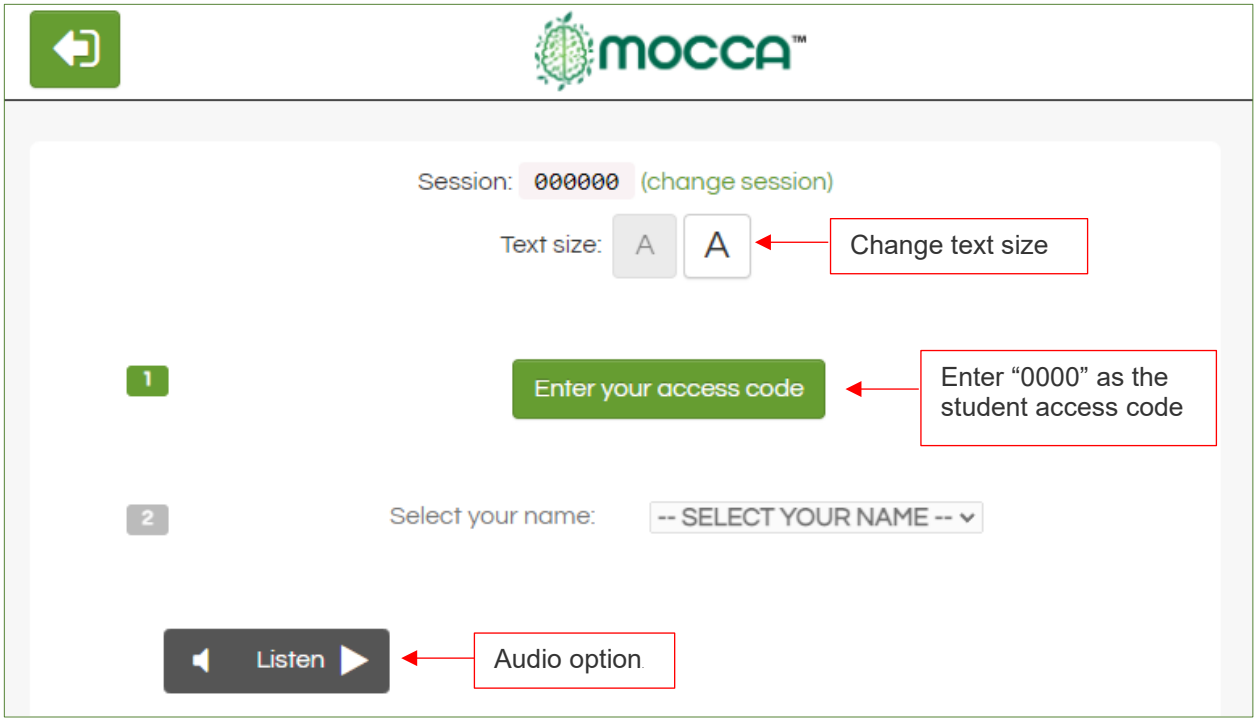
2. Click on "Assessment (students)".



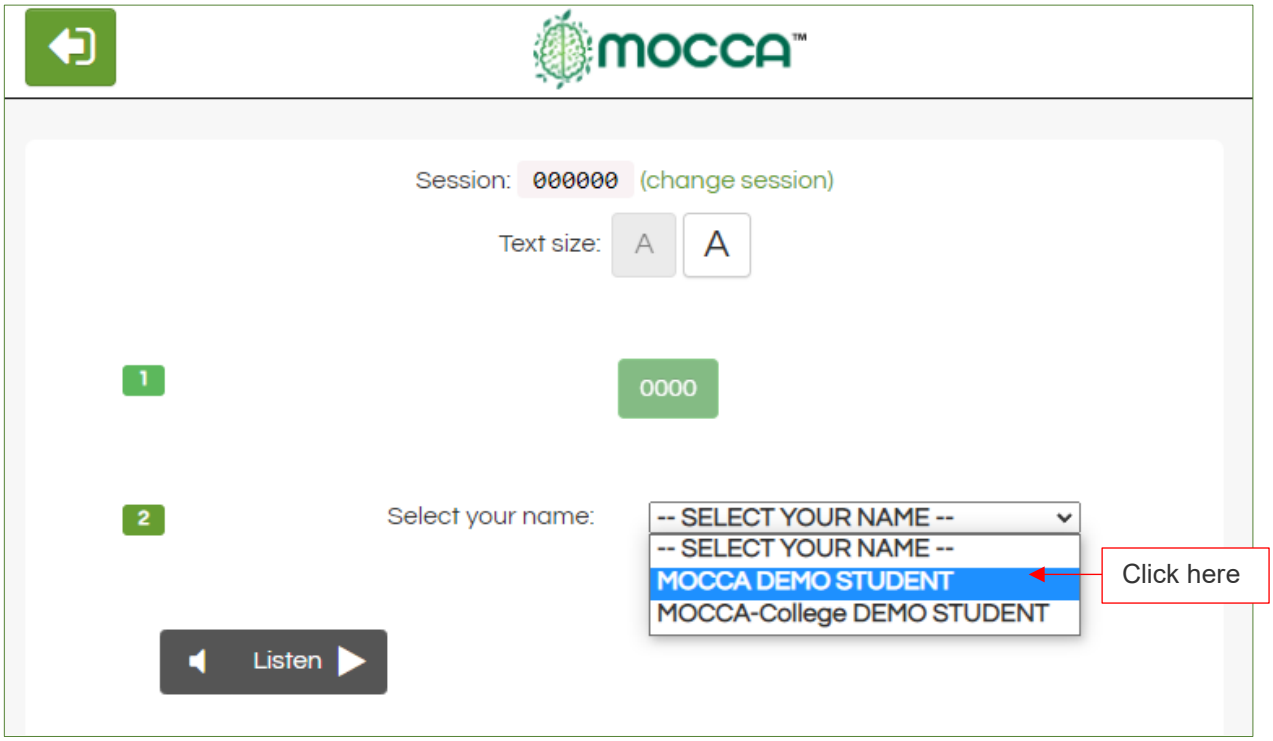
3. Enter "000000" as the session code.



4. Enter the 4-digit access code "0000".
   - The text size can be changed from big to small and vice versa.
   - There is an option to listen to the directions. After the directions and sample problems are complete, the option to listen is not available. We recommend that students use headsets to listen to the instructions.



5. After entering the access code, select "MOCCA DEMO STUDENT" in the "SELECT YOUR NAME" dropdown menu.



6. Now you can continue through the instructions and practice items, just as a real student would. Note that the wording for the instructions may have changed.

2. Administration

Intended Uses

MOCCA is currently intended as a screener for reading comprehension difficulties in Grades 3 to 6. In addition, MOCCA is intended to provide a diagnostic classification of readers who do not comprehend well based on their predominant approach to comprehension.

MOCCA does not provide diagnostic information about decoding or other "low-level" component reading skills. Nor is MOCCA designed for making high stakes decisions. As with any assessment, MOCCA scores are most meaningful and useful in decision making when used in combination with other data sources.

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 personnel 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

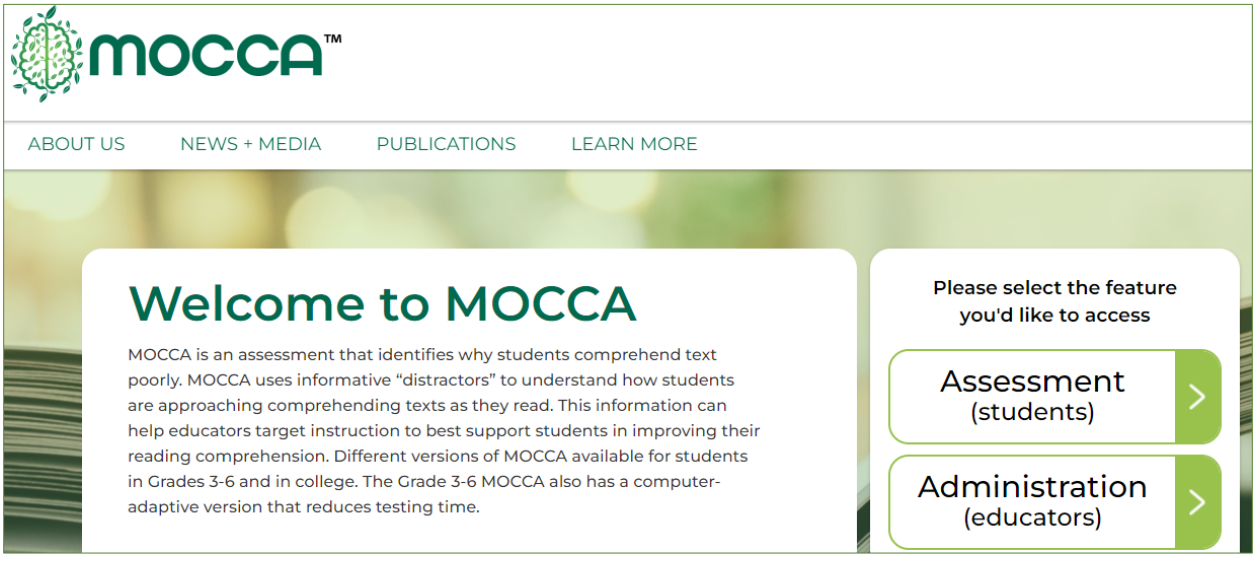
over the school network. Headphones are optional, though recommended for students to receive clear introductory instructions. You may also play the instructions over a speaker for all to hear. Regardless, test administrators must monitor that individual students are keeping up by clicking Next appropriately.

Getting Started

To administer MOCCA, users must have an account in the MOCCA system at mocca.uoregon.edu. To create a multi-user and/or multi-classroom account, contact mocca@uoregon.edu. Otherwise, you can create a free, single-user account by following the online instruction when they click Administration on mocca.uoregon.edu. Following are the steps required to administer MOCCA.

Logging in

1. Navigate to mocca.uoregon.edu.
2. Click "Administration" to login to your account.



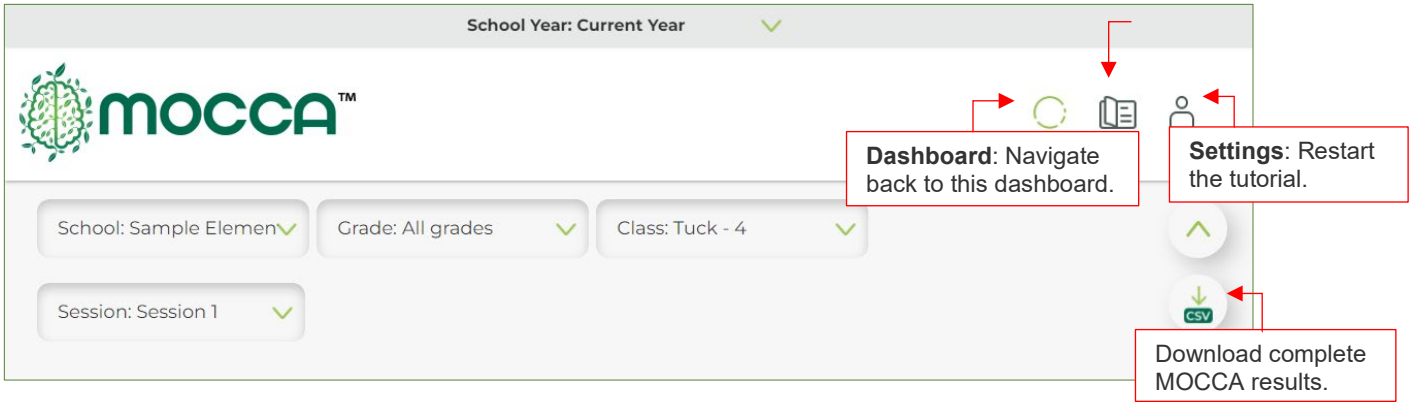
3. Login using your email and password.

Dashboard

After signing in at https://mocca.uoregon.edu, you will see a dashboard, which will initially be blank. Once students are rostered and begin testing, there will be three donut dials populated with information about the administration of MOCCA to students in your classroom, school or district.

At the top of the screen, you can select "School Year". By default, the current, or most recent school year will be displayed. Dropdown menus for school, grade, class, and session are available, depending on your account access level. For example, accounts with district-level access will have all menu selections available, including school, grade, class and session. School-level accounts will have access to all menus except school (i.e., school-level accounts will not be able to access information about other schools in the same district.) Class-level or teacher accounts will have access only to their own class. Teachers with access to more than one class are considered school-level users.

Resources: Navigate to resources for interpreting MOCCA.

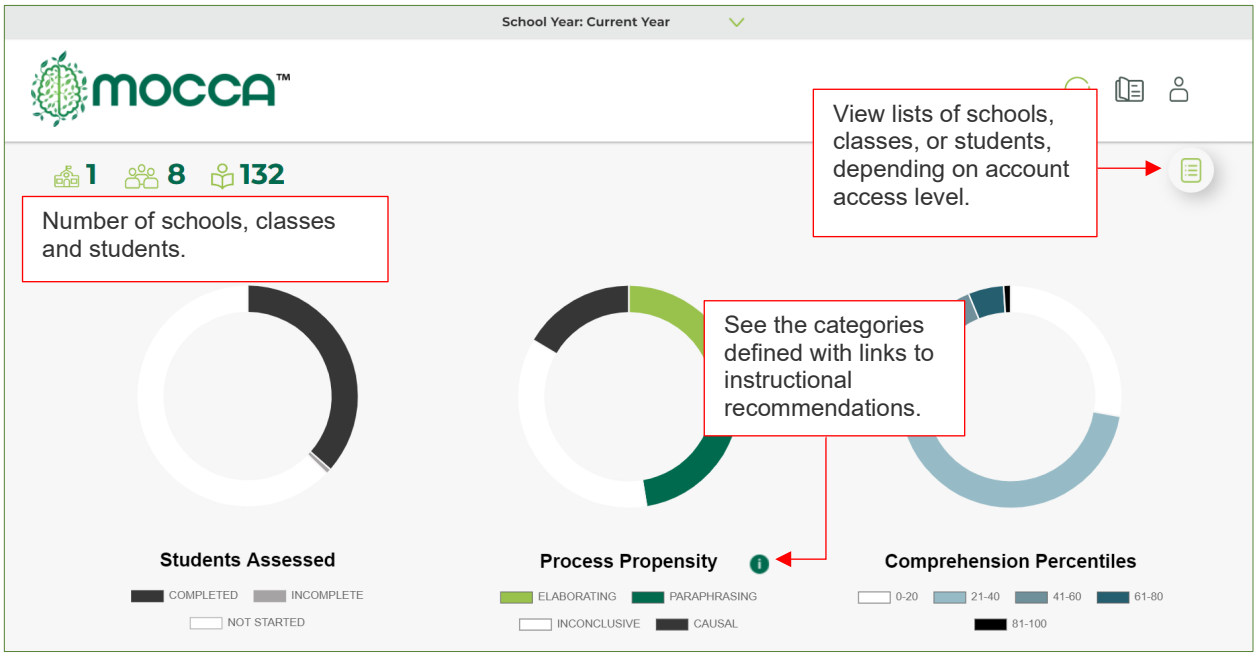


Features of the dashboard, depending on account-access level, include a display of the number of:

- Schools, classes and students (depending on your account access level)
- Students rostered, started and completed MOCCA
- Students in each "Process Propensity" category
- Students in each "Comprehension Percentiles" range.

District-level users will see all of these numbers. School level users will not see school counts.

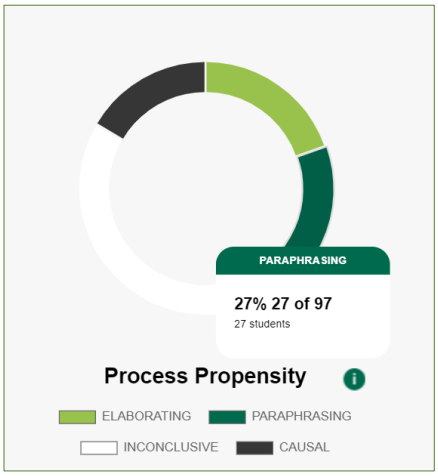
Classroom-level users, which includes most teachers, will not see school or class counts.



- View the School List, Class List, or Student List by clicking on the notebook icon, depending on account access level. (District-level users are taken to the School List;

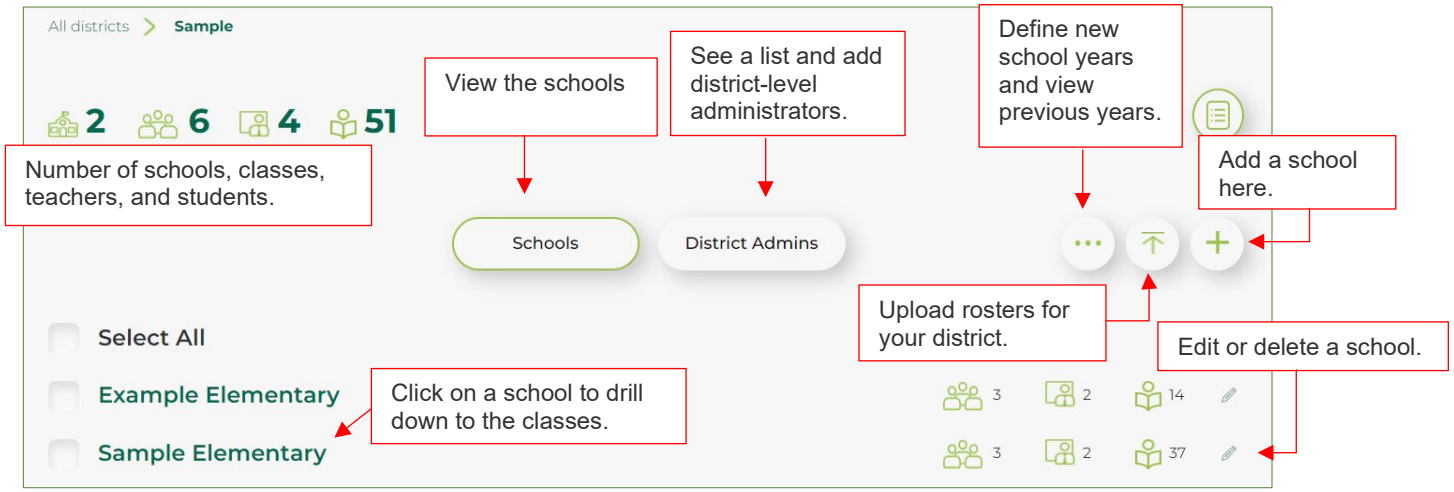
school-level users are taken to the Class List; and class-level or teacher users are taken to the Student List.)

- Hovering the cursor over a donut chart will display the number or percentage of students represented in that section of the chart.



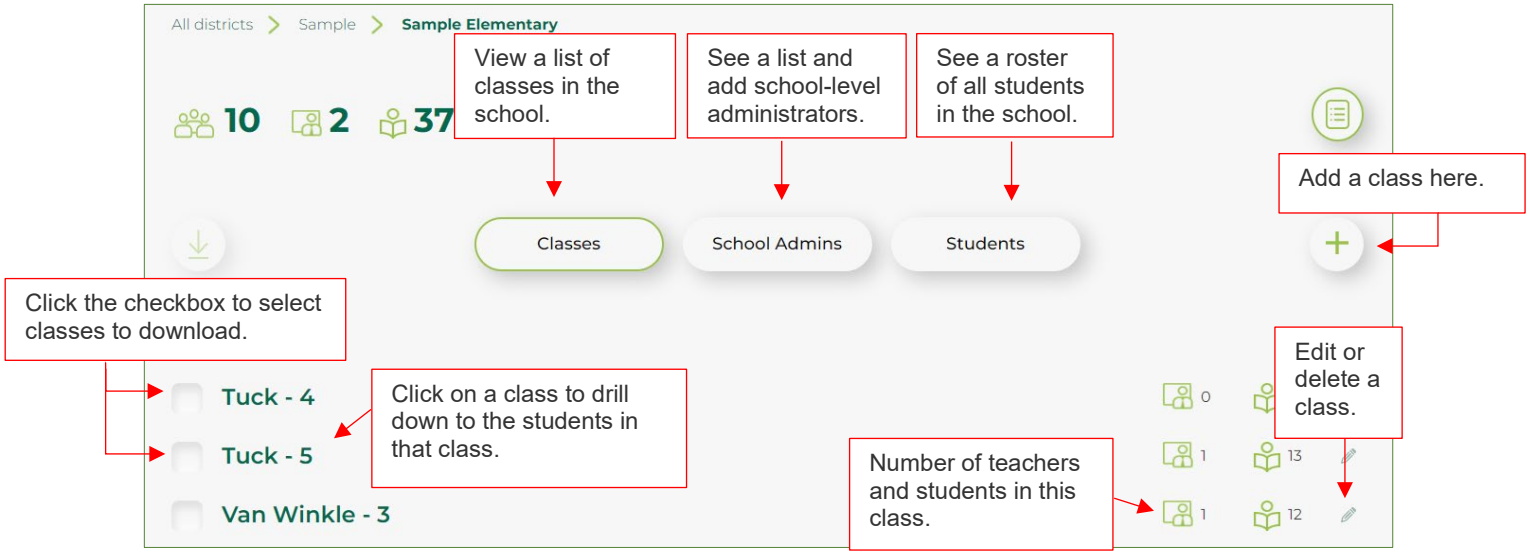
School List View (for district-level users only)

- Access the list of schools and district administrators in your district
- View the number of schools, classes, teachers, and students
- Define new school years and view previous years
- View, edit, and add schools
- For help uploading rosters, contact mocca@uoregon.edu



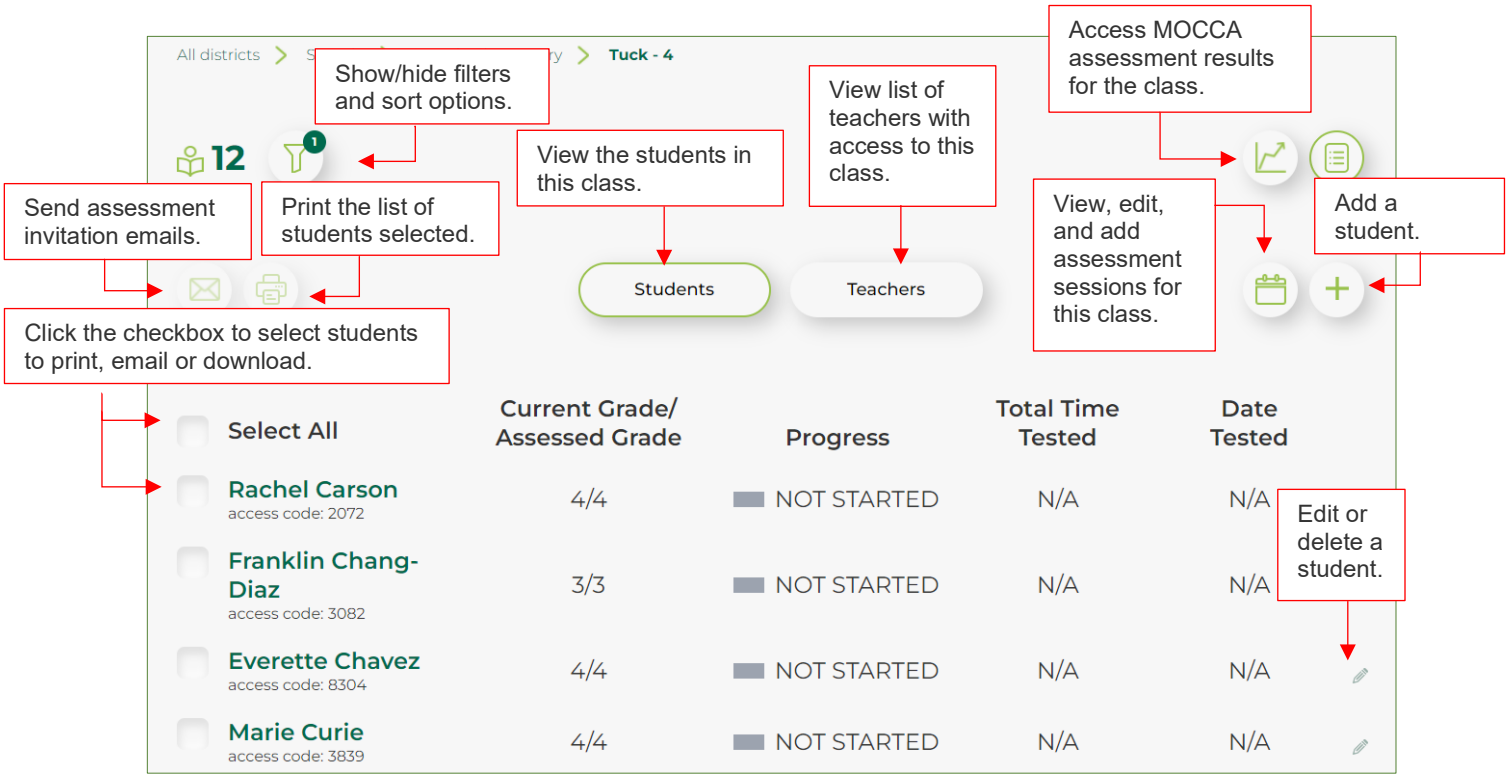
Class List View (for school- and district-level users only)

- Access the list of classes in a school and school-level administrators
- View number of classes, teachers, and students
- View, edit, and add classes
- Select and download a list of classes



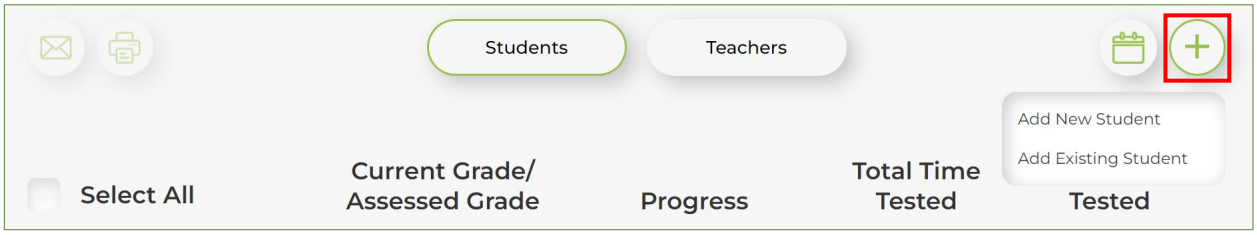
Student List View (for class-, school- and district-level users.)

- Access the student roster and students' access codes for a class
- View number of students in a class using filter and sort options
- See students' enrolled and assessed grades, assessment progress, total time tested, and test dates
- Send assessment invitations via email to students
- Print or download a list of selected students
- Access MOCCA assessment results for a class
- View, edit, and add assessment sessions, students, and teachers.



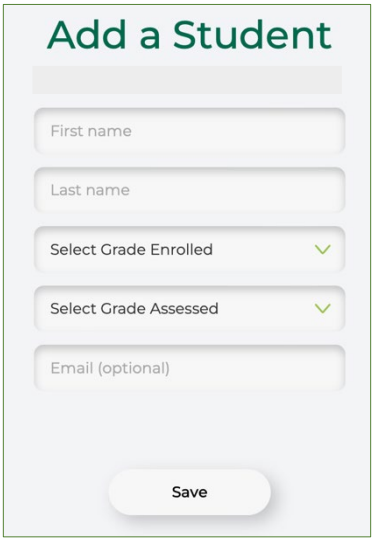
Add Individual Students to a Class

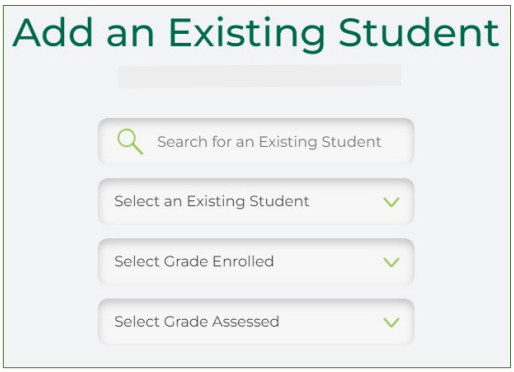
- To add individual students to a class, in the Student List View, click the "Students" button, then click the "+" button.
- Select if you want to add a new or existing student.



- To add a new student, fill in the student's first name, last name, email (optional), enrolled grade and assessment grade. Then click "Save".

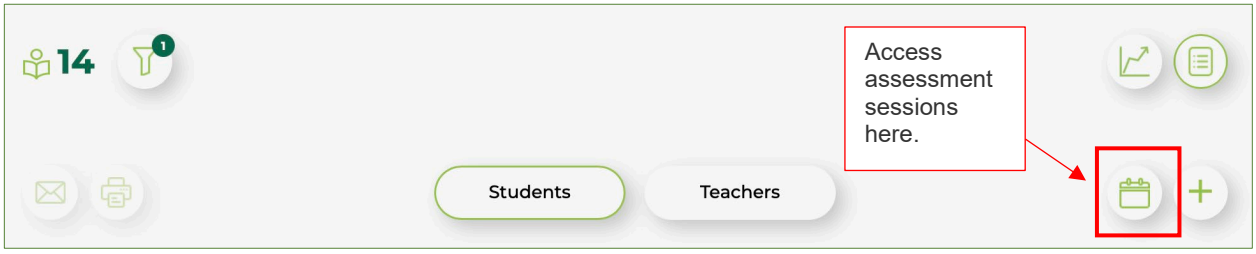
- To add an existing student (i.e., a student already in the school and enrolled in another class), search for the student or use the dropdown menus.



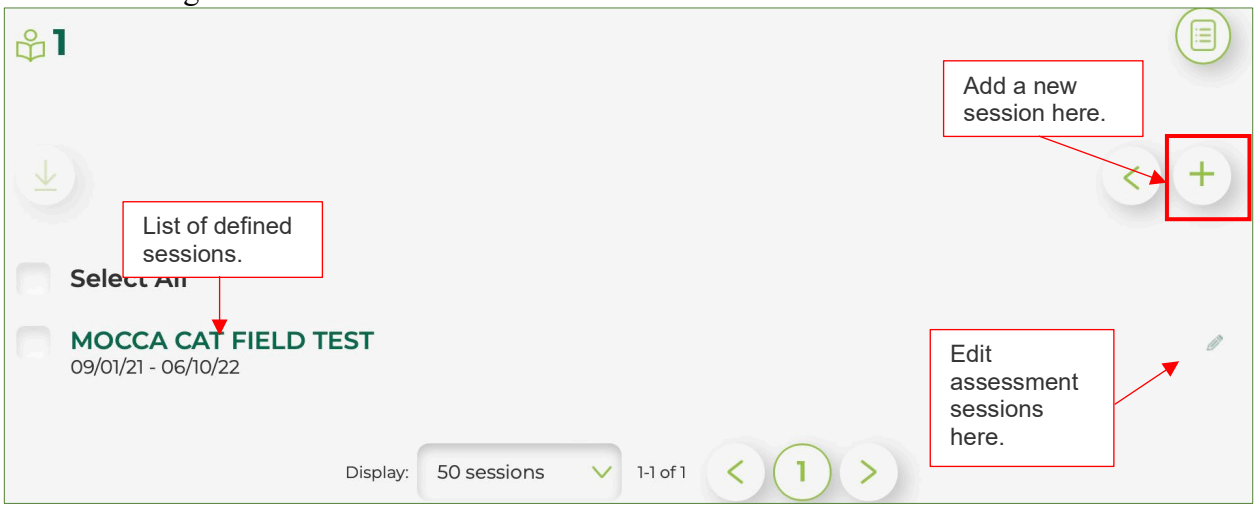


Create an Assessment Session

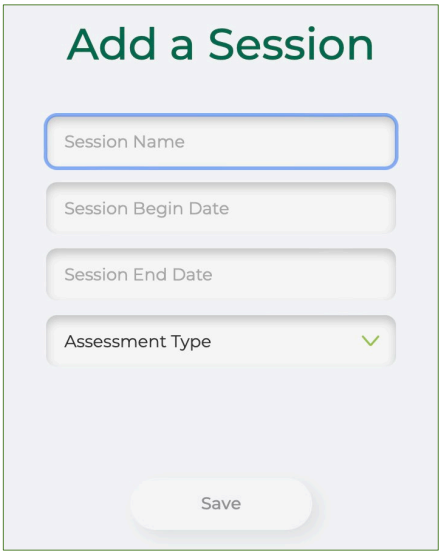
- To create an assessment session, within the Student List View, click the calendar icon.



- On the next screen, click the "+" sign to add a session. You can also choose to edit existing sessions from this screen.



- Name the session something meaningful to you and set the beginning and end dates for the session. Choose the assessment type listed. You should see only one listed. Then click "Save".



Administering MOCCA

An assessment session must be defined before MOCCA can be administered, and the date that MOCCA is administered must fall within the date range of the defined session.

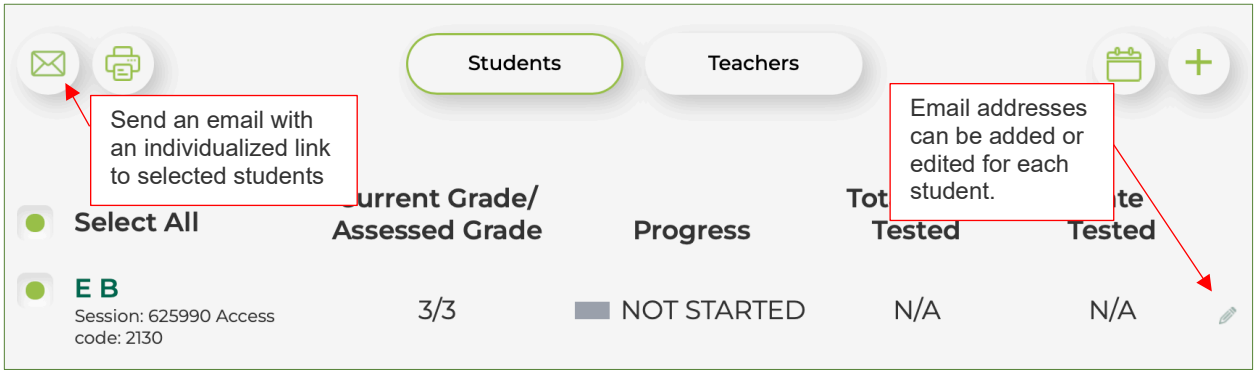
There are two methods for administering MOCCA. Teachers can either send emails to students rostered in a class, or they can distribute access codes to their students.

Administer MOCCA by Sending an Email Link

From the Student List view, do the following.

- Select the students who will take MOCCA.
- Click the envelope icon to send each of those students an email with an individualized link.
- Students will click on the link in the email to start MOCCA.
- The link is personalized and only requires a student to confirm their name before beginning the assessment.

Note: Email addresses must be entered for each student who will access MOCCA via email. Emails can be entered when uploading rosters, when manually adding students, or when editing student information from the Student List View.

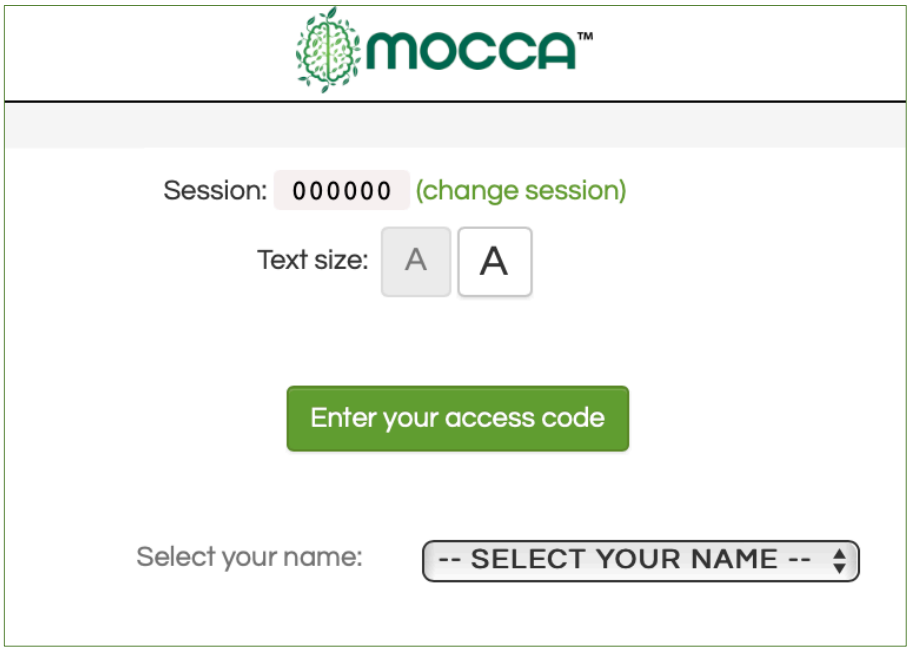


Administer MOCCA Using Access Codes

From the Student List view, do the following.

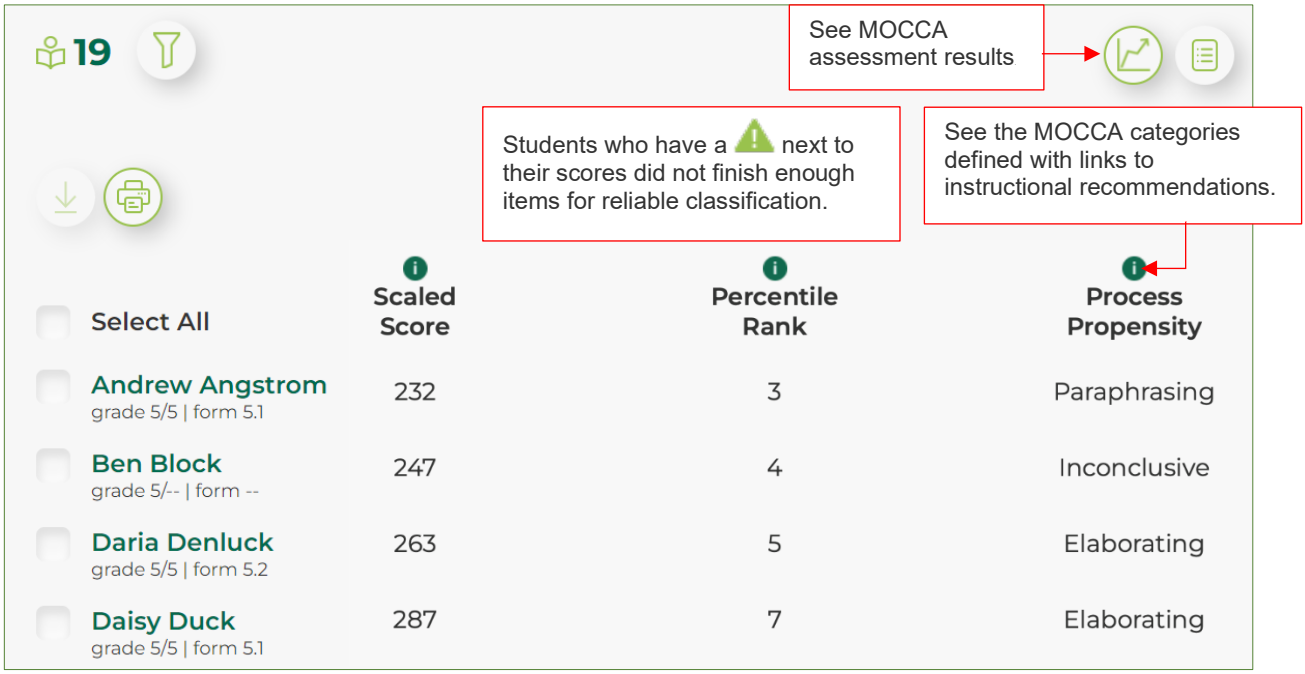
- Select the students who will take MOCCA.
- Click the printer icon to print the individualized student access codes.
- Cut the printed paper in strips, one for each student, and distribute to students.
- Ask students to navigate to mocca.link
- Direct students to:

- Enter their six-digit session code and click “OK”.
- Click on “Enter your access code” and then enter their four-digit access code.
- Select their name from a dropdown list. The selected name must match their individualized code to gain access to the assessment.



Accessing Assessment Results

- Access assessment results from within the Student List View by clicking on the graph icon.
- View number of students in the class based on filters and sort options.
- Print or download the results for the selected students.
- See students' scaled scores, percentiles, and Comprehension Process Propensities.
- Comprehension Process Propensity is a classification unique to MOCCA that allows you to tailor comprehension instruction to a student's needs. There are four process propensities: Elaborating, Paraphrasing, Inconclusive, and Causal. Information about the Process Propensities is available in the MOCCA Interpretive Guide: https://blogs.uoregon.edu/mocca/mocca-comprehender-types/



3. Interpretation

MOCCA identifies students who are and are not comprehending well. When students are not comprehending well, MOCCA offers diagnostic information that can inform instruction. MOCCA categorizes readers into four types. This guide helps you understand the scores that MOCCA provides, as well as their implications for instruction.

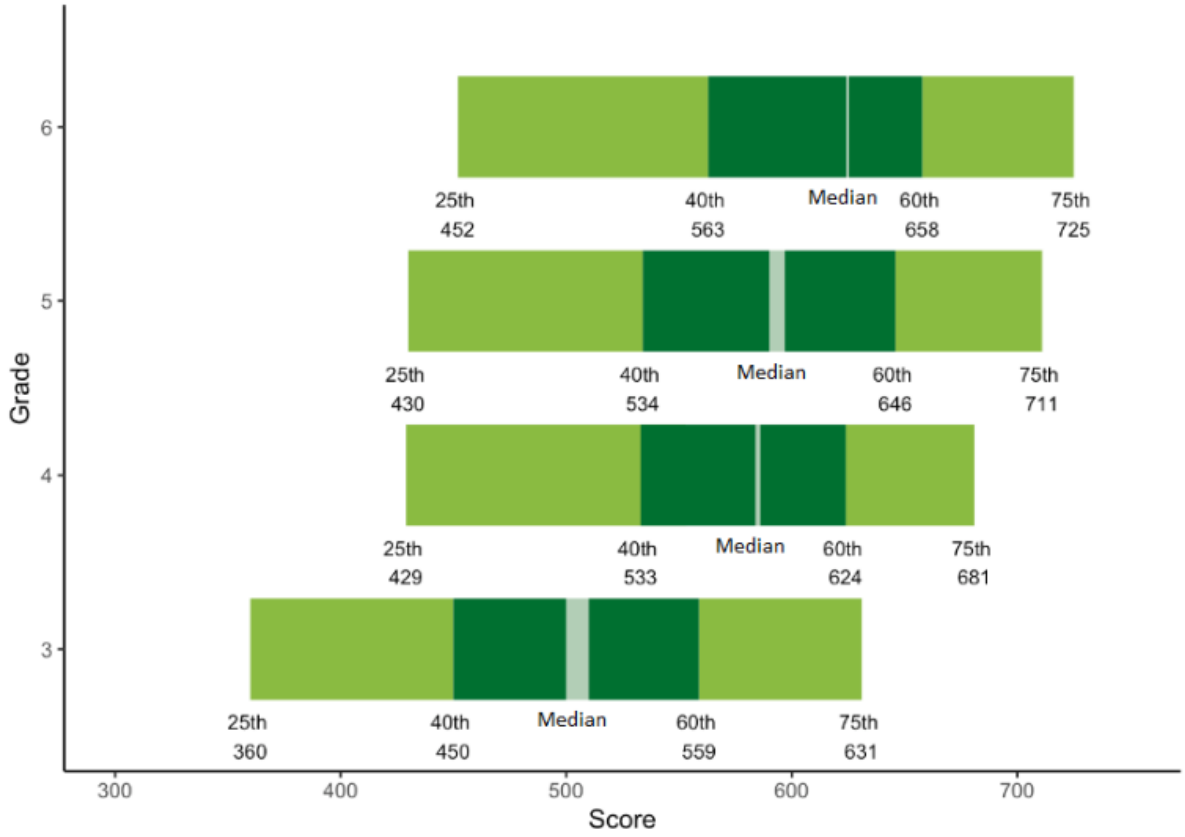
MOCCA Scaled Scores

The MOCCA Scale Score reports a student's reading comprehension performance on a scale from 50 to 950, where 500 is the average score and 150 is the standard deviation. The MOCCA Scaled Score can be used in Grades 3 to 6 to track student improvement in reading comprehension from the beginning of Grade 3 through the end of Grade 6. Students who are making progress in reading comprehension should have scores that increase across Grades 3 to 6 because the range of typical performance increases across these grades.

In the figure below, you can see how the range of typical scores changes from grade to grade. The median, or average, performance is depicted with a very light green band, while the scores for the next 10% of students in either direction is in dark green and an additional 15% are in light green. The full length of each bar represents how the middle 50% of students perform at a given grade level. As the figure makes clear, scores shift up on the MOCCA scale across Grades 3 to 6. The shift is larger between third and fourth grade, than across the subsequent grades.

Figure 3.1

MOCCA Scaled Score Ranges by Grade



Percentile Ranks

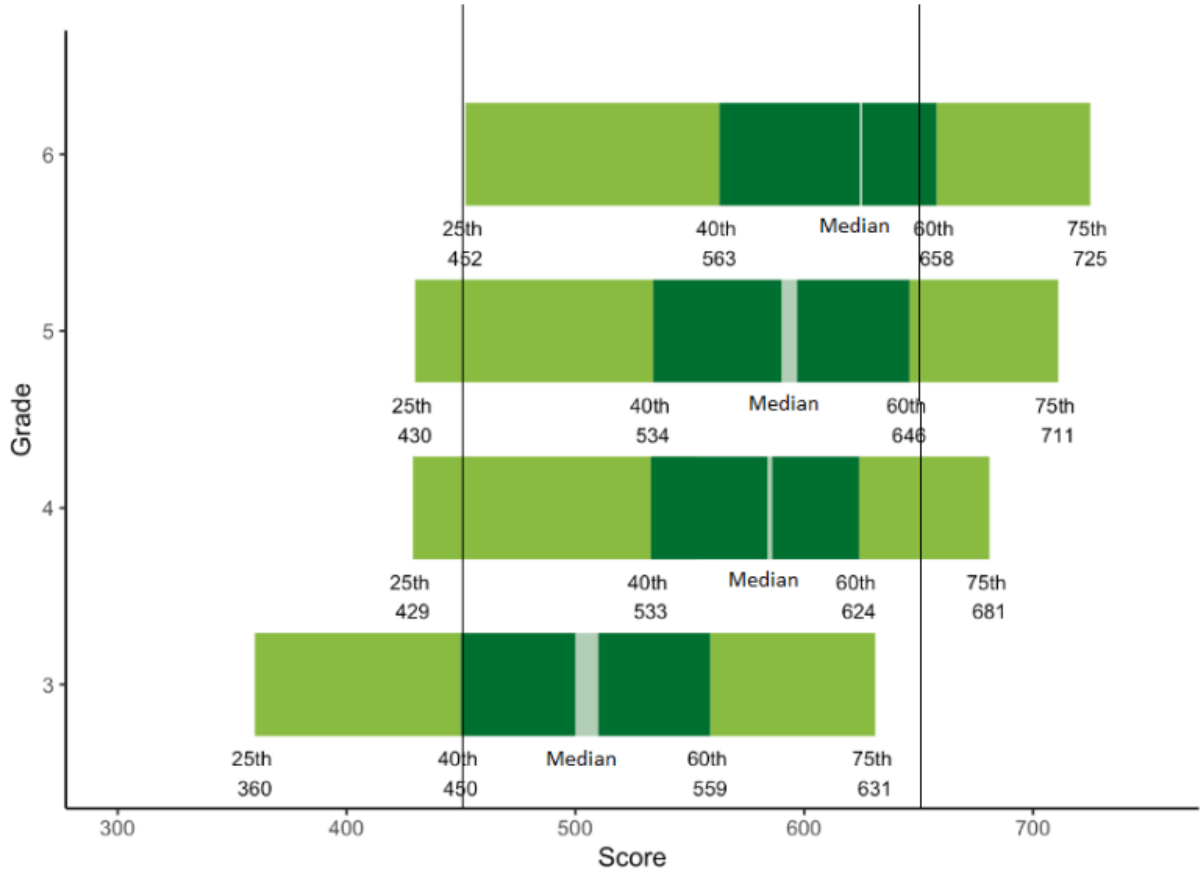
Percentile ranks convey the percentage of students that a student performed as well as, or better than. This score can be used to understand how a student is performing relative to similar students in the same grade. A percentile rank of 50 means a student scored as well as or better than 50% of students at their grade level (see figure below).

Because MOCCA scaled scores are on a continuous scale across Grades 3 to 6 and percentile ranks are based on grade level, the same scaled score will result in different percentile ranks depending on a student's grade (i.e., the grade level to which they are compared). For example, a scaled score of 650 is the 81st percentile rank in Grade 3, the 68th in Grade 4, the 61st in Grade 5, and the 58th in Grade 6. Similarly, a scaled score of 450 is the 40th percentile

rank in Grade 3, the 27th in Grades 4, the 26th in Grade 5, and the 24th in Grade 6. This effect is also visible in the figure below.

Figure 3.2

Percentiles by Grade for MOCCA Scaled Scores of 450 and 650



Percentile ranks are not always as meaningful as they appear. Teachers should pay particular attention to the grade level of the form of MOCCA a specific student has taken and whether a student has finished the full assessment.

Percentiles are only meaningful when students take MOCCA for the same grade level in which they are enrolled. If a student takes MOCCA off-grade, the percentile rank compares that student to students in their grade who took a *same-grade* form, which is not a valid comparison.

If a student does not complete MOCCA, percentile ranks need to be interpreted with great caution. The fewer items a student has completed, the less reliable their score. You can confirm that a student completed the full assessment by checking their status on the MOCCA roster screen, where their status will read "Not started," "Incomplete," or "Complete." *Unless a student's status reads Complete, the percentile rank should be interpreted with caution.*

MOCCA Comprehender Types

To understand MOCCA Comprehender Classifications, you need to understand causal inferences. You also need to understand the research behind MOCCA.

What is a causal inference?

Causal inferences are inferences that are required for a text to make sense. They are sometimes called *necessary* inferences. The following very short story serves as a great example.

> Tyrese decided to bake a pumpkin pie for dessert. He looked in the pantry. Tyrese was disappointed.

In this story, one can infer that Tyrese did not find all the needed ingredients in the pantry. Most proficient readers will immediately and effortlessly infer that. In fact, they also infer that Tyrese must think the ingredients will be in the pantry.

Proficient readers may be entirely unaware that they have made those two *necessary* inferences. What makes them necessary is that the three sentences become *non sequiturs* without those two inferences.

There are other "unnecessary" inferences that readers can make as well. One might infer what Tyrese's gender, what Tyrese looked like, and even what specific ingredients were missing or the type of dish the pie might be baked in. While all of these inferences elaborate and enrich,

or flesh out, the story, making it more interesting and complete, they are not necessary to understanding the central core of the text. Thus, they would not be considered *causal inferences*.

Where did the comprehender types come from?

The MOCCA comprehender types are based on reading research conducted with a method called a "think aloud." Think alouds have been used in instruction for many years now, but before that, they were primarily a means of understanding what readers were thinking while they read. The method has also been used in many kinds of psychology studies, in marketing research, and many other kinds of studies. In reading, researchers ask readers to stop now and then — for example, after every sentence, paragraph, or page — while reading to say what they are thinking aloud.

Think aloud research in reading has taught us much of what we know about reading comprehension. It is how we first identified the cognitive strategies that proficient readers use to understand what they read and the source of the comprehension strategies we teach our students. Although the list of strategies we teach may vary based on the curriculum we use, two of the most common strategies are paraphrasing and making inferences, especially elaborative inferences. It turns out that readers who are struggling specifically with comprehension (but not word reading) tend to rely on one or the other of these strategies *a lot*. MOCCA helps to identify whether a reader is "stuck" on using one of these two strategies, which gives teachers the information they need to get them "unstuck" and making progress in comprehension.

Causal Comprehenders

Causal Comprehenders regularly make the causal inferences necessary for comprehension. Causal Comprehenders do not require help making these inferences.

They will benefit from continuing to read texts they enjoy and find challenging. Like all readers, they also benefit from talking to others and writing about what they have read. These practices will help them maintain and grow their comprehension of and engagement with texts.

Paraphrasing Comprehenders

Paraphrasing Comprehenders do not regularly make the causal inferences necessary for comprehension. Instead, Paraphrasing Comprehenders tend to paraphrase or repeat what they have read, sticking to a relatively literal interpretation of texts.

To improve their comprehension, these readers will benefit from being prompted to make various kinds of inferences. These inferences can connect ideas from different places in a text (often called bridging, connecting, or text-to-text inferences). They can also use a reader's background knowledge (often called elaborative, text-to-self, or gap-filling inferences). All inferences require a reader to supply information that is not explicitly stated in a text, which is what helps deepen comprehension.

We recommend explicit instruction in making inferences, where the teacher models inferences and engages students in group and individual practice at making inferences. Here is an example of defining and modeling a bridging (also known as connecting or text-to-text) inference:

> We know [state an event or fact from the text] happened, and we know [state another event or fact from the text] happened. When we *connect* these ideas, we can *infer* [a detail that is not explicitly statement].

You can also use questioning to prompt Paraphrasing Comprehenders to make inferences when reading. These questions work in whole class, small group, and individual settings. Here are two questions that specifically prompt connecting inferences:

> How does what we just read connect to what we read earlier?
> How does what we just read connect to [state an earlier important idea from the text]?

Here are two questions that specifically prompt elaborative inferences (including a predictive inference):

> What do we already know about [state main topic from the text]? How does that connect to what we just read?
> What do you think will happen next? What clues in the text make you think that?

Elaborating Comprehenders

Elaborating Comprehenders do not regularly make the causal inferences necessary for comprehension. Elaborating Comprehenders do make inferences about information not explicitly stated in a text, but that information is not strictly needed to comprehend a text. These inferences are nice to make and may enrich the reading experience, but they do not provide information that is necessary for comprehension.

To improve their comprehension, these readers will benefit from being prompted to make causal (i.e., necessary) inferences. These inferences use a reader's background knowledge to fill in *important* information that is not explicitly stated in a text. Causal inferences tend to focus on *why* events occur as they do or characters behave as they do.

We recommend explicit instruction in making causal inferences, where the teacher models inferences and engages students in group and individual practice at making causal inferences. Here is an example of defining and modeling a causal inference:

> We know [state a goal or problem from the text], and we know [state an event or behavior from the text] happened. When

> we *connect* these ideas, we can *infer* that [the event/behavior] happened *because* [restate the goal or problem].

You can also use questioning to prompt Elaborating Comprehenders to make causal inferences when reading. These questions work in whole class, small group, and individual settings. Here are some questions that specifically prompt causal inferences:

> Why do you think that just happened? What clues in the text make you think that?
> Why do you think that character did that? What clues in the text make you think that?
> Why do you think that character said that? What clues in the text make you think that?
> Why do you think that character wants that? What clues in the text make you think that?

Inconclusive Comprehenders

Inconclusive Comprehenders do not regularly make the causal inferences necessary for comprehension. Inconclusive Comprehenders do not demonstrate a reliable pattern in what they do when they do not make causal inferences. There are a number of reasons a reader may end up classified as inconclusive.

Some Inconclusive Comprehenders may be struggling with lower level reading skills, such as reading words and fluency. If you have data from other assessments that suggests a reader is struggling with lower level skills, we recommend focusing intervention on those skills. In the meantime, it may help to offer the student easier to read texts that allow them to practice and enjoy reading for comprehension. Audio books at grade level can also be an important way to give the student access to rich, grade-level content while they work on their skills.

Inconclusive Comprehenders who demonstrate strong lower-level reading skills on other assessments may be struggling with literal comprehension, especially if other assessments also

indicate a comprehension problem. In some cases, the student may not be prioritizing reading for meaning, which can occur when instruction is overly focused on lower-level reading skills. Difficulties with literal comprehension are best addressed through engaging students in active reading practices. Having the student read aloud and stop to discuss what they are reading. Occasionally asking literal (i.e., Who? What? When? Where?) and inferential questions about a text will help the reader focus on meaning while reading. See the suggestions under Paraphrasing and Elaborating Readers for information on how to encourage inference making.

Other Inconclusive Comprehenders may be guessing when they take MOCCA. If you feel a student may have been guessing or otherwise disengaged while taking MOCCA, you may wish to reassess after encouraging the student to take their time and do their best work.

Otherwise, these readers will benefit from continuing to read texts they enjoy and find challenging. Like all readers, they also benefit from talking to others and writing about what they have read. These practices will help them maintain and grow their comprehension of and engagement with texts.

4. Development and Validation Process

The results reported in this manual are the product of a multi-year program of research. Development activities included four main objectives.

First, we aimed to experiment with item specifications to expand the range of difficulty of MOCCA items. Specifically, we allowed for MOCCA narratives to have anywhere from 5 to 10 sentences, as opposed to the original 7 sentences. We also wrote two versions of a subsample of items to have 3 or 5 answer choices. We anticipated that shorter items would be easier, longer items more difficult, and items with more answer choices would also be more difficult. Results are reported in the New Items Pilot chapter in this manual.

Second, we aimed to increase the overall size of the MOCCA item pool to accommodate more frequent testing. As a result, after testing how manipulating item specifications might affect item difficulties, we performed a larger calibration study to more robustly estimate item difficulties and bring all new items onto the same scale as the existing MOCCA item bank. Results are reported in the Item Calibration Study chapter in this manual.

Third, we aimed to make MOCCA a computerized adaptive test (CAT). To support this aim, we conducted simulation studies and two pilots of the resulting CAT algorithms. These results are reported in a separate technical manual (Davison et al., 2021). The resulting CAT was validated by (a) comparing it to the more traditional fixed-item format, in which the same items are encountered in the same order for students taking the same form, and (b) using MOCCA scores to predict performance on a range of criterion measures of reading. These results are reported in the Comparison and Validation Study chapter in this manual. Normative information for the CAT of MOCCA is reported in the MOCCA CAT Norms chapter.

Fourth, we set out to test the sensitivity of MOCCA to change in student comprehension. We applied both a longitudinal growth modeling approach and an adaptive measurement of change (AMC; Kim-Kang & Weiss, 2008; Finkelman et al., 2010; Lee, 2015; Phadke et al., 2015) approach to interpreting change in students' scores. Results are reported in the Measurement of Change chapter.

5. New Items Pilot

New test items were developed from Summer 2019 to Spring 2020 by published children's authors. Items were revised by the research team to ensure alignment to the MOCCA guidelines from Spring 2020 to Fall 2020. These items were piloted with students in November and December of 2020.

Test Bank

Six authors were recruited to write stories for MOCCA. Authors were instructed to create short stories, seven to ten sentences long, that require a causal chain of events with a goal that is central to the story's development. Authors input their stories into Excel sheets, with fixed sentence lengths to control reading time as much as possible. 380 stories were written, with no author writing over 100. One of the principal investigators wrote an additional 50 MOCCA stories, and stories that did not make it into the 2019 version of MOCCA were also revised. Authors submitted complete stories to the research team, and the principal investigator and a research assistant edited each story to ensure they aligned with MOCCA guidelines. Some stories were removed if they could not be edited to align with these guidelines.

The stories ranged in length from 5 to 10 sentences. The distribution of sentences per item was not experimentally manipulated for a few reasons. First, items from the 2019 version of MOCCA were all seven-sentences long and 13 of these items were anchor items for the scaling of the pilot data. Second, it would be impossible to create parallel stories with different numbers of sentences that also met item specifications in terms of sentence length and equivalent readability. Third, many items originated in prior MOCCA development that did not make it to the 2019 version were also included in this study, resulting in more seven-sentence stories than other length stories.

Within each story, one sentence was targeted to be removed to create the MOCCA test items. The missing sentence removed was always the second to the last sentence. This sentence was integral to the causal coherence of the story. This sentence, the correct answer choice, was used to format the incorrect answer responses. The correct answer was the only item response that created a causally coherent story.

Incorrect answer choices included two different response types: an elaborative statement or a paraphrase statement. For new MOCCA stories, there were two responses per type; all these incorrect answer choices were developed by the principal investigator and a research assistant. For previous MOCCA stories, additional elaborative and paraphrase statements were created. Elaborative statements provided extra details or associations regarding information presented in the two sentences prior to the missing sentence. These statements did not relate to the goal of the story. Paraphrase statements summarized the main idea or goal of the story, including updated goals as applicable. For incorrect answer choices, the length of the sentence reflected the length of the causally coherent sentence, and the use of names or pronouns was also aligned to the correct answer choice.

Within the MOCCA assessment, a random selection of items included both three- and five-answer choices across all forms. These items were used to evaluate if the number of response options impacted the students' score on the MOCCA assessment.

To ensure the readability of the story was maintained regardless of answer choice, the stories' readability statistics were evaluated using Word for Macintosh. Each response sentence was placed in the missing sentence location in the story and then readability was assessed. If the difference in readability was greater than 0.5 between the story with the correct response and an incorrect response, that response sentence was edited to align readability with the correct answer

choice. For complete MOCCA stories (i.e., stories with their causally coherent sentence included), readability ranged from 0.2 to 7.6, with an average of 4.4.

Methods

Sample

A total of 210 third grade, 215 fourth grade, and 231 fifth grade students participated in the pilot of new MOCCA items. Students attended four schools in three states representing the Pacific Northwest, Mountain West and South Atlantic regions. Across grades, 52% were female and 48% were male. 65.1% of students were white, 12.0% were Hispanic, 8.4% were Black or African American, 4.7% were Asian, and 0.2% were American Indian or Alaskan Native. Approximately 7% of students were identified as English language learners or former English learners, although this information was not available for 30% of the sample. 2.6% of the sample was eligible for Special Education. 17.2% of the sample was eligible for Free or Reduced Price lunch program, although this information was not available for 7.3% of the sample. Students within a grade were randomly assigned to one grade-level form of the MOCCA assessment.

MOCCA Assessment

A total of 40 items were included on each form. Items were assigned to the form based on readability, gender of the main character, and ending (i.e., positive, negative, neutral). The principal investigator assigned items to forms and the research team confirmed equivalency of forms within the same grade level.

A total of 10 different forms were created. Third grade students were exposed to five forms, fourth grade students to all ten forms, and fifth grade students to the five forms third grade students did not see. Thus, the ten forms for fourth grade overlapped with the forms for third and fifth grade.

Across all ten forms, 252 items were piloted. Of these, 13 were items from the 2019 version of MOCCA and had item parameters and thus by appearing on all forms could serve as anchor items for the scaling of the 239 new items. Of the 239 new items, 25 items had two versions, one of which had 3 answer choices and the other 5 choices, which appeared on multiple forms so that the effect of number of answer choices could be examined.

Analyses

Item response theory item parameter scores for the reading comprehension dimension and process propensity dimension were estimated based on the data from the pilot. This included item discrimination and difficulty parameters for both dimensions. Each 5- alternative item had one Causal Coherent (CC) alternative, two Paraphrase alternatives, and two Elaboration alternatives. For purposes of measuring the RC Dimension, each item (3- and -5-alternative) was scored correct and incorrect: 1 = Causal Coherent response, 0 otherwise. For purposes of scaling the PP Dimension, each item (both 3- and 5-alternative) was scored as 1 = a Paraphrase alternative, 0 = an Elaboration response, and missing = Causal Coherent. In scoring the 5-alternative items for the PP Dimension, no distinction was made between the two Paraphrases responses nor between the two Elaboration responses. A Paraphrase response was coded 1 no matter which Paraphrase response they chose. Likewise, an Elaboration response was coded as 0 irrespective of which Elaboration response they chose. In addition, a series of analyses of variance (ANOVAs) were used to evaluate if the (a) number of response choices provided, (b) number of sentences within a story, and (c) readability of a story impacted students' theta estimates.

Results

Testing Administration

Participants had an unlimited amount of time to complete the test. An approximate average of 65 students took each form. Across all forms, an average of 37 items were completed by participants (average range 34-40). On average, 84% of participants completed their test across all forms (range 68-100%).

Guessing Parameter Estimation

We found little difference in fit between a 3PL model with all lower asymptote parameters constrained equal and a 3PL model in which all lower asymptote parameters allowed to vary. In the constrained model, the common lower asymptote parameter equaled .24.

Number of Answer Choices

One-way, within-subjects ANOVA were used to evaluate the impact of the number of response 25 items were given. Overall, item parameters for the three- and five-choice conditions appeared similar, but ANOVAs for item discrimination, difficulty, and guessing parameters were all statistically significant. The number of incorrect choices accounted for approximately 16% of the variance in the item discrimination parameter, $F(1, 24) = 4.46$, $p = .045$, $\eta^2 = .156$; about 20% of the variance in the item difficulty parameter, $F(1, 24) = 6.2$, $p = .02$, $\eta^2 = .206$); and approximately 49% of the variance in the guessing parameter, $F(1, 24) = 22.7$, $p < .01$, $\eta^2 = .486$). Items with five answer choices had higher item discrimination and item difficulty parameters, but lower guessing parameters compared to items with three answer choices. This demonstrates the addition of answer choices increases an item's discrimination and difficulty and reduces the likelihood that a student will select the correct answer by chance.

Number of Sentences within Stories

A one-way, between-subjects ANOVA was run to evaluate the impact of the item length, or number of sentences within each MOCCA story. Item discrimination scores ranged from 1.93

to 2.03 and the effect of the number of sentences per item was not significant, $F(5, 271) = 1.41$, $p = .22$, $\eta^2 = .025$. For the guessing parameter, scores ranged from .22 to .29 and the effect of the number of sentences per item was also not significant, $F(5, 271) = 1.38$, $p = .23$, $\eta^2 = .025$. Finally the item difficulty ranged from -.22 to -.01 and the overall effect of the number of sentences per item was significant and explained a small amount of variance, $F(5, 271) = 2.4$, $p = .038$, $\eta^2 = .042$. After utilizing the Benjamini and Hochberg (1995) correction procedure, none of the pairwise comparisons were statistically significant. These results indicate the number of sentences within an item does not significantly contribute to item discrimination, item difficulty, or guessing.

Readability and Theta Scores

We evaluated the correlations between the item parameter estimates, including reading comprehension discrimination, difficulty, and guessing as well as process propensity discrimination and difficulty and readability as measured by Flesch-Kincaid using Word for Macintosh. All correlations were below .15 and there was only one significant correlation, between reading comprehension difficulty and readability ($r = .14$, $p < .05$). Despite its statistical significance, this correlation is weak (Cohen, 1988), meaning readability is a poor predictor of item difficulty.

Conclusions

Increasing the number of answer choices was deemed the only predictable way to increase the difficulty of MOCCA items. As a result, prior to the Item Calibration Study discussed in the next section, all new items were rewritten to have five answer choices where possible, resulting in 132 such items with only 7 new items having three answer choices. Also based on these results, it was determined that the guessing parameter exhibited minimal variability, and thus it would be set to a constant of .24 for the Item Calibration Study.

6. Item Calibration Study

Items piloted in 2020 were subjected to a large-scale field test from Spring 2021 through Spring 2022. Given results of the New Items Pilot demonstrating that items with five answer choices were more difficult and more discriminating than the same items with three answer choices, the rest of the new item pool was revised to have five answer choices where possible. The purpose of the calibration study was to scale the 239 new items with the existing pool of 352 items using a large, nationally representative sample.

Methods

Sample

The field test/calibration sample consisted of 1,741 students, including 556 third graders, 653 fourth graders and 532 fifth graders. Students attended 16 schools in 11 states with representation in the following census regions: South Atlantic (two schools), the West North Central Midwest (one school), the East North Central Midwest (two schools), the Mountain West (three schools), the Pacific West (two schools) and the Middle Atlantic Northeast (one school). Additionally, 51 individual students were recruited via social media and weren't associated with participating schools. 48.9% of the sample was female, 46.2% was male and 5% was not reported. The race distribution of the sample was 61.8% White, 13.5% Black or African American, 7.6% Hispanic, 4.6% Two or more races, 3.4% Asian, 1.1% American Indian or Alaskan Native and 8.0% unknown. 16.9% of the sample was known to be eligible for free or reduced price meals, although this information wasn't known for 38.7% of the sample. 7.5% of the sample was reported to be eligible for Special Education and 4.0% were English learners. Students within a grade were randomly assigned to one grade-level form of the MOCCA assessment.

MOCCA Assessment

A total of 40 items were included on each form. Items were assigned to the form based on readability, gender of the main character, ending (i.e., positive, negative, neutral), and preliminary item difficulty from the New Items Pilot. The principal investigator assigned items to forms and the research team confirmed equivalency of forms within the same grade level.

A total of 8 forms were created. Third grade students were exposed to four forms, fourth grade students to all eight forms, and fifth and sixth grade students to the four forms third grade students did not see. Thus, the eight forms for fourth grade overlapped with the forms for the other grades.

Across all eight forms, 249 items were field tested. Of these, 10 were items from the 2019 version of MOCCA and had item parameters and thus by appearing on all forms could serve as anchor items for the scaling of the 239 new items. A single item appeared on two forms to balance the number of items across forms.

Criterion Measures

Dynamic Indicators of Basic Early Literacy Skills 8th Edition (DIBELS 8). DIBELS 8 is a set of short, individually administered measures of early literacy development. These measures are designed to assess the acquisition of early reading skills from kindergarten through sixth grade. DIBELS 8 includes a series of screening and monitoring assessments that track skills such as phonemic awareness, alphabetic principle, accuracy and fluency, vocabulary, and comprehension.

EasyCBM. EasyCBM is an online assessment system designed to help educators monitor and evaluate student progress in reading and mathematics. The reading component encompasses measures for letter names, letter sounds, phoneme segmentation, word reading fluency, and passage reading fluency, among others.

Analyses

Item response theory item parameter scores for the reading comprehension dimension and process propensity dimension were estimated based on the data from the pilot. This included item discrimination and difficulty parameters for both dimensions. Each 5- alternative item had one Causal Coherent (CC) alternative, two Paraphrase alternatives, and two Elaboration alternatives. For purposes of measuring the RC Dimension, each item (3- and -5-alternative) was scored correct and incorrect: 1 = Causal Coherent response, 0 otherwise. For purposes of scaling the PP Dimension, each item (both 3- and 5-alternative) was scored as 1 = a Paraphrase alternative, 0 = an Elaboration response, and missing = Causal Coherent. In scoring the 5-alternative items for the PP Dimension, no distinction was made between the two Paraphrases responses nor between the two Elaboration responses. A Paraphrase response was coded 1 no matter which Paraphrase response they chose. Likewise, an Elaboration response was coded as 0 irrespective of which Elaboration response they chose. For purposes of estimating the RC Dimension, responses were scaled using a three-parameter logistic model (3PL) with the scaling constant $D = 1$ and with all lower asymptote parameters set equal to .24.[1] For purposes of estimating the PP Dimension, responses were scaled using a two-parameter logistic model (2PL) with the scaling constant $D = 1$. Resulting difficulty and discrimination parameters were also compared for items with three versus five alternative responses.

[1] In earlier research, we found little difference in fit between a 3PL model with all lower asymptote parameters constrained equal and a 3PL model in which all lower asymptote parameters allowed to vary. In the constrained model, the common lower asymptote parameter equaled .24.

Results

Testing Administration

Participants had an unlimited amount of time to complete the test. An approximate average of 132 students took each form. Across all forms, an average of 34.76 items were completed by participants with a range over forms from 32.39 to 38.59. On average 73% of participants completed all 40 items. Across the several forms, the percentage who completed all 40 items ranged from 61% - 85%.

Calibration Results

Fit statistics. For the RC dimension, the -2 log likelihood (-2LL) statistic was 608.30. The χ^2 fit statistic was 3741.64 (p = 0.00). With 2988 degrees of freedom (df), the χ^2 fit statistic would be sensitive to small deviations from model expected response probabilities. For the PP dimension, -2LL = 30665, and χ^2 = 3715.04. With 3237 df, however, this statistic would be sensitive to small deviations from model expected response probabilities. None of the items in the pool were flagged as misfitting, and as a result, fit was deemed adequate for both dimensions.

Marginal reliability and error of measurement. Table 2 shows mean estimated measurement error variance by form for the RC and PP dimensions. The mean error variances are smaller for the RC dimension. This is because the RC score for each student is based on responses to all items completed by the student, whereas the PP score is based only on the smaller number of items that the student answered incorrectly.

Table 2. Mean Measurement Error Variance (MEV) and Marginal Reliability (MR) by Form for the Field Test Study for Forms with 40 Items.

	All Students				Students with RC θ ≤ 0	
	MEV		MR		MEV	MR
Form	RC	PP	RC	PP	PP	PP
3S	0.19	0.40	0.84	0.71	0.42	0.70
3T	0.23	0.36	0.81	0.74	0.36	0.74

3U	0.20	0.52	0.83	0.66	0.49	0.67
3V	0.17	0.50	0.85	0.67	0.51	0.66
4S	0.15	0.51	0.87	0.66	0.56	0.64
4T	0.14	0.49	0.88	0.67	0.50	0.67
4U	0.17	0.54	0.85	0.65	0.59	0.63
4V	0.16	0.65	0.86	0.61	0.63	0.61
4W	0.16	0.46	0.86	0.68	0.54	0.65
4X	0.19	0.55	0.84	0.65	0.45	0.69
4Y	0.15	0.33	0.87	0.75	0.30	0.77
4Z	0.14	0.51	0.88	0.66	0.62	0.62
5W	0.20	0.65	0.83	0.61	0.63	0.61
5X	0.19	0.66	0.84	0.60	0.63	0.61
5Y	0.18	0.57	0.85	0.64	0.59	0.63
5Z	0.19	0.63	0.84	0.61	0.70	0.59

Note. MEV = Mean error variance, MR = Marginal reliability, RC = Reading Comprehension Dimension 1, PP = Process Propensity Dimension 2

Table 2 also shows the marginal reliability estimates for the RC and PP dimensions. Reflecting the differences in the mean error variances, the marginal reliabilities are higher for the RC dimension than for the PP dimension. For the RC dimension, marginal reliabilities are good to excellent, ranging from .81 to .88 across forms. For the PP dimension, they range from .75 to .60. There is a small decline in the PP marginal reliability over grades. This decline results from a decline in the number of incorrect responses for students in higher grades. For the 3rd, 4th and 5th grade forms, the average marginal reliabilities were .69, .67, and .61.

Since process propensity scores are used to classify only students for whom the RC scores is at or below zero, we computed the average estimated error variance and the marginal reliabilities for these students. These are shown in the last two columns of Table 2. They differ little from those of the sample as a whole.

Test information and range of item difficulty. Results also revealed that we have successfully, but modestly extended the range of ability covered by MOCCA items (see Figure 1.1) and the difficulty of individual items (see Figure 1.2).

Figure 1.1.

Theta Range for the 352 Original Item Pool (left panel) and 591 New Item Pool (right panel)

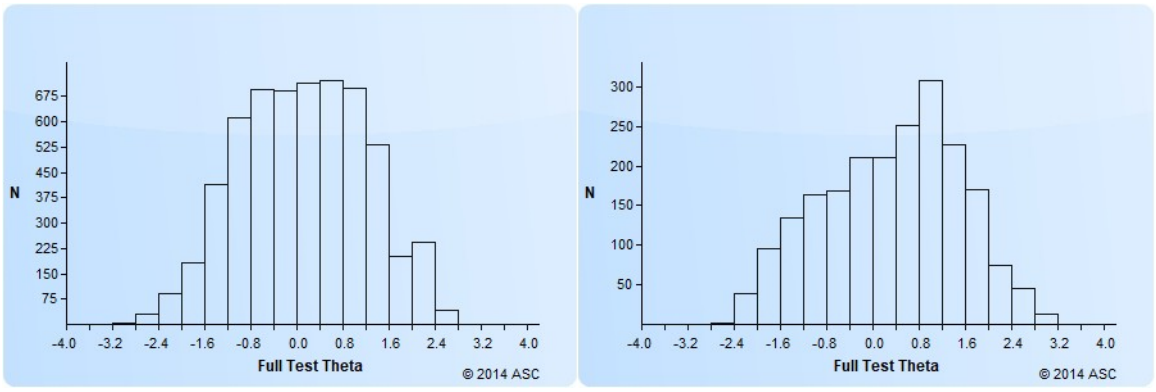
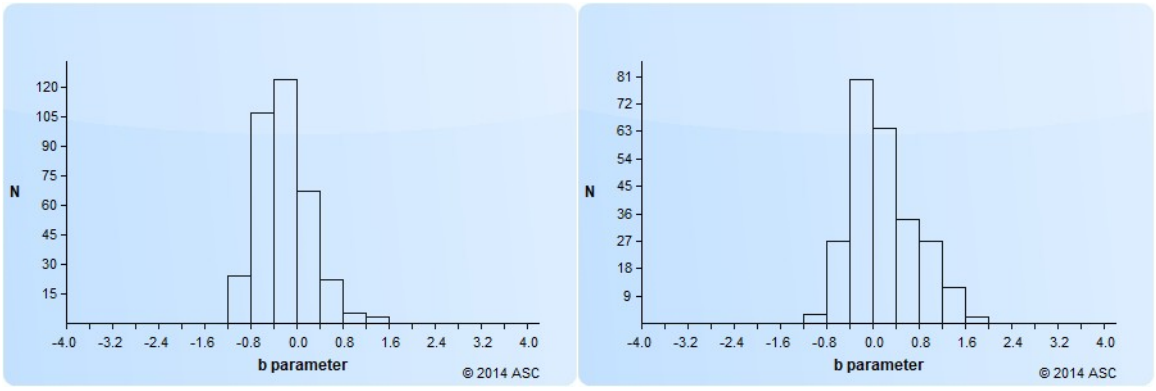


Figure 1.2.

Item Difficulties for the 352 Original Item Pool (left panel) and the 249 New Item Pool (right panel)



Validity Correlations

Table 1 reports concurrent and predictive convergent validity correlations of MOCCA reading comprehension dimension scores with other measures of reading used in the participating schools that were shared with the research team. Table 2 reports concurrent and predictive divergent validity correlations with scores on math measures shared by participating schools. In all cases, we only report correlations where the sample size was 40 or greater. Among the convergent measures reported are measures of component reading skills (i.e., DIBELS and

easyCBM) and more comprehensive measures of reading comprehension (i.e., ILEARN, IREAD-3, and Milestones). Divergent measures included easyCBM Math and Milestones Math.

Table 1

Convergent Validity of Field Study: Correlations of MOCCA with Reading and ELA tests with Sample Sizes Provided in Parentheses

Tests		Concurrent	Predictive		
			EOY-EOY[a]	BOY-EOY	MOY-EOY
		Grade 3			
DIBELS 8	Composite	.47** (53)	-	-	-
easyCBM	Reading – PRF	-	-	.59** (130)	-
	Proficient reading	-	-	.47** (129)	-
	Vocabulary	-	-	.43** (129)	-
ILEARN	ELA	.78 ** (79)	-	-	-
IREAD-3	Total reading	.78** (79)	-	-	-
Milestones	ELA	-	-	.66** (138)	-
		Grade 4			
DIBELS 8	Composite	.77** (43)	-	-	-
easyCBM	Reading – PRF	.54** (126)	.63** (126)	-	.55** (124)
	Proficient reading	.62** (126)	.57** (126)	-	.52** (124)
	Vocabulary	.50** (126)	.58** (126)	-	.49** (124)
Milestones	ELA	.76** (137)	.71** (134)	-	.69** (134)
		Grade 5			
DIBELS 8	Composite	.46** (73)	-	-	-
Milestones	ELA	.71** (138)	-	-	-
		Grade 6			
DIBELS 8	Composite	.61** (49)	-	-	-

Note. DIBELS 8 = Dynamic Indicators of Basic Early Literacy Skills 8th Edition, PRF = passage reading fluency, IREAD-3 = Indiana Reading Evaluation and Determination, Milestones = Education Records Bureau Milestones Assessment.
** p <.01

As expected, convergent correlations are generally strong and positive. Relations with reading component skills are weaker than those with more comprehensive measures. In contrast, divergent correlations are also positive and range from moderate to strong. Stronger correlations are observed for higher grades, perhaps reflecting an increasing reliance on word problems in upper grade levels.

Table 2

Divergent Validity of Field Study: Correlations of MOCCA with Mathematics tests (Sample Sizes in Parentheses)

Test	Subtest	Concurrent	Predictive		
			EOY-EOY[a]	BOY-EOY	MOY-EOY
		Grade 3			
easyCBM	Math	-	.52** (129)	-	-
Milestones	Math	-	-	.54** (138)	-
		Grade 4			
easyCBM	Math	.54** (126)	.53** (126)	-	.34** (124)
Milestones	Math	.56** (137)	.54** (132)	-	.52** (134)
		Grade 5			
Milestones	Math	.67** (138)	-	-	-

Note. Milestones = Education Records Bureau Milestones Assessment.
** p <.01

Conclusions

The 239 new items were successfully calibrated with the existing pool of 352 items that made up the 2019 version of MOCCA. Reliability was adequate at above .80 for reading comprehension dimension scores and .60 for process propensity scores. Validity correlations were also sufficient with the strongest relationships being observed for more comprehensive measures of reading comprehension.

7. Comparison and Validation Study

During the 2022-2023 academic year, a study was conducted to compare the psychometric qualities of a variable-length computerized adaptive test (VCAT) version of MOCCA with the fixed item test (FIT) version of MOCCA. For a full report on how the CAT model was derived, see Davison et al. (2021). The psychometric qualities of interest were standard errors of measurement, number of items taken, testing time, and validity correlations. In addition, the study aimed to determine whether 5-choice items demonstrated greater difficulty and discrimination and to set new grade-specific cut scores for which students would be classified as causal comprehenders versus one of the diagnostic categories. Finally, it was also designed to provide scores with norm-referencing for a nationally representative sample.

Methods

Participants

In all, 2,563 students participated in the study, including 855 students in Grade 3, 774 in Grade 4, 713 in Grade 5 and 221 in Grade 6. Classrooms were randomly assigned to either the VCAT or FIT MOCCA, resulting in 1,335 students who took the VCAT and 1,228 students who took the FIT. Students attended 21 schools in 12 states, representing eight of the nine geographic census regions. Of the total, 47.1% of the sample was female, 50.8% were male, and 2.1% were unknown. The sample was racially diverse; 42.4% of the sample was White, 20.1% was Hispanic, 13.8% was Black or African American, 9.9% was Asian, 7.4% was American Indian or Alaskan Native, 3.7% was two or more races, 0.3% was Native Hawaiian or Other Pacific Islander, and 2.5% unknown. Only 16.2% of the sample was reported to be eligible for free or reduced-price meals, although this information was not available for 60.2% of the sample. Finally, 13.2% of students were eligible for Special Education and 23.3% were English learners.

FIT MOCCA

The CAT was programmed such that the maximum number of items was capped at 25. As a result, new 25-item FIT forms of MOCCA were created for the comparison study. Grade-specific forms were created with an average Flesch-Kincaid readability of about 3.4 for Grade 3, 4.3 for Grade 4, 4.8 for Grade 5, and 5.4 for Grade 6 and an average item difficulty of about -.17 for Grade 3, -.05 for Grade 4, .04 for Grade 5, and .12 for Grade 6. There were three FIT forms per grade, which were randomly assigned to those randomly assigned to the FIT condition. Each form contained 25 operational items and 5 new, non-operational items administered solely for calibration purposes. The results below are based on scale scores and diagnostic classifications computed using only the 25 operational items.

VCAT MOCCA

The VCAT administration consisted of one or two phases for each student. The goal of Phase 1 was to place the student along the RC Dimension. Based on the student's current estimate of θ_1, the next item chosen was the one that would maximize Fisher information. Testing in Phase 1 continued until the student had completed 15 items or their estimated SEM fell below 0.35. At the end of Phase 1, testing stopped completely for students whose estimated $\theta_1 \geq 0$. The students whose estimated $\theta_1 \geq 0$ were placed in the Causal category. Students for whom estimated $\theta_1 < 0$ proceeded to Phase 2.

In Phase 2, each succeeding item was chosen to maximize a weighted Fisher information along θ_2, Fisher information multiplied by the estimated probability that the student would get the item wrong. A response supplies information about Dimension 2 only if answered incorrectly, and the weighted Fisher information function is more likely to select an item that the student will answer incorrectly than is simple Fisher information. After each item, the student's θ_2 was updated as was the LR. The LR was based on all the items that the students had missed in

Phase 1 and Phase 2. Testing stopped as soon as the LR indicated a classification or the student reached 25 items (the number of items in the FIT condition), whichever came first. After completing Phase 2, the student would receive a classification of Paraphrase, Elaboration, or Inconclusive based on their LR statistic. Students in the Inconclusive category completed all 25 items without being placed in either the Paraphrase or Elaboration category at any point in Phase 2.

Criterion Measures

Measures of Academic Progress (MAP). MAP is a computer-adaptive assessment tool produced by the Northwest Evaluation Association (NWEA) that offers a comprehensive overview of student academic growth. In terms of reading and ELA, MAP tests cover various areas like vocabulary acquisition, literary analysis, and comprehension strategies for both informational and literary texts.

For grades 2-6 in mathematics, MAP covers a broad range of content. This can include foundational arithmetic concepts, geometry, measurement, data analysis, algebraic thinking, and problem-solving strategies. The breadth and depth of topics expand as the grade level increases. Results from MAP can assist educators in personalizing learning, setting student goals, and understanding students' readiness levels for particular content.

Rapid Online Assessment of Reading (ROAR). ROAR is an open-access assessment platform developed by the Stanford Reading & Dyslexia Research Program. It is delivered online and includes assessments of single word recognition, phonological awareness, sentence reading efficiency, and vocabulary.

Analysis

The experimental design for each dependent variable was a two-way ANOVA design, with grade (i.e., 3, 4, 5, or 6) crossed with format (i.e., VCAT vs. FIT). To account for the

random assignment to format at the classroom level, the analysis employed a hierarchical linear model with students nested within classrooms, and two independent variables (Grade and Format) at level 1. Because several classrooms included students in more than one grade, we treated grade as a student variable (level 1). Dependent variables included the number of items completed, minutes of testing time, overall Reading Comprehension IRT score along Dimension 1, standard error of measurement (SEM) along Dimension 1, Process Propensity score along Dimension 2, and SEM along Dimension 2. Two models were fitted for each dependent variable: one that included only main effects for test format and grade level and a second that accounted for the interaction of format and grade. In addition to investigating the effects of format on continuous variables, we also examined how the incidence of categorical diagnostic classifications (i.e., Causally Coherent, Paraphrasing, Elaborating, and Inconclusive Comprehenders) differed between the two formats for each grade. We used χ^2 tests to determine whether observed differences in classification were greater than would be expected due to random sampling variability. Finally, we compared convergent validity correlations for the two formats with MAP and ROAR for each grade. To test whether observed differences between correlations across format were greater than would be expected due to random sampling variability, we first converted each correlation coefficient into a *z*-score using Fisher's *r*-to-*z* transformation. Then, making use of the sample size employed to obtain each coefficient, *z*-scores were compared using formula 2.8.5 from Cohen and Cohen (1983, p. 54).

Results

The goal of this research project was to use VCAT to reduce testing time and items without loss of accuracy for overall reading comprehension, with some of the saved time being allocated to additional item responses for classification purposes. A major part of the research

was to compare measurements obtained via VCAT and FIT to determine if VCAT would produce equivalent measurements while reducing testing time and number of items administered without sacrificing measurement accuracy.

Measurement Equivalence and Accuracy Comparisons

In this portion of the research, there were six continuous dependent variables: number of items completed, minutes of testing time, overall Reading Comprehension IRT score along Dimension 1 (Theta1), standard error of measurement (SEM1) along Dimension 1, Process Propensity score along Dimension 2 (Theta2), and SEM2 along Dimension 2. The experimental design for each dependent variable was a two-way ANOVA design with grade (3, 4, 5, or 6)[2] crossed with format (VCAT vs. FIT). Since random assignment to the VCAT and FIT was by classroom, the analysis employed a hierarchical linear model with students nested within classrooms, and two independent variables (Grade and Format) at level 1. Since a few classrooms had more than one grade, we treated grade as a student variable (level 1). In addition to the continuous variables, there was one categorical dependent variable, diagnostic classification: Causal Coherent, Paraphrase, Elaboration, and Inconclusive.

Number of items. Figure 2 shows the mean number of items completed by grade and format. The hierarchical linear model that included the interaction terms proved significant (see Table 1). Students in the VCAT condition took fewer items. Averaging across grades, students in the VCAT condition took an average of 14.93 items while those in the FIT condition took an average of 24.45. There is an interaction, because students in the VCAT condition took fewer items as grade increased, whereas those in the FIT condition took about the same number of

[2] Second graders were included in the study, but were only administered MOCCA in VCAT format. Therefore, we could not compare VCAT and FIT results for second graders.

items regardless of grade. In the VCAT condition, the mean number of items ranged from 13.87 in Grade 6 to 16.12 in Grade 3. In the FIT condition, the mean number of items ranged from 24.34 in Grade 4 to 24.83 in Grade 6. As hypothesized, VCAT reduced the number of items student took regardless of grade.

Table 1

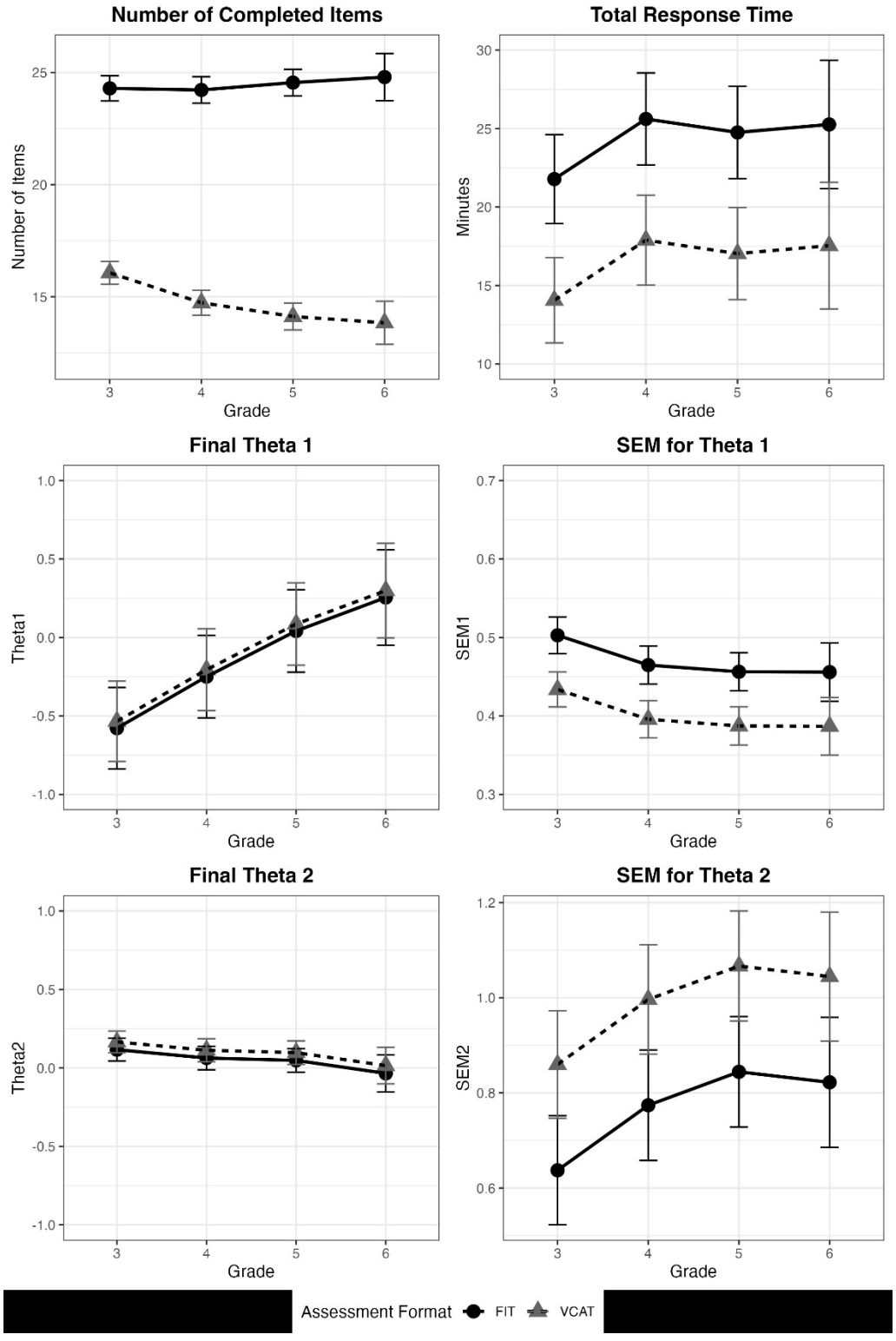
Hierarchical linear model results comparing variable-length adaptive and fixed item test formats by grade level for six dependent variables

	Items completed		Testing time		$\hat{\theta}_1$		SEM$\hat{\theta}_1$		$\hat{\theta}_2$		SEM$\hat{\theta}_2$	
	b	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Grade 4	-1.33	< .001	3.96	< .001	0.33	< .001	-0.03	.026	-0.55	.156	0.13	< .001
Grade 5	1.95	< .001	3.51	.028	0.62	< .001	-0.04	< .001	-0.07	.074	0.18	< .001
Grade 6	-2.23	< .001	3.54	.136	0.83	< .001	-0.05	.040	-0.15	.013	0.18	< .001
Fixed	8.23	< .001	7.83	< .001	-0.02	.834	0.07	< .001	-0.05	.091	-0.21	< .001
Fixed × Grade 4	1.26	.011	--	--	--	--	--	--	--	--	--	--
Fixed × Grade 5	2.21	< .001	--	--	--	--	--	--	--	--	--	--
Fixed × Grade 6	2.73	< .001	--	--	--	--	--	--	--	--	--	--

Note. $\hat{\theta}_1$ = Estimated reading comprehension ability. $\hat{\theta}_2$ = Estimated process propensity. SEM = Standard error of measurement.

Figure 2

Estimated Marginal Means with Confidence Intervals for Six Dependent Variables



Testing time. Figure 2 also shows the average testing time in minutes by grade and format. As reported in Table 1, the interaction effect of format and grade was not significant, thus format had a main effect on testing time such that students taking the VCAT tested for almost eight minutes less than did FIT students. Averaging over grades, the mean testing time for those in the VCAT condition was 17.00 minutes; for those in the FIT condition, it was 25.09. Over grades in the VCAT condition, the average testing time ranged from 14.50 minutes in Grade 6 to 18.38 minutes in Grade 4. In the FIT condition, the average testing time ranged from 21.98 minutes in Grade 3 to 27.87 minutes in Grade 4. As predicted, VCAT reduced testing time at all grades.

Reading comprehension score. Figure 2 also shows mean reading comprehension scores ($\hat{\theta}_1$) by grade and format. As reported in Table 1, only the grade terms were significant. Mean reading comprehension scores increased as grade increased for both formats, and there was no main effect of format. Thus, there is no evidence to suggest differences in scores for the CAT and FIT conditions in the population.

Reading comprehension SEM. Figure 2 also shows the SEM for θ_1 by grade and format. There was no significant interaction, but there was a significant effect for both grade and format. In both the VCAT and FIT conditions, the mean SEM declined as grade increased. At all grades, the mean SEM was lower in the VCAT condition than it was in the FIT condition. In the VCAT condition, mean SEM ranged from 0.371 in Grade 6 to 0.434 in Grade 3. In the FIT condition, the SEM ranged from 0.455 in Grade 4 to 0.502 in Grade 3. Thus, despite being a shorter test, as measured by mean number of items completed or time in minutes, VCAT led to smaller SEMs at every grade than did FIT.

Process propensity score. Figure 2 also shows the Process Propensity mean $\hat{\theta}_2$ scores by grade and format. Table 1 reports the results for the main effects model, however neither the main effects of grade and format, nor their interaction were significant with exception of Grade 6 students having a significantly lower process propensity than did Grade 3 students. As depicted in Figure 6, this effect is not large in magnitude.

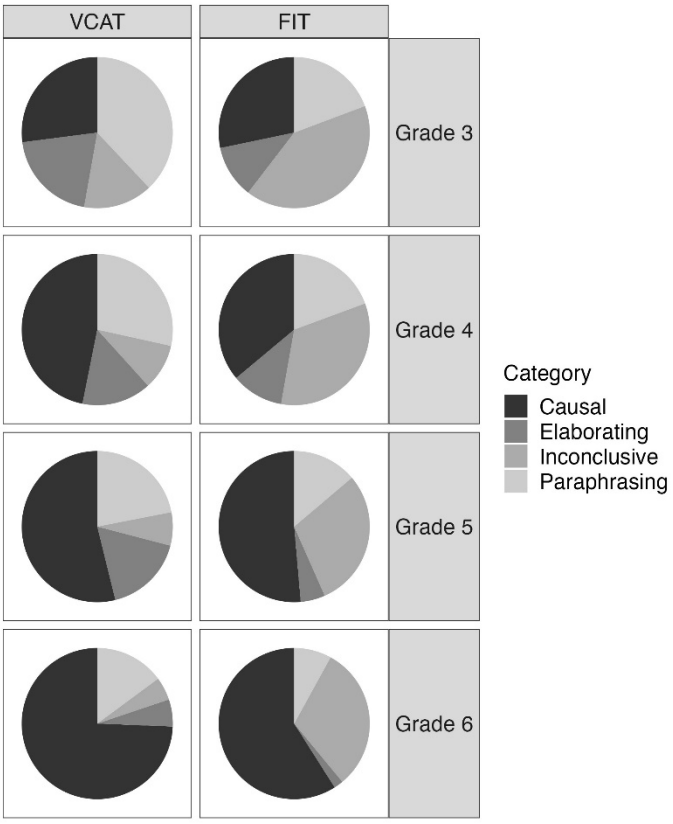
Process propensity SEM. Figure 2 also shows the mean SEM for θ_2. Only the main effects for grade and format were significant. The SEM for θ_2 was lower in the FIT condition than in the VCAT condition Contrary to our hypotheses, students in the VCAT condition had larger SEMs than in the FIT condition.

Classification Comparisons

Figure 3 contains pie charts showing the percentage of readers classified into one of four categories using the LR statistic above: Causal, Elaborate, Inconclusive, and Paraphrase. The Causal category includes students at or above the student mean in the calibration sample, roughly equal to the 4th grade mean in that sample. These are proficient readers unlikely to be classified as struggling by their teachers. The Elaborate and Paraphrase categories include students who might be considered struggling readers ($\theta_1 < 0$) with Elaborate or Paraphrase propensities respectively based on their θ_2 scores and the likelihood ratio test described above. The inconclusive category contains less proficient readers with no clear propensity toward one or the other type of incorrect response.

Figure 3

Diagnostic Classification by Grade and Format



At each grade, we performed a χ^2 test to determine if classification was independent of format. The χ^2 statistic was significant with $p < .001$ at every grade (χ^2= 90.45, 64.63, 77.83, and 28.23 at grades 3 – 6 respectively) leading to rejection of the null hypothesis of independence between test format and classifications at every grade. At every grade, a higher percentage of students were classified as Inconclusive in the FIT condition than in the CAT condition. Correspondingly, at every grade, more students are classified in the Causal, Elaborate, or Paraphrase categories in the CAT than in the FIT condition. Given that these data are real data, the "true" categorization of each student is unknown. The earlier simulation (Davison et al., 2022) found high rates of correct classification using CAT, except near the indifference point. However, that simulation study did not compare CAT and FIT. Ultimately there is a question of

whether the CAT is overclassifying, or the FIT is under-classifying in real data, and this question must remain a matter for future validity research with live data.

Measurement Error Variances and Marginal Reliabilities

Table 1 shows the mean squared measurement error variances (MEV) and the marginal reliabilities (MR) for both the RC Dimension and the PP Dimension, Fixed Form and VCAT formats, and the various fixed forms within the Fixed Format. The top nine rows refer to forms within the Fixed Length condition. Each form is designated by a two-digit number. The first number is the grade, and the second number is the form within the grade. For example, 3.1F is fixed (F) form 1 for grade 3. The last five rows refer to the CAT for grades 2 – 6 respectively. Grade 2 students only took the CAT, so there is MEV and MR data for CAT2 but no fixed form for Grade 2. Grade 2 students had a limit of 15 items (as compared to 25 items for CAT3 – CAT6), so the CAT2 RC MEV is larger than that for CAT3 – CAT6, and the CAT2 RC MR is lower than that for CAT3 – CAT 6.

The first two columns contain the mean error variance (MEV) for the RC and PP dimensions, all students. The most obvious trend in column 1 is that the MEV are lower for the CAT format (except CAT2) than for the Fixed Forms in corresponding grades 3 - 6 along the RC dimension. However, in column 2, the MEVs for the PP Dimension are lower for the Fixed Forms in corresponding grades. The next two columns contain marginal reliabilities for the RC and PP Dimensions. These mirror the results for the MEVs, in that along the RC dimension, marginal reliabilities are higher for the CAT in corresponding grades. The MR range for .74 to .85 for the RC Dimension and fixed forms for forms in grades 3 – 6, whereas they range from .81 to .87 for CAT 3 – CAT6. Along the PP dimension MRs are higher for the Fixed Forms. In the fixed forms for grades 3 – 6, the PP MR range from .44 to .68, whereas for CAT3 – CAT6, they range from .39 to .50. In a later section, we examine the Fixed Form vs. CAT MEV differences

in more detail.

Table 3

Mean Measurement Error Variance (MEV) and Marginal Reliability (MR) by Grade, Format, and Form within the Fixed Length Format

	All Students				Students with RC θ ≤ 0	
	MEV		MR		MEV	MR
Form	RC	PP	RC	PP	PP	PP
3.1F	0.30	0.67	0.77	0.60	0.18	0.85
3.2F	0.34	0.55	0.75	0.64	0.20	0.83
3.3F	0.28	0.71	0.78	0.59	0.20	0.83
4.1F	0.26	0.90	0.79	0.53	0.21	0.83
4.2F	0.25	1.05	0.80	0.49	0.20	0.83
4.3F	0.22	0.48	0.82	0.68	0.18	0.84
5.1F	0.30	1.10	0.77	0.48	0.19	0.84
5.2F	0.24	1.29	0.81	0.44	0.23	0.82
5.3F	0.23	1.04	0.81	0.49	0.20	0.83
6.1F	0.34	0.60	0.74	0.62	0.16	0.86
6.2F	0.18	0.90	0.85	0.53	0.20	0.83
6.3F	0.30	0.76	0.77	0.57	0.16	0.86
CAT2	0.40	0.59	0.71	0.63	0.39	0.72
CAT3	0.23	0.99	0.81	0.50	0.41	0.71
CAT4	0.19	1.44	0.84	0.41	0.49	0.67
CAT5	0.18	1.47	0.85	0.41	0.52	0.66
CAT6	0.15	1.58	0.87	0.39	0.32	0.76

Note. RC = Reading Comprehension Dimension 1, PP = Process Propensity Dimension 2, MEV = Measurement Error Variance, MR = Marginal Reliability

In CAT, only students with RC $\theta \leq 0$ participated in Phase 2. Therefore, we computed the MEV and MR for students in the CAT Format with RC $\theta \leq 0$. These values are shown in the last five rows of the last two columns. The MEVS are clearly smaller for students who took Phase 2 than for all students. For instance, for CAT 2, the mean MEV is .59 for all students and .39 for students who took Phase 2. Also, the marginal reliability is higher for those who took Phase 2.

For instance, the MR for the PP Dimension in CAT2 is .63 for all students and .72 for students who took Phase 2. Among the CAT students who took Phase 2, the marginal reliabilities range from .66 with CAT4 to .76 for CAT6. Table 1 also shows the MEV and MR for students with RC $\theta \leq 0$ taking fixed length forms. Their MEVs are generally lower and their reliabilities are generally higher than those taking CAT in corresponding grades. With adjustments to the stopping rules regarding Phase 2, the CAT MEVs and MR for the PP dimension could likely be made more comparable to those for fixed forms, but these adjustments would likely increase the length of the CAT, at least for some students.

Three versus Five Alternative Responses

Table 2 shows the mean IRT difficulty and discrimination parameters for 3- and 5-alternative items along both the RC Dimension and the PP Dimension. For the RC Dimension 1, there was a significant difference between the 3- and 5-alternative items for both the difficulty and discrimination parameters. The 3-alternative items were more difficult. The mean item difficulties for 3- and 5- alternative items were -.138 and .188 respectively. The 3-alternative items were also less discriminating. The mean item discriminations were 1.99 and 2.10 respectively for the 3- and 5-alternative items.

Along the PP dimension, the mean item difficulty parameters were not significantly different for the 3- and 5-alternative items. However, the discrimination parameters were significantly different: 1.14 for the 3-alternative items and 1.39 for the 5-alternative items. The 5-alternative items were significantly more discriminating.

Table 4

IRT Item Difficulty and Discrimination Parameters for 3- and 5-alternative Items along Two Dimensions

Options	*N*	*M*	*SD*	Min	Max	*t*	*df*	*p*
RC dimension								
Item difficulty (IRT *b*)								
3	191	-0.14	0.46	-1.07	1.45	-5.43	241.74	< .001
5	132	0.19	0.57	-1.16	1.66			
Item discrimination (IRT *a)*								
3	191	1.99	0.34	1.28	2.97	-2.97	321	.006
5	132	2.10	0.33	1.51	3.23			
PP dimension								
Item difficulty (IRT *b*)								
3	191	-0.03	0.41	-2.23	0.89	-0.89	321	.374
5	132	-0.01	0.36	-0.91	1.50			
Item discrimination (IRT *a*)								
3	191	1.14	0.19	0.70	1.67	-10.85	321	< .001
5	132	1.38	0.22	0.37	1.82			

Validity Correlations

To evaluate the convergent and divergent validity of MOCCA, convergent validity of MOCCA was evaluated by correlating scores with other reading or English language arts tests (i.e., DIBELS 8, easyCBM, mCLASS, MAP, and ROAR) both concurrently and predictively. Only those correlations where both the FIT and CAT versions of MOCCA had sample sizes of 40 or greater are reported. Measures in the tables are listed alphabetically by grade with MAP Reading being the most comparable measure in that it assesses reading comprehension in an untimed manner. Correlations were estimated separately for the CAT and FIT versions of MOCCA and compared statistically.

Table 3.3.3 reports concurrent convergent validity correlations of MOCCA reading comprehension dimension scores with other measures of reading used in the participating schools that were shared with the research team. Most comparisons failed to reject the null hypothesis. Exceptions, which are in bold font, included easyCBM PRF in Grade 3 and DIBELS Composite, Maze, and ORF Rate scores in Grade 5. In each case, the FIT version of MOCCA correlated more strongly with the criterion than did the CAT version and sample sizes were relatively small. Correlations with the strongest criterion measure, MAP, were all above .60 for both versions.

Table 3.3.3

Comparison of Concurrent Convergent Validity Correlations of MOCCA Versions for a Range of Reading Measures in Grades 3-6

		n		*r*			
Grade	Criterion	VCAT	FIT	VCAT	FIT	*z*	*p*
3	MAP Reading	315	244	0.67	0.69	0.63	0.53
	ROAR Word Reading	295	227	0.60	0.50	1.58	0.11
4	MAP Reading	204	265	0.77	0.73	1.11	0.27
	ROAR Word Reading	161	214	0.62	0.61	0.08	0.93
5	MAP Reading	228	181	0.72	0.75	0.68	0.49
	ROAR Word Reading	163	127	0.50	0.60	1.14	0.25

Note. VCAT = computerized adaptive test. FIT = fixed item test. MAP = Measures of Academic Progress. ROAR = Rapid Online Assessment of Reading.

Table 3.3.4 reports predictive convergent validity correlations of MOCCA reading comprehension dimension scores with other measures of reading. Once again, most comparisons failed to reject the null hypothesis. Exceptions included easyCBM Vocabulary from BOY and

MOY to EOY in Grade 3 and DIBELS Composite and ORF Rate scores in Grade 5 from BOY to MOY. In each case, the FIT version of MOCCA correlated more strongly with the criterion than did the CAT version and sample sizes were relatively small. Correlations with the strongest criterion measure, MAP, were again all above .60 for both versions.

Table 3.3.4

Comparison of Predictive Convergent Validity Correlations of MOCCA Versions for a Range of Reading Measures in Grades 3-6

Grade	Criterion	MOCCA TOY	Criterion TOY	VCAT *n*	FIT *n*	VCAT *r*	FIT *r*	*z*	*p*
3	DIBELS 8 Composite	BOY	MOY	69	45	0.63	0.42	1.49	0.13
		BOY	EOY	72	48	0.64	0.44	1.5	0.13
		MOY	EOY	66	48	0.67	0.65	0.19	0.85
	DIBELS 8 Maze	BOY	MOY	69	46	0.7	0.59	0.91	0.36
		BOY	EOY	72	48	0.64	0.58	0.44	0.66
		MOY	EOY	66	48	0.7	0.68	0.24	0.81
	DIBELS 8 NWF-CLS	BOY	MOY	67	46	0.6	0.39	1.4	0.16
		BOY	EOY	64	46	0.61	0.39	1.49	0.14
		MOY	EOY	60	46	0.68	0.59	0.74	0.46
	DIBELS 8 NWF-WRC	BOY	MOY	67	46	0.6	0.49	0.8	0.43
		BOY	EOY	64	46	0.62	0.42	1.34	0.18
		MOY	EOY	60	46	0.68	0.62	0.53	0.59
	DIBELS 8 ORF Accuracy	BOY	MOY	69	46	0.41	0.23	1.01	0.31
		BOY	EOY	72	48	0.31	0.17	0.81	0.42
		MOY	EOY	66	48	0.35	0.16	1.07	0.29
	DIBELS 8 ORF Rate	BOY	MOY	69	46	0.58	0.43	1.06	0.29
		BOY	EOY	72	48	0.6	0.39	1.42	0.16
		MOY	EOY	66	48	0.64	0.61	0.28	0.78
	DIBELS 8 WRF	BOY	MOY	67	45	0.64	0.42	1.52	0.13
		BOY	EOY	64	46	0.66	0.44	1.63	0.10
		MOY	EOY	60	46	0.74	0.65	0.93	0.35
	easyCBM - PRF	BOY	MOY	44	45	0.46	0.72	1.84	0.07
		BOY	EOY	44	45	0.47	0.71	1.72	0.09
		MOY	EOY	44	42	0.62	0.66	0.35	0.73
	easyCBM – Proficient Reading	BOY	MOY	44	46	0.29	0.64	2.13	0.03
		BOY	EOY	44	45	0.32	0.61	1.72	0.09
		MOY	EOY	44	42	0.6	0.81	1.99	0.05
		BOY	MOY	44	46	0.42	0.65	1.48	0.14

Grade	Criterion	MOCCA TOY	Criterion TOY	VCAT *n*	FIT *n*	VCAT *r*	FIT *r*	*z*	*p*
	easyCBM - Vocabulary	**BOY**	**EOY**	**44**	**45**	**0.16**	**0.59**	**2.35**	**0.02**
		MOY	**EOY**	**44**	**42**	**0.39**	**0.71**	**2.13**	**0.03**
	MAP Reading	MOY	EOY	321	241	0.68	0.66	0.51	0.61
	ROAR Word Reading	MOY	EOY	251	161	0.56	0.59	0.34	0.74
	ROAR Vocabulary	MOY	EOY	235	146	0.34	0.43	1.03	0.30
4	DIBELS 8 Composite	BOY	MOY	49	50	0.72	0.67	0.49	0.63
		BOY	EOY	47	48	0.49	0.62	0.91	0.36
	DIBELS 8 Maze	BOY	MOY	49	50	0.75	0.65	0.94	0.35
		BOY	EOY	47	48	0.7	0.58	1.03	0.30
	DIBELS 8 ORF Accuracy	BOY	MOY	49	50	0.23	0.25	0.11	0.91
		BOY	EOY	47	48	0.36	0.28	0.4	0.69
	DIBELS 8 ORF Rate	BOY	MOY	49	50	0.72	0.67	0.46	0.65
		BOY	EOY	47	48	0.48	0.62	0.98	0.33
	easyCBM – PRF	MOY	EOY	41	46	0.57	0.61	0.24	0.81
	easyCBM – Proficient Reading	MOY	EOY	41	46	0.68	0.64	0.31	0.75
	easyCBM – Vocabulary	MOY	EOY	41	46	0.74	0.65	0.77	0.44
	MAP Reading	MOY	EOY	197	220	0.73	0.74	0.14	0.89
	ROAR Word Reading	MOY	EOY	144	145	0.62	0.46	1.94	0.05
	ROAR Vocabulary	MOY	EOY	116	135	0.49	0.37	1.07	0.29
5	DIBELS 8 Composite	**BOY**	**MOY**	**62**	**43**	**0.44**	**0.72**	**2.1**	**0.04**
		BOY	EOY	59	43	0.4	0.6	1.28	0.20
	DIBELS 8 Maze	BOY	MOY	62	43	0.49	0.65	1.14	0.25
		BOY	EOY	59	43	0.49	0.68	1.42	0.16
	DIBELS 8 ORF Accuracy	BOY	MOY	62	44	0.41	0.42	0.04	0.97
		BOY	EOY	59	43	0.45	0.27	1.02	0.31
	DIBELS 8 ORF Rate	**BOY**	**MOY**	**62**	**44**	**0.43**	**0.72**	**2.2**	**0.03**
		BOY	EOY	59	43	0.39	0.59	1.26	0.21
	MAP Reading	MOY	EOY	173	175	0.7	0.67	0.4	0.69

Grade	Criterion	MOCCA TOY	Criterion TOY	VCAT *n*	FIT *n*	VCAT *r*	FIT *r*	*z*	*p*
	ROAR Word Reading	MOY	EOY	79	82	0.56	0.53	0.28	0.78
	ROAR Vocabulary	MOY	EOY	73	80	0.54	0.38	1.24	0.21
6	DIBELS 8 Composite	BOY	MOY	44	54	0.51	0.68	1.26	0.21
		BOY	EOY	43	55	0.51	0.62	0.82	0.41
	DIBELS 8 Maze	BOY	MOY	44	54	0.63	0.71	0.78	0.44
		BOY	EOY	43	55	0.51	0.72	1.67	0.09
	DIBELS 8 ORF Accuracy	BOY	MOY	44	55	0.32	0.5	1.01	0.31
		BOY	EOY	43	55	0.33	0.3	0.16	0.88
	DIBELS 8 ORF Rate	BOY	MOY	44	55	0.5	0.66	1.2	0.23
		BOY	EOY	43	55	0.51	0.62	0.79	0.43

Note. CAT = computerized adaptive test. FIT = fixed item test. TOY = time of year. BOY = beginning of year. MOY = middle of year. EOY = end of year. NWF = nonsense word fluency. CLS = correct letter sounds. WRC = words recoded correctly. ORF = oral reading fluency. WRF = word reading fluency. PRF = passage reading fluency. MAP = Measures of Academic Progress. ROAR = Rapid Online Assessment of Reading. Bold font = statistical significance.

Conclusions

Results from the Item Calibration and Validation Study broadly support the superiority of the CAT version of MOCCA when compared to a FIT version. While reliability and validity evidence is largely commensurate across the two versions, the shorter testing times, fewer items delivered, and smaller standard errors of measurement associated with the CAT mean the same or better psychometric qualities are achieved with a less onerous testing experience and better precision.

8. Criterion and Norm Referencing of CAT MOCCA

Criterion and norm referencing of MOCCA Cat were achieved by using results from the Comparison and Validation Study. Specifically, those students who took the CAT version of MOCCA, including an additional sample of second grade students, contributed to a receiver operating curve analysis predicting to the 40th percentile rank on the Measures of Academic Progress reading test, as well as to calculation of percentile ranks for their corresponding grade.

Methods

Participants

Students who took the MOCCA CAT in the Comparison and Validation Study numbered 1,443. Of these, 560 students were in Grade 3, 411 in Grade 4, 351 in Grade 5 and 121 in Grade 6. In addition, 436 Grade 2 students also took MOCCA CAT. Demographically, 46.4% were female, 51.2% were male and 2.3% were unknown; 40.8% were White, 19.3% were Hispanic, 16.1% were Black or African American, 9.5% were Asian, 7.7% were American Indian or Alaskan Native, 3.7% were Two or more races, 0.2% were Native Hawaiian or other Pacific Islander, and 2.8% were unknown. In addition, 13.5% of the sample were reported to be eligible for Special Education, 22.7% were English learners, and 14.9% of students were reported to be eligible for free or reduced-price lunch, although eligibility for the latter wasn't known for 59.9% of the sample. For calculating norms, only those students completing 10 or more items were included; the adjusted sample sizes by grade were 369 students in Grade 2, 464 in Grade 3, 397 in Grade 4, 351 in Grade 5, and 121 in Grade 6.

Measures of Academic Progress

MAP is a computer-adaptive assessment tool produced by the Northwest Evaluation Association (NWEA) that offers a comprehensive overview of student academic growth. In

terms of reading and ELA, MAP tests cover various areas like vocabulary acquisition, literary analysis, and comprehension strategies for both informational and literary texts.

Results

Receiver Operating Curve Analysis Results

Concurrent and predictive accuracy of the MOCCA CAT Reading Comprehension (RC) score were empirically evaluated using the MOCCA RC score to predict whether students scored below the 20th, 30th, or 40th percentiles on the MAP Reading assessment in each grade. These empirically derived cut scores were then compared to the performance of a theoretically derived cut score of 0, which previous versions of MOCCA used to determine when to categorize students as a causal comprehender or report their process propensity category as *paraphraser*, *elaborator*, or *inconclusive*. The data for these analyses were collected from one large school district in the 2022-23 MOCCA Comparison Study that provided middle of year (MOY) and end of year (EOY) data on MAP Reading for Grades 2-5.

These analyses employed Receiver Operating Characteristic (ROC) curve analyses, which describe the relation between true positive rates (i.e., scores that correctly identify students who are not on track for attaining proficiency) and false positive rates (i.e., scores that indicate a student was not on-track when they really were). ROC analyses yield an area under the curve (AUC) estimate, which summarizes a test's classification accuracy. An AUC of .5 indicates the test predicts no better than chance, whereas an AUC of 1.0 indicates that a test is perfectly predictive (Habibzadeh, Habibzadeh, & Yadollahie, 2016).

Separate ROC analyses were conducted for each combination of grade (i.e., grades 2-5), criterion threshold (i.e., 20th, 30th, or 40th percentile) and assessment occasion (i.e., middle of year [MOY] and end of year [EOY]). Students in the MAP Reading analytic sample took both the CAT version of MOCCA and MAP Reading at MOY and EOY, resulting in two sets of

concurrent comparisons and one set of predictive comparisons (i.e., MOY MOCCA predicting EOY MAP Reading) in each grade. Across three comparisons, three thresholds, and four grades, 36 ROC analyses were conducted. The results of each analysis are summarized in Table 7.

Table 7

Sample sizes and area under the curve (AUC) values from Receiver Operating Characteristic Curve analyses of the MOCCA Reading Comprehension score predicting performance above or below the 20th, 30th, and 40th percentile on the MAP Reading assessment in grades 2 - 5.

Grade	Comparison	Type	N	Threshold		
				20th	30th	40th
2	MOY –MOY	Concurrent	436	.801	.820	.848
	MOY – EOY	Predictive	423	.763	.790	.830
	EOY – EOY	Concurrent	111	.793	.870	.908
3	MOY –MOY	Concurrent	560	.804	.837	.852
	MOY – EOY	Predictive	548	.813	.819	.852
	EOY – EOY	Concurrent	134	.870	.824	.830
4	MOY –MOY	Concurrent	411	.882	.882	.868
	MOY – EOY	Predictive	398	.868	.853	.878
	EOY – EOY	Concurrent	118	.804	.797	.834
5	MOY –MOY	Concurrent	329	.835	.834	.826
	MOY – EOY	Predictive	321	.867	.852	.835
	EOY – EOY	Concurrent	101	.898	.879	.873

Across analyses, the AUC exceeded .76, with all but 3 (92%) exceeding .80, values that indicate the MOCCA RC score is a strong predictor of MAP reading performance in all grades evaluated (i.e., 2-5), particularly when predicting concurrently to the MAP 40th percentile. The 40th percentile analyses displayed the best AUC confidence intervals and make pragmatic sense, given that they minimize the number of students misclassified as causal comprehenders when they likely need additional instructional support. Similarly, given that the concurrent analyses minimize the potential confound of instructional impact, and sample sizes were largest for the MOY analyses, the MOY – MOY 40th percentile analyses were used to identify an empirical cut

score for each grade. Those results are summarized in Table 7. In Grades 2-4, the lower bound of the AUC confidence interval exceeded .80, and in all cases, sensitivity and specificity values exceeded .7, with all but the sensitivity value in Grade 5 exceeding .75.

Table 8

Summary of Receiver Operating Characteristic (ROC) Curve analyses predicting performance above or below the 40th percentile on the MAP Reading assessment in grades 2 – 5.

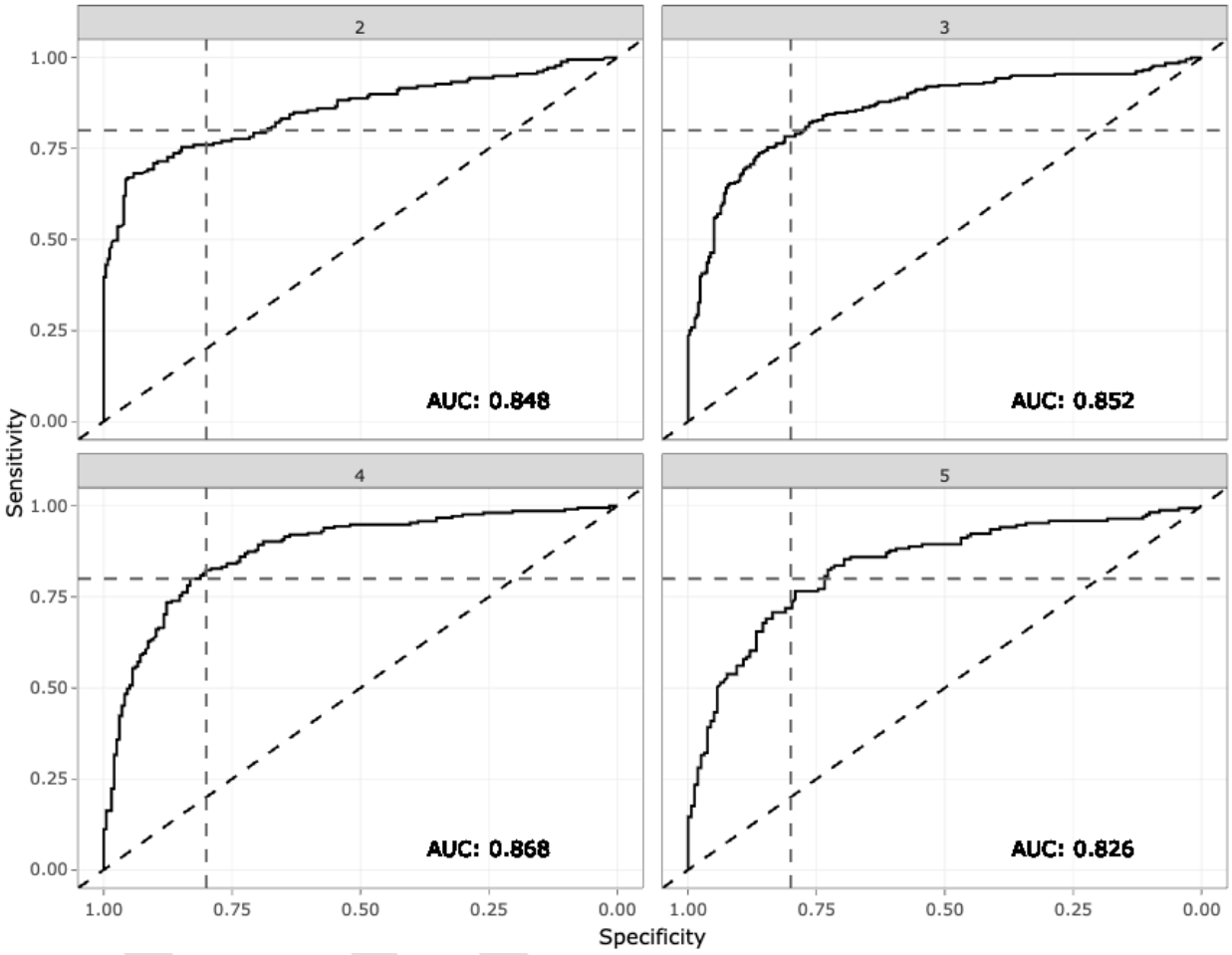
Grade	N	AUC	AUC CI	Cut	Sensitivity	Specificity
2	436	.85	.81 – .89	-.97	.76	.82
3	560	.85	.82 – .88	-.78	.78	.81
4	411	.87	.83 – .90	-.51	.82	.80
5	329	.83	.78 – .87	-.13	.72	.81

Note. AUC = Area under the ROC curve, CI = confidence interval, cut = empirically derived cut score, Sensitivity = proportion of students who scored below the 40th percentile on the MAP Reading who also scored below the specified MOCCA RC cut, Specificity = proportion of students who scored above the 40th percentile on the MAP Reading who also scored above the specified MOCCA RC cut.

Figure 11 visualizes, for each grade, the relationship between sensitivity and specificity across the range of MOCCA RC scores. Each quadrant of the plot represents a separate grade level, as indicated in the gray bar across the top of each quadrant. In all four grades, the optimal cut score meets or nearly meets the highest technical standards for sensitivity and specificity as specified by the National Center on Intensive Intervention (NCII, 2023), as illustrated by the horizontal and vertical dashed lines in Figure 11.

Figure 11

Receiver Operating Characteristic (ROC) Curves for middle of year (MOY) MOCCA Reading Comprehension scores predicting MOY MAP Reading performance below the 40th percentile.



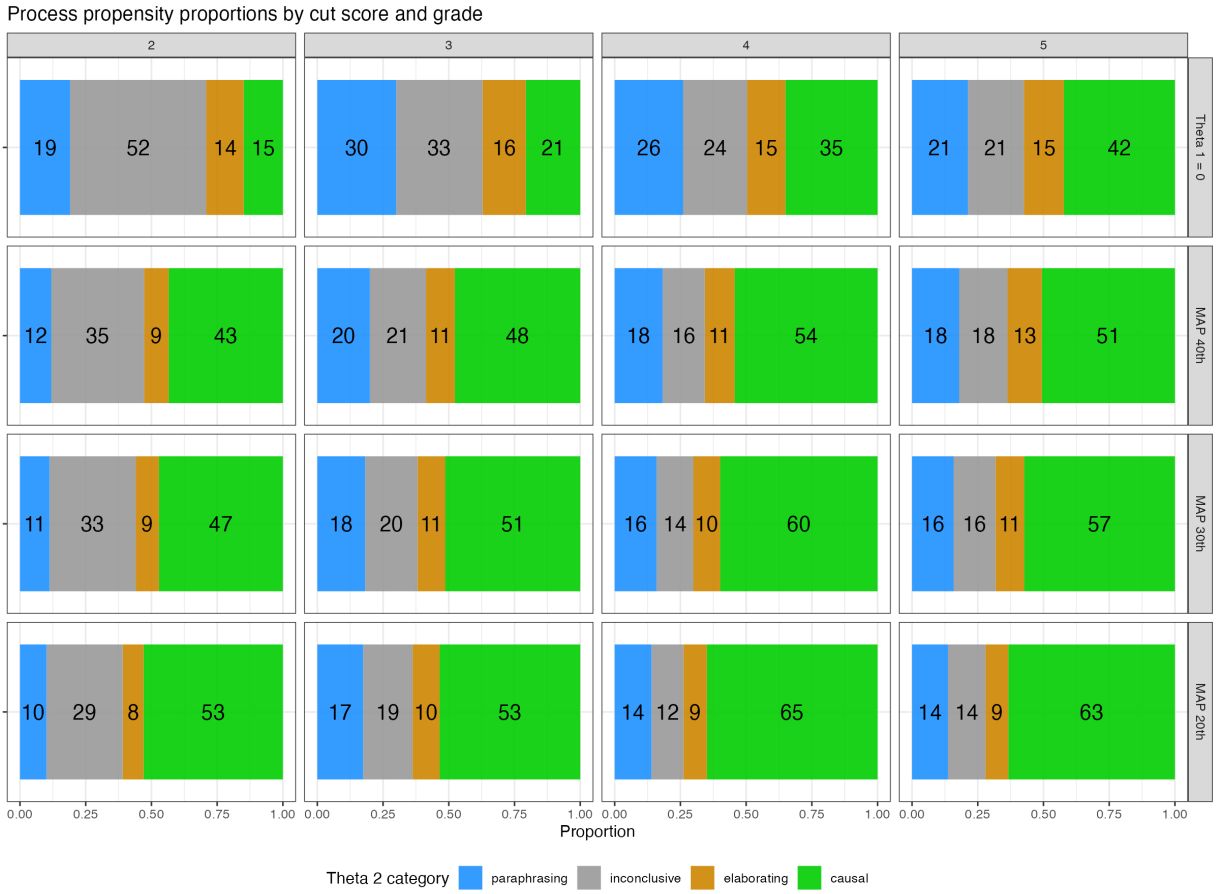
Note. AUC = Area under the ROC curve, CI = confidence interval, cut = chosen cut score, Sensitivity = proportion of students who scored below the 40th percentile on the MAP Reading who also scored below the specified cut on the MOCCA RC score, Specificity = proportion of students who scored above the 40th percentile on the MAP Reading who also scored above the specified cut on the MOCCA RC score. The horizontal and vertical dashed lines represent the National Center on Intensive Intervention (NCII) technical standards for sensitivity and specificity (NCII, 2023). The diagonal dashed line represents chance prediction.

Figure 12 provides a visual comparison of the implications of using the theoretically derived cut score of 0 to determine process propensity classifications versus each of the three empirically derived cuts. As Figure 12 shows, the empirically based cut scores decrease the

proportion of students classified as inconclusive and increases the proportion of students classified as causal comprehenders in all grades.

Figure 12

Proportion of students in each MOCCA process propensity category by cut score type.

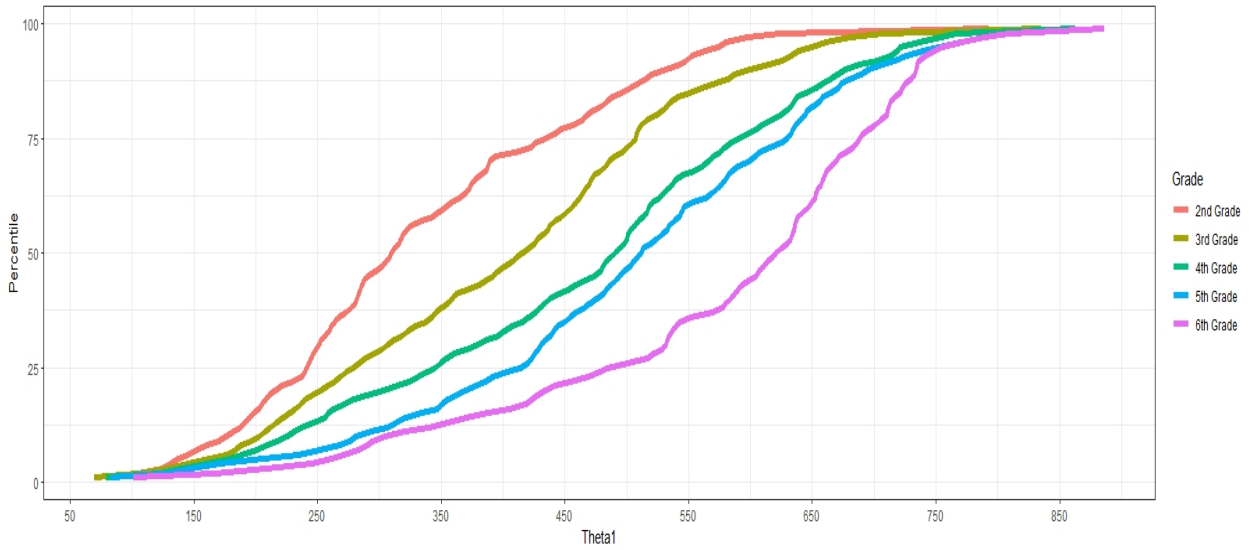


MOCCA CAT Norms

In an earlier technical report, we reported norm tables based on fixed length, 40 item forms of MOCCA. Going forward, the CAT version of MOCCCA will be the major operational format. Therefore, we have renormed MOCCA based on the CAT sample within our larger calibration sample. Specifically, we have developed norms for the RC scale scores for grades 2 – 6. Since scores along the PP Dimension are not reported, no norm tables have been developed for the second dimension. While there is little reason to think that norms will be different for fixed-length and CAT versions, there is some reason to expect small differences at the extreme high

and low ends of the distribution, because there are more marked ceiling and floor effects with the fixed length version than with the CAT version. In general, scores do not "bunch up" at the top and bottom endos of the CAT distribution like they do for the Fixed Forms.

Figure 9. Plot of Percentile Ranks by Reading Comprehension Scale Scores for Grades 2 – 6



The norms are user norms developed without any weighting to make them more representative of a national population. The sample consisted of those students who took the CAT in the Comparison and Validation Study and is described in more detail there. In terms of major ethnic groups (Black, Hispanic, and White), the user group is reasonably representative of the U.S. student population. However, it has a larger proportion of boys than the national population. It has students from all four major geographic regions, but not in the same proportions as in the U.S. student population.

Appendix Table A.1 shows the full norm table. At each of grades 2 – 6, it shows the interval of scale scores associated with each percentile rank 1 – 99. For each of grades 2 – 6, Figure 1 contains a line graph showing the midpoint of the scale score interval (horizontal axis) associated with each percentile rank (vertical axis). To derive the percentile ranks, we did not use a formal smoothing function. Rather, we used a formula based on scale scores and their

cumulative percentiles to derive the upper and lower limit for the interval associated with each percentile rank. This formula yields curves (Figure 1) that is not perfectly smooth but nearly so.

Conclusions

ROC Analyses demonstrated that MOCCA CAT is an excellent predictor of average or better performance on the MAP. These results informed the adjustment of the CAT cut-score between good and poor comprehension from being the same (theta-1 = 0) to being grade-specific ().

9. Measurement of Change

In addition to providing reading comprehension scores and diagnostic classifications of processing modes for struggling readers, MOCCA can also be used for measuring growth in reading comprehension at both the group and single student level. Toward this end, MOCCA was administered at least twice to students from Grades 3 through 6 during the period from September 2022 through May 2023. The results from the first administration of this series of tests were reported in earlier chapters. This chapter analyzes the data from successive tests on the same students to evaluate growth in reading comprehension for both groups of students and individual students. Test forms—CAT versus FIT—were randomly assigned to classrooms. All tests were computer-administered. To avoid repeating items across the three administrations, students in the FIT group received different alternate forms of the MOCCA at each administration; for the CAT group, the item bank was divided into three smaller independent banks with similar information functions, with a separate bank used for each testing occasion. Group-level analyses were primarily based on the θ estimates on the RC dimension; individual-level analyses included results from the PP dimension.

Methods

Participants

The sample for this study consisted of participants in the Comparison and Validation Study who took MOCCA CAT or FIT more than once. The sample for the growth analyses included 1,434 Grade 3 through 6 students who completed at least 10 items in two or more tests. Table 1 shows the distribution of students by Grade and test Type. These students were spread across 15 schools in 13 districts. As Table 2 shows, across Grades, 48.4% of students for whom gender were reported were female and 51.6% were male; 54.5% were white, 6.9% were Hispanic, 15.2% were Black or African American, 6.5% were Asian, and 11.5% were American

Indian or Alaskan Native. Approximately 9.1% of students identified as English language learners. Free and reduced-price meal status was available for 63.4% of the sample, but of those reporting it, 42.0% were eligible.

Table 1. Sample Size by Grade and Test Format

	Grade				
Format	3rd	4th	5th	6th	Total
CAT	216	233	203	115	767
FIT	200	179	194	94	667
Total	416	412	397	209	1434

Table 2. Demographic Characteristics of the Longitudinal Student Sample

Characteristic	N	Proportion
Gender		
Female	676	0.471
Male	721	0.503
NA	37	0.026
Race/Ethnicity		
American Indian/Alaskan Native	160	0.112
Asian	91	0.063
Black/African American	212	0.148
Hispanic	96	0.067
Multiracial	69	0.048
Native Hawaiian/ Pacific Islander	5	0.003
White	758	0.529
NA	43	0.030
English language learners		
No	1204	0.840
Yes	120	0.084
NA	110	0.077
Free and reduced-price meal status		
No	527	0.368
Yes	381	0.266
NA	526	0.367
Free and reduced-price meal eligibility		
No	527	0.58
Yes	381	0.42
NA	527	0.58

The individual change progress monitoring subsample included 134 Grade 3 through 6 students who completed at least 10 items in four or five tests. These students were spread across four schools, with 49 in Grade 6, 39 in Grade 4, 31 in Grade 5, and 15 in Grade 6. Table 3 shows the demographic characteristics of the group of students. Across Grades, 45% of students for whom gender were reported were female and 52% were male; 34% of students were white, 14% were Hispanic, 2% were Black or African American, 2% were Asian, and 39% were American Indian or Alaskan Native. Approximately 3% of students identified as English language learners. Free and reduced-price meal status was available for 59.0% of the sample, but of those reporting it, 75% were eligible.

Table 3. Demographic Characteristics of the Progress Monitoring Sample of Students

Characteristic	*N*	Proportion
Gender		
Female	60	0.45
Male	70	0.52
NA	4	0.03
Race/Ethnicity		
American Indian or Alaskan Native	52	0.39
Asian	2	0.02
Black/African American	2	0.02
Hispanic	19	0.14
Multiracial	10	0.08
Native Hawaiian/ Pacific Islander	0	0
White	45	0.34
NA	4	0.03
English language learners		
No	126	0.94
Yes	4	0.03
NA	4	0.03
Free and reduced-price meal status		
No	20	0.15
Yes	59	0.44
NA	55	0.41

FIT MOCCA

The CAT was programmed such that the maximum number of items was capped at 25. As a result, new 25-item FIT forms of MOCCA were created for the comparison study. Grade-specific forms were created with an average Flesch-Kincaid readability of about 3.4 for Grade 3, 4.3 for Grade 4, 4.8 for Grade 5, and 5.4 for Grade 6 and an average item difficulty of about -.17 for Grade 3, -.05 for Grade 4, .04 for Grade 5, and .12 for Grade 6. There were three FIT forms per grade, which were randomly assigned to those randomly assigned to the FIT condition. Each form contained 25 operational items and 5 new, non-operational items administered solely for calibration purposes. The results below are based on scale scores and diagnostic classifications computed using only the 25 operational items.

VCAT MOCCA

The VCAT administration consisted of one or two phases for each student. The goal of Phase 1 was to place the student along the RC Dimension. Based on the student's current estimate of θ_1, the next item chosen was the one that would maximize Fisher information. Testing in Phase 1 continued until the student had completed 15 items or their estimated SEM fell below 0.35. At the end of Phase 1, testing stopped completely for students whose estimated $\theta_1 \geq 0$. The students whose estimated $\theta_1 \geq 0$ were placed in the Causal category. Students for whom estimated $\theta_1 < 0$ proceeded to Phase 2.

In Phase 2, each succeeding item was chosen to maximize a weighted Fisher information along θ_2, Fisher information multiplied by the estimated probability that the student would get the item wrong. A response supplies information about Dimension 2 only if answered incorrectly, and the weighted Fisher information function is more likely to select an item that the student will answer incorrectly than is simple Fisher information. After each item, the student's

θ_2 was updated as was the LR. The LR was based on all the items that the students had missed in Phase 1 and Phase 2. Testing stopped as soon as the LR indicated a classification or the student reached 25 items (the number of items in the FIT condition), whichever came first. After completing Phase 2, the student would receive a classification of Paraphrase, Elaboration, or Inconclusive based on their LR statistic. Students in the Inconclusive category completed all 25 items without being placed in either the Paraphrase or Elaboration category at any point in Phase 2.

Results

Group-level Average Growth

The group-level growth analyses were based on three primary timepoints—Fall, Winter, and Spring. The data were analyzed by hierarchical linear modeling (HLM, also known as multilevel models or mixed-effects modeling; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), with time nested within students, nested within classrooms. The following variable names are used in reporting the results:

- **Clock** represents time points since first test session (i.e., growth; 0 is Test 1, 1 is Test 2, 2 is Test 3).
- **Type** represents the difference in intercept between FIT and CAT, averaged across Grades, where CAT is the reference group.
- **Clock × Type** represents the difference in growth between the CAT and the FIT group, i.e., the interaction between Clock and Type.
- Sum contrasts were used for **Grade** level, so other effects were averaged across grade level.

- **Wave** indicates whether the first test was in Fall (Wave 1) or Winter (Wave 2). This term is also used for sum contrasts, so the primary effects of interest (Clock and Type) are averaged over Wave.

ANOVA Results. Table 3 presents the results of an analysis of variance based on the HLM model to test the overall effects of each variable in the design (Appendix Tables xxx – xxx provide means and standard deviations for all main effects and interactions). The results show that there was no significant difference between mean θ estimates on RC due to test Type—mean θs for the CAT, combined across Grade and Wave, were not different from those from the FIT tests, nor was the Type variable included in any significant interactions. The most highly significant effect ($p < 0.001$) was for the Clock variable. These differences reflect mean changes in RC θ estimates across the three testing occasions. The Clock × Wave interaction was also significant at $p < 0.001$ and the Clock × Grade interaction was significant at $p = 0.022$. Additional significant effects were observed for the main effects of Grade ($p = 0.002$) and Wave ($p < 0.001$), and for the two-way interaction of Clock × Grade ($p = 0.022$). There were no significant differences in RC across time due to test Type (Type $p = 0.25$ or Clock × Type ($p = .24$).

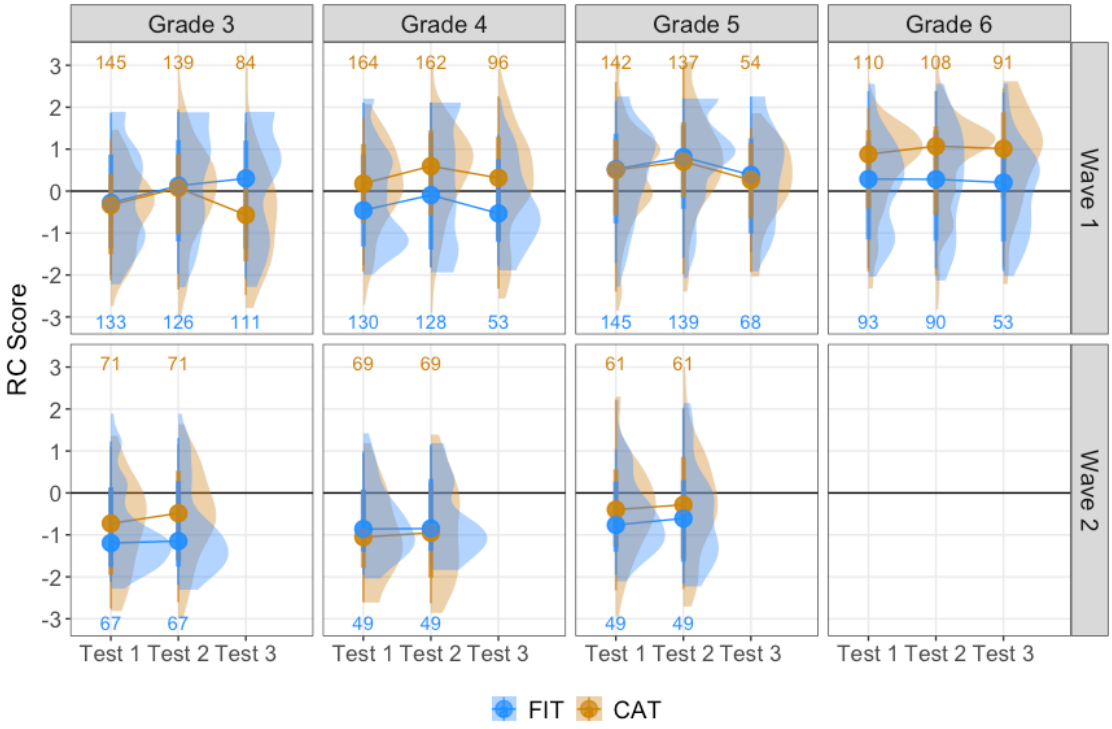
Table 4. Results of the Analysis of Variance

	Degrees of Freedom			
Source of Variation	Numerator	Denominator	*F*	*p*
Clock	2	2032	69.01	<0.001
Type	1	68	1.32	0.254
Grade	3	77	5.32	0.002
Wave	1	155	18.14	<0.001
Clock × Type	2	2033	1.43	0.240
Clock × Grade	6	2033	2.47	0.022
Clock × Wave	1	2014	17.23	<0.001
Type × Grade	3	76	0.95	0.421
Type × Wave	1	156	0.02	0.893
Grade × Wave	3	263	2.89	0.036

Source of Variation	Degrees of Freedom Numerator	Denominator	F	p
Clock × Type × Grade	6	2034	1.34	0.236
Clock × Type × Wave	1	2014	0.07	0.794
Clock × Grade × Wave	3	2013	0.12	0.949
Type × Grade × Wave	3	276	0.64	0.589

Figure 1 displays the data analyzed by the ANOVA as median IRT RC scores (θ estimates) and their distributions for all Grades, Waves, and Clock groups (Tests 1, 2, and 3), separately for CATs and FITs. The numbers in the figure are the number of students in each distribution for the FITs (blue) and CATs (orange). Median RC scores were higher for CATs across all three tests in four of six comparisons, although the difference did not reach statistical significance. The distributions of scores differed between the two test types. Specifically, the vast majority of FIT distributions display truncation at the lower end and in many cases a similar truncation was observed at the upper end of the score distributions, indicating an inability to measure students with scores at the lower and upper ends of the θ distribution. The truncation was especially problematic for the Wave 2 group of students who were of lower reading ability; the distributions show that the FIT was unable to provide scores for students who were below approximately $\theta = -2$. By contrast, CAT was able to provide scores for students with θs as low as −3; For almost all tests and grades, CATs were also able to measure students whose scores were above the scores at which FITs were truncated, thereby providing better capability of measuring students at both ends of the score distributions. In addition there were differences between FIT and CAT score distributions. For a number of the Wave 1 tests, FITs resulted in score distributions that were bimodal, whereas CAT scores tend to be more evenly distributed. For the Wave 2 group, FIT scores for Grades 3 and 4 were highly peaked with long upper tails in comparison to CAT scores.

Figure 1. Medians and Distributions of RC θ Estimates by Grade, Wave, and Clock (Test 1, 2, 3) for FITs and CATs



HLM Results. The HLM regression-based analysis provides results that permit further analysis of the data from the ANOVA. In particular, they allow analysis of the categories involved in the categorical main effects and interactions by treating each category of a variable as a separate variable in the regression. Figure 2 displays the model-predicted means from the HLM analysis. Comparison of Figure 2 with Figure 1 shows that the model-predicted means are similar to the observed medians, but as is expected are all regressed somewhat. However, the similarity between the model-predicted means and the observed medians indicates that the HLM model was a good fit to the data.

Figure 2. Model-Predicted Means and Their Standard Errors from the HLM Analysis

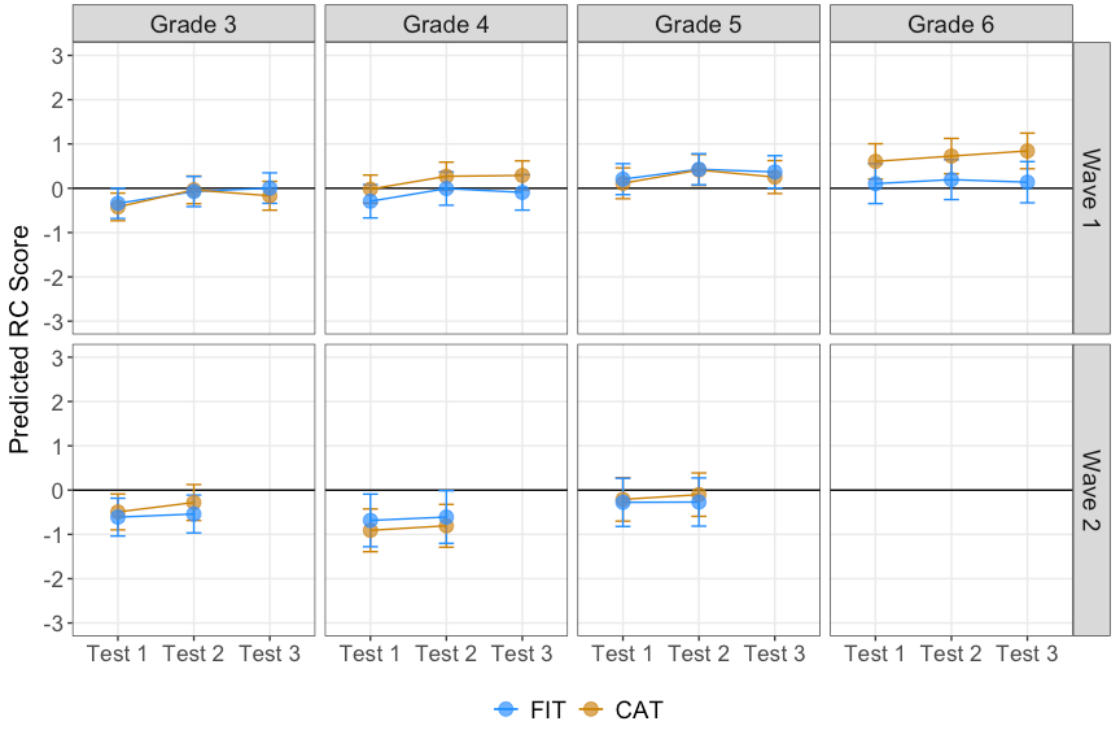


Table 4 displays the regression coefficients (β) and their confidence intervals. The analysis also provides a term for each level of a categorical variable. The intercept represents the RC score at first test—Fall for Wave 1 or Winter for Wave 2 for the CAT group. A piecewise growth model was used (i.e., one growth term for Time 1 to Time 2 and a second growth term for Time 2 to Time 3). The following acronyms are used in the table:

- **Clock C1** is the growth in RC scores from Test 1 to Test 2 for the CAT group.
- **Clock C2** is the change from Test 1 to Test 3 for the CAT group.
- **Type FIT** is the difference in initial average reading comprehension between the CAT and FIT group.
- **Clock C1 × Type FIT** is the difference in Test 1 to Test 2 change between CAT and FIT groups

- **Clock C2 × Type FIT** is the difference in Test 1 to Test 3 change between CAT and FIT groups

Table 4 shows significant βs only for C1 ($\beta = 0.20$) and C2 ($\beta = 0.23$), Grade 4 ($\beta = -0.42$) and Grade 6 ($\beta = 0.85$). These effects were further analyzed using contrasts, with results displayed in Table 5. Averaged over test types and other variables, Test 1 to Test 2 growth (C1 – C0) was significant at $p < 0.001$, with a model-estimated mean growth in RC scores of $\widehat{\theta} = 0.170$. The same contrast within Wave 1 showed significant mean growth of $\widehat{\theta} = 0.246$, but the same growth contrast in the Wave 2 group had non-significant mean growth of 0.094. When analyzed using contrasts within test type within the Wave 1 group, the C1 – C0 contrast was significant for both CAT and FIT, with mean growth of $\widehat{\theta} = 0.274$ for CAT and 0.217 for FIT. But within the Wave 2 group the same contrast (C1 – C0) resulted in mean growth of 0.134 for CAT and 0.54 for FIT, neither of which was statistically significant.

Table 5 also shows model-estimated growth contrasts for Time 1 to Time 3 (C2 – C0) and Time 2 to Time 3 in the Wave 1 group (the Wave 2 group did not test at Time 3). Combining the data from CAT and FIT, the model-estimated growth from Time 1 to Time 2 (C2 – C0) was 0.210, which was significant at $p < 0.001$, and estimated growth from Time 1 to Time 3 (C2 – C0) was also significant. When examined separately by test type, both CAT and FIT showed significant ($p < 0.001$) growth from Time 1 to Time 2 (C1 – C0) and Time 1 to Time 3 (C2 – C0), but no significant growth from Time 2 to Time 3. CAT growth from Time 1 to Time 3 was $\theta = 0.234$ and that for FIT was $\theta = 0.185$. In all cases, however, the difference in growth between CAT and FIT was not significant, even though CAT growth was higher.

It is noteworthy that the similarity between CAT and FIT results was achieved with substantial differences in both numbers of items used by the two test types, and the amount of

time required by the students. Table 6 displays average testing time and average number of items used by grade. The results show that at each grade the FIT required from 24.2 to 24.8 items on average, with a combined mean of 24.5 items. By contrast, the CAT required from an average of 12.0 to 12.4 items with a combined average of 12.2 items—a 50% reduction in number of items administered. Similarly, the CAT resulted in a 40.6% reduction in testing time—a combined mean of 15.9 minutes for the CAT versus a mean of 25.4 minutes for the FIT, with the difference in the two reduction percentages possibly due to the increased difficulty of the CATs as a result of the adaptive process.

Table 4. HLM β Regression Coefficients, 95% Confidence Intervals (CI), and *p*-Values

Characteristic	β	95% CI[1]	*p*
(Intercept)	-0.04	-0.24, 0.16	0.70
Clock			
C0	—	—	
C1	0.20	0.11, 0.29	<0.001
C2	0.23	0.16, 0.31	<0.001
Type			
CAT	—	—	
FIT	-0.10	-0.46, 0.25	0.57
Grade			
Grade 4	-0.42	-0.71, -0.14	0.004
Grade 5	-0.01	-0.31, 0.29	0.95
Grade 6	0.85	0.40, 1.3	<0.001
Wave			
Wave 1	0.22	-0.11, 0.55	0.19
Clock × Type			
C1 × FIT	-0.07	-0.16, 0.03	0.16
C2 × FIT	-0.05	-0.16, 0.07	0.41
Clock × Grade			
C1 × Grade 4	-0.01	-0.12, 0.11	0.90
C2 × Grade 4	0.08	-0.05, 0.20	0.24
C1 × Grade 5	0.00	-0.12, 0.12	>0.99
C2 × Grade 5	-0.09	-0.24, 0.05	0.21

Characteristic	β	95% CI[1]	p
C1 × Grade 6	-0.09	-0.32, 0.15	0.47
C2 × Grade 6	0.00	-0.13, 0.13	0.98
Clock × Wave			
C1 × Wave 1	0.14	-0.04, 0.32	0.13
Type × Grade			
FIT × Grade 4	0.08	-0.40, 0.56	0.75
FIT × Grade 5	0.12	-0.38, 0.61	0.64
FIT × Grade 6	-0.28	-1.2, 0.62	0.54
Type × Wave			
FIT × Wave 1	-0.10	-0.72, 0.53	0.77
Grade × Wave			
Grade 4 × Wave 1	0.67	0.17, 1.2	0.009
Grade 5 × Wave 1	0.10	-0.41, 0.62	0.69
Grade 6 × Wave 1	-0.63	-1.4, 0.10	0.090
Clock × Type × Grade			
C1 × FIT × Grade 4	0.05	-0.08, 0.18	0.43
C2 × FIT × Grade 4	-0.06	-0.26, 0.14	0.55
C1 × FIT × Grade 5	-0.02	-0.15, 0.11	0.77
C2 × FIT × Grade 5	0.07	-0.13, 0.28	0.49
C1 × FIT × Grade 6	0.03	-0.14, 0.20	0.75
C2 × FIT × Grade 6	-0.15	-0.36, 0.05	0.14
Clock × Type × Wave			
C1 × FIT × Wave 1	0.02	-0.16, 0.21	0.79
Clock × Grade × Wave			
C1 × Grade 4 × Wave 1	0.05	-0.15, 0.25	0.64
C1 × Grade 5 × Wave 1	0.05	-0.15, 0.26	0.61
C1 × Grade 6 × Wave 1	-0.13	-0.59, 0.33	0.57
Type × Grade × Wave			
FIT × Grade 4 × Wave 1	-0.40	-1.3, 0.46	0.36
FIT × Grade 5 × Wave 1	0.26	-0.58, 1.1	0.55
FIT × Grade 6 × Wave 1	-0.14	-1.7, 1.4	0.86

Table 5. Contrast Estimates from the HLM Analysis by Type and Wave, Their Standard Errors (SE), Degrees of Freedom (df), Z Ratios (Estimate/SE), and *p* Values

Contrast	Type	Wave	Estimate	SE	df	Z	*p*
Test 1 to Test 2 growth, averaged over test Type							
C1 - C0	--	--	0.170	0.043	Inf	3.980	< 0.001
Test 1 to Test 2 (C1 – C0), Test 1 to Test 3 (C2 – C0), and Test 2 to Test 3 (C2 – C1) growth by Wave							
C1 – C0	--	1	0.246	0.023	Inf	10.554	<0.001
C2 – C0	--	1	0.210	0.029	Inf	7.179	<0.001
C2 – C1	--	1	-0.036	0.030	Inf	-1.210	0.447
C1 – C0	--	2	0.094	0.082	Inf	1.146	0.486
C1 – C0	CAT	1	0.274	0.032	Inf	8.600	<0.001
C2 – C0	CAT	1	0.234	0.040	Inf	5.905	<0.001
C2 – C1	CAT	1	-0.040	0.040	Inf	-0.991	0.582
C1 – C0	FIT	1	0.217	0.034	Inf	6.403	<0.001
C2 – C0	FIT	1	0.185	0.043	Inf	4.319	<0.001
C2 – C1	FIT	1	-0.032	0.043	Inf	-0.731	0.745
C1 – C0	CAT	2	0.134	0.086	Inf	1.560	0.263
C1 – C0	FIT	2	0.054	0.097	Inf	0.552	0.846

Table 6. Mean and *SD* of Number of Items and Testing Time (in Minutes) by Grade and Test Type

Variable and Test	Grade 3	Grade 4	Grade 5	Grade 6	Total Group
No. of Items					
FIT					
Mean	24.2	24.7	24.5	24.8	24.5
SD	2.8	1.6	2.2	1.4	2.2
CAT					
Mean	12.0	12.3	12.1	12.4	12.2
SD	2.2	2.2	2.2	2.2	2.2
TOTAL	18.1	17.6	18.3	17.7	17.9
Testing time					
FIT					
Mean	23.0	24.8	27.6	27.6	25.4
SD	12.9	14.3	20.4	13.8	15.9
CAT					
Mean	15.0	15.3	15.7	14.0	15.1
SD	9.8	10.1	11.1	11.9	10.6
TOTAL	19.0	19.3	21.6	19.8	19.9

Discussion and Conclusions. The ANOVA analysis resulted in significant main effects for the Clock variable, indicating significant growth in RC scores across the three testing occasions. The HLM analysis allowed this effect to be further analyzed. The results showed that combining across test type, average RC growth from Test 1 to Test 2 was estimated at $\hat{\theta} = 0.17$. When the results were analyzed within Wave, the average growth was not significant for the Wave 2 group, but was significant for Wave 11, with an estimated $\hat{\theta} = 0.246$.

There was also a significant effect for the Wave variable, indicating that the second group (Wave 2) of students had significantly lower reading comprehension scores than the first group (Wave 1) of students, and a significant interaction action between Clock and Wave. The latter result reflected the observed growth in RC scores in the Wave 1 group but not in the Wave 2

group. This result might have been the results of when the Time 1 data were collected: In the Wave 1 group, the Time 1 tests were obtained during the fall months of the year (September-December), whereas for the Wave 2 group, the Time 1 tests were obtained later in the school year, during the months of January and February. This hypothesis is supported by the results of the HLM analysis that showed significant increases in RC scores from Test 1 to Time 2 in the Wave 1 group, but not Time 2 to Time 3, with the latter period roughly corresponding to the Time 1 to Time 2 time period in the Wave 2 group.

Results also showed that the two types of tests—FIT and CAT—were able to detect growth in reading comprehension using the MOCCA RC scores. Gains in reading comprehension were significant between the initial testing and the second and third test, but there were no significant gains in average RC scores between the second and third tests. Although there were no significant differences in mean RC scores between the FITs and the CATs, CAT growth estimates were consistently higher than those obtained from FITs, and were achieved with substantial reductions in both testing time and number of items administered to students.

Individual Growth

The previous section analyzed the RC growth data from MOCCA at the group level. This section reanalyzes the same data by examining change for individual students using a recently developed method called Adaptive Measurement of Change (AMC; Kim-Kang & Weiss, 2007, 2008; Weiss & Kingsbury, 1984). This method was developed within the context of CAT, but is also applicable to IRT-based scores derived from FITs—the MOCCA RC analyses presented below represent a first application of AMC to FITs. It is expected that the method will perform better for CATs as compared to FITs because of the higher levels of measurement precision obtainable using CATs.

The AMC procedure was first applied to the same group of students who were analyzed at the group level by ANOVA and HLM over three testing occasions. A smaller group of students completed the MOCCA on four or five testing occasions, extending the AMC analysis across those testing occasions—this group is referred to as the "progress monitoring" group to evaluate MOCCA as a potential tool for monitoring student reading comprehension over longer time periods, providing data designed to inform instruction at each testing occasion.

The AMC Procedure. AMC is designed to identify significant intraindividual change in θ estimates across two or more testing occasions. AMC uses a form of null hypothesis significance testing to determine whether differences in an individual's θ estimates at multiple testing occasions measuring the same trait are greater than would be expected by chance. Using AMC, individuals' θ estimates are analyzed to determine whether their observed change is "psychometrically significant" (Kim-Kang & Weiss, 2007, 2008; Weiss & Kingsbury, 1984). Although typical tests of significance are based in statistical theory to examine $\widehat{\theta}$ changes at the group level, based on assumptions of random sampling from a population of individuals, AMC uses psychometric theory to examine changes in θ estimates for a single individual measured at multiple occasions using items drawn by CAT from an item bank (Wang et al., 2021).

AMC research has demonstrated that the method can accurately identify significant intraindividual change across numerous longitudinal measurement designs, including across two (Finkelman et al., 2010; Lee, 2015) or more (Cooperman et al., 2021; Phadke, 2017) testing occasions, with unidimensional (Cooperman et al., 2021; Finkelman et al., 2010; Lee, 2015; Phadke, 2017) and multidimensional (Wang & Weiss, 2018; Wang et al., 2021) CATs. It has also been demonstrated to be effective in the presence of item parameter estimation error (Cooperman et al., 2021) and in detecting various patterns of individual change (Tai et al., 2023).

AMC has demonstrated higher power to detect significant change with larger changes in θ (Finkelman et al., 2010; Wang et al., 2021) and using banks with higher bank information (Cooperman et al., 2021). Multiple significance tests have been evaluated to use in AMC (Finkelman et al., 2010; Lee, 2015; Phadke, 2017; Wang & Weiss, 2018; Wang et al., 2021), but the likelihood ratio test has been identified as the best of the methods evaluated because it maintains a nominal error rate while providing the best rate of identifying true significant change (e.g., Cooperman et al., 2021; Finkelman et al., 2010; Wang et al., 2021). The multidimensional version of AMC is currently in use in a medical environment measuring change in hospitalized patients' reported symptoms, to help direct medical treatment by evaluating symptom change over time (Weiss et al., 2021).

The Likelihood Ratio Test. The LRT statistic used in AMC is the ratio of the likelihood of observing the response patterns under the null hypothesis (H_0: $\theta_1 = \theta_2 = \cdots = \theta_t$) over the likelihood of observing the response patterns under the alternative hypothesis (H_a: at least one of the equal signs does not hold). The denominator of the ratio is the product of the separate likelihoods evaluated at the corresponding θ estimate (Finkelman et al., 2010; Phadke, 2017). Formally, the LRT test statistic is defined as

$$\Lambda_O = \frac{L\left(\boldsymbol{u}_{1+2+\cdots+t}\middle|\hat{\theta}_{\text{Pool}_t}\right)}{L\left(\boldsymbol{u}_1\middle|\hat{\theta}_1\right) \times L\left(\boldsymbol{u}_2\middle|\hat{\theta}_2\right) \times \ldots \times L\left(\boldsymbol{u}_t\middle|\hat{\theta}_t\right)}, \tag{1}$$

where $\hat{\theta}_{\text{Pool}_t}$ is the maximum likelihood estimate of θ under H_0, $\boldsymbol{u}_{1+2+\cdots+t}$ is the combined or pooled response vector across *t* testing occasions, $\boldsymbol{u}_i$ is the response vector at testing occasion T_i $(i = 1, \ldots, t)$, and $L(\cdot)$ is the likelihood of a response vector for the 3PL model evaluated at a $\hat{\theta}$ value. Under the null hypothesis, $-2\log_e \Lambda_O$ follows a chi-square distribution with $(t-1)$ degrees of freedom. Because the likelihoods can be very small numbers

below zero, Equation 1 is estimated using its log version, as shown below for three testing occasions:

$$\mathrm{LRT}_3 = -2\left\{LL\left(\hat{\theta}_{pooled}\,|\,\mathbf{u}_{1+2+3}\right) - \left[LL\left(\hat{\theta}_1|\mathbf{u}_1\right) + LL\left(\hat{\theta}_2|\mathbf{u}_2\right) + LL\left(\hat{\theta}_3|\mathbf{u}_3\right)\right]\right\}. \quad (2)$$

Equation 2, with 2 degrees of freedom, was used to evaluate overall significant change across the three testing occasions for the total group of N = 1,434 students for both FITs and CATs.

Individual Change Across Three Testing Occasions. Table 7 shows the proportion of students with psychometrically significant change across the three testing occasions. The results show that for CAT the proportion of significant change varied somewhat by grade from 0.25 in Grade 3 to 0.23 in Grade 6, with an average proportion of 0.23. For FIT, proportion of significant change was highest in grade 4 (0.22) and constant at 0.19 in the other three grades; average proportion for FIT was 0.197.

Table 7. Proportion of Students with Psychometrically Significant Change Across Three Testing Occasions

Test Type	Grade 3	Grade 4	Grade 5	Grade 6
CAT	0.25	0.22	0.22	0.23
FIT	0.19	0.22	0.19	0.19

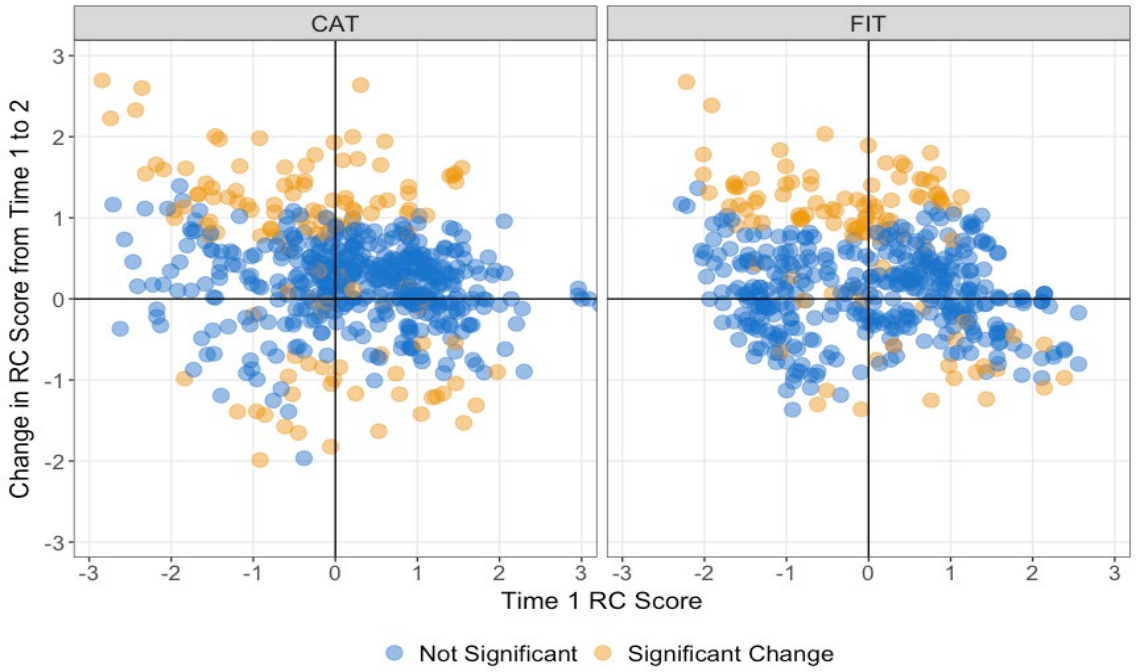
Figure 3 displays changes in θ estimates for individual students as a function of their initial (Time 1) θ estimates for the Wave 1 group (the testing occasions for the Wave 2 group are labeled Time 2 and Time 3 to coordinate with the time in the school year when the data were collected in that group), separately for the CAT and FIT groups; these figures indicate for each student whether the change plotted on the vertical axis was significant based on the LRT. Figure

3a provides results for changes from Time 1 to Time 2. In these figures, plotted points above the horizontal line reflect positive change and points below that line reflect negative change. Points to the right of the vertical line are for students whose θ estimates at Time 1 were above the mean of zero; those to the left of the line had initial θ estimates that were below zero. The results first reflect the greater range in θ estimates for the CAT group versus the FIT group. They also reflect more significant change in the CAT scores than FIT scores, including the identification of greater numbers of significant negative change as well as more significant positive change. For both FIT and CAT, the majority of change—and significant change—was positive, as expected (negative significant change will be discussed further, below). It is notable that for both FIT and CAT, significant change of almost three standard deviations was observed for some students, with higher levels of large magnitude change primarily for students who Time 1 θ estimates were below average.

Figure 3b shows the results for change from Time 1 (winter) to Time 3 (spring) for the Wave 1 group. As the figure shows, significant change for the FIT group and the CAT group were very similar in distribution and magnitude, except that the CAT group identified a few more students who had very large significant negative change.

Figure 3. Changes in Estimated θ From Time 1 to Time 2 for CATs and FITs for Individual Students for the Wave 1 Group

a. Time 1 to Time 2 Change



b. Time 1 to Time 3 Change

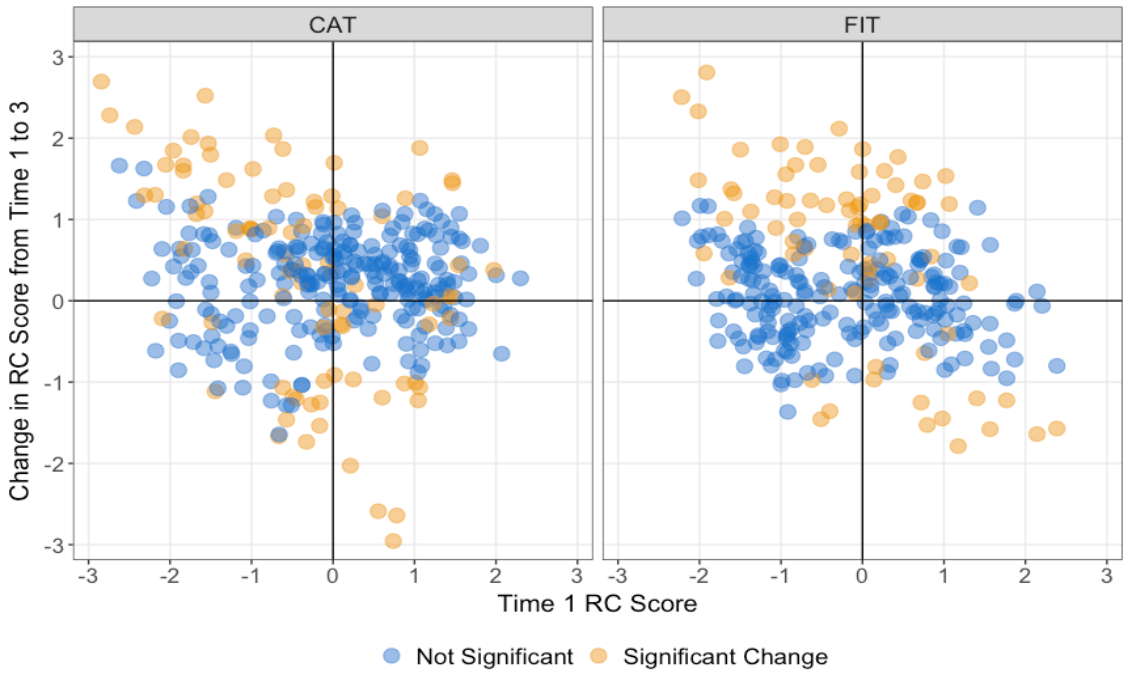
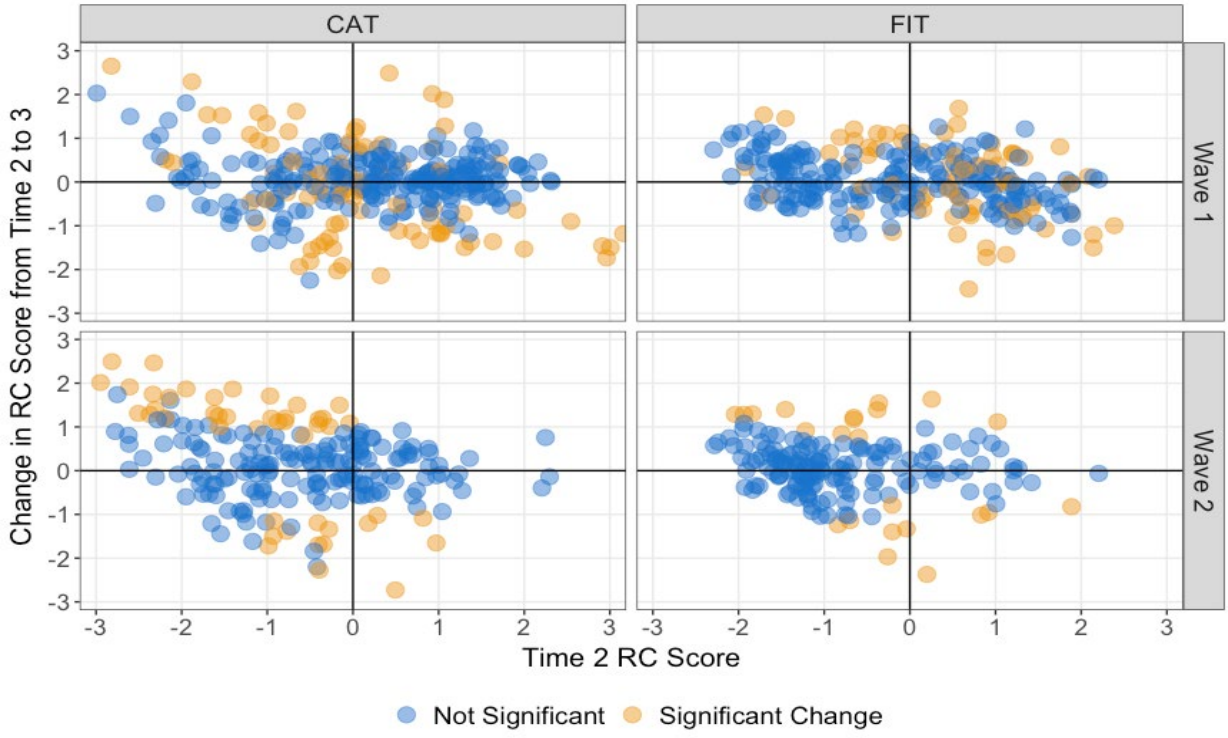


Figure 4 provides a comparison of individual change for both CAT and FIT for the Wave 1 and Wave 2 groups from Time 2 (winter) to Time 3 (spring). Although the ANOVA and HLM analyses indicated that there was not significant group change in RC scores across this time period, the AMC analyses shows that there were individual students whose θ estimates changed positively by as much as 1 to almost 3 standard deviations over that same time period, with those large significant changes occurring primarily within the CAT group. There were, however, also some significant negative changes in that group that would have reduced the mean change estimates observed in the group analyses. The magnitudes and patterns of change were similar in both the Wave 1 and Wave 2 groups, with the largest amount of significant positive change observed for students whose θ estimates at Time 1 were below average in the Wave 1 group. Appendix D provides descriptive statistics for groups demonstrating significant and not significant change in RC for various combinations of Wave, Time Interval, Test Type, and Grade.

Figure 4. Changes in Estimated θ From Time 2 to Time 3 for CATs and FITs for Individual Students for Wave 1 and Wave 2 Groups



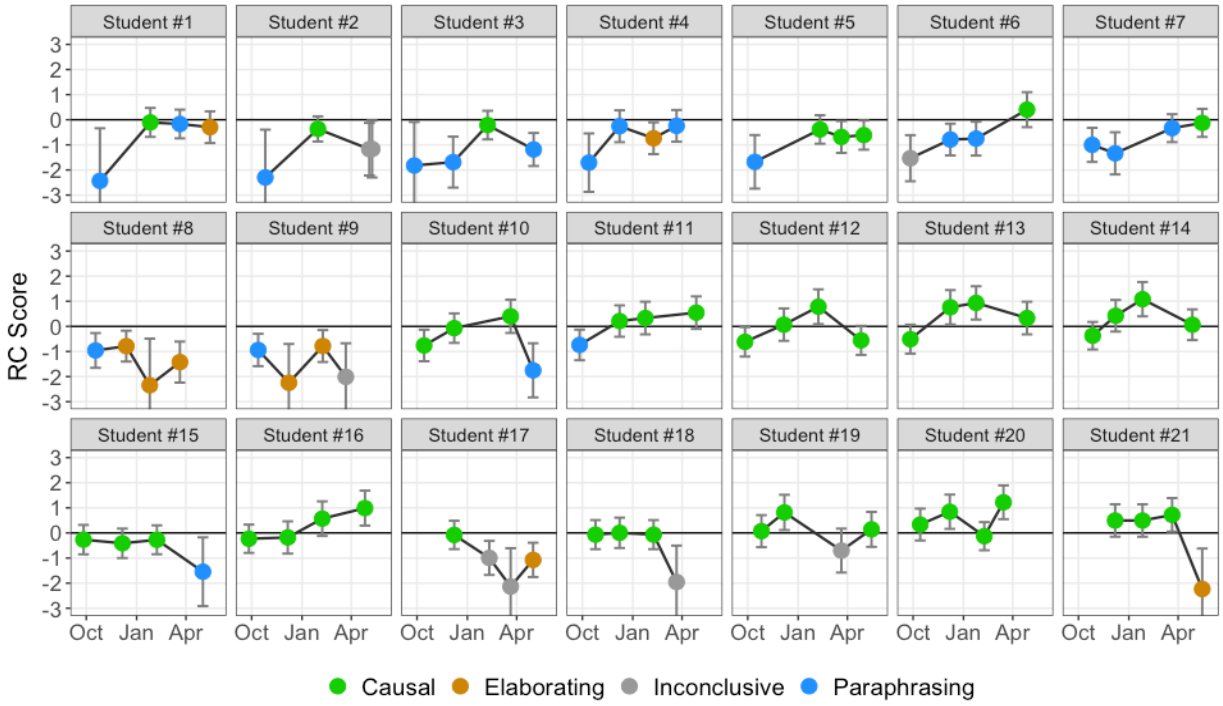
Progress Monitoring

Figure 5 shows plots of individual change for the 21 students who had significant LRTs across four testing occasions. In addition to plotting their θ estimates (plus a two SEM confidence interval around each), each plotted point is identified by their Dimension 2 PP classification (causal, elaborating, inconclusive, or paraphrasing) to illustrate how those classifications changed over time, along with the θ estimates from the RC dimension. As Figure 5 shows, there were many different psychometrically significant patterns of change across the four tests. Three students (Students 6, 11, and 16 had essentially linear growth. Students 1, 4, and 7 had a single growth period and then remained at a high θ level. The remainder of the students demonstrated mixed growth patterns, with some increases in score followed by

decrease, or vice-versa. Notable in this group are students 10, 15, 18, and 21 whose RC scores dropped significantly on the fourth testing occasion.

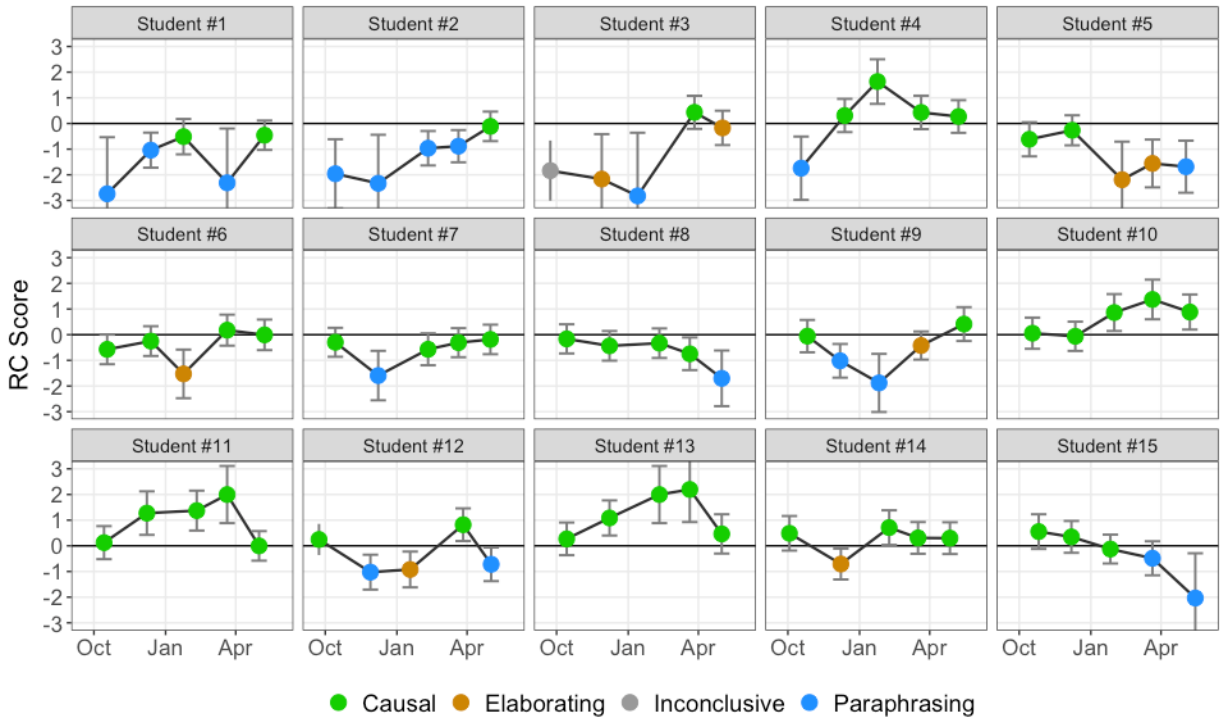
PP classifications also demonstrated a wide range of patterns. Students 12, 13, 14, 16, and 20 functioned as Causal for all four tests, even though their RC change was significant. Students 8 changed from a Paraphraser to an Elaborator, while Students 10, 15, and 21were Causal for the first three tests then changed to a Paraphraser, Paraphraser, and Elaborator, respectively as their Time 4 θ estimate dropped. Student 3 started as a Paraphraser before moving to Causal as their Time 3 θ estimate increased, then moved back to Paraphraser as their Time 4 θ estimate decreased. Student 5 started as a Paraphraser before moving to Causal at Time 2 as their θ estimate increased, and in contrast to Student 3, maintained that higher θ estimate and Causal classification across the remaining testing occasions. Other students showed mixed patterns of PP classification as their RC θ estimates changed.

Figure 5. RC θ Estimates and PP Classifications for Progress Monitoring Students With Significant LRTs From the AMC Analysis Across Four Tests



Another group of students completed the MOCCA CAT at five occasions. Their SC θ estimates and PP classifications at each testing occasion are shown in Figure 6. In this group there were no students whose θ estimates increased linearly over time—the closest was Student 2 whose RC scores were stable at Time 1 and Time 2, increased at Time 3 and remained the same at Time 4, then increased again at Time 5. Several students (4, 11, and 13) showed steady increases in scores for the first three or four occasions, then a drop in scores for Test 5. Student 15's scores decreased consistently from Test 1 through Test 5. As for the 4-occasion group, PP classifications changed for most students over time in various patterns.

Figure 6. RC θ Estimates and PP Classifications for Progress Monitoring Students With Significant LRTs From the AMC Analysis Across Five Tests



Addressing Significant Negative Change

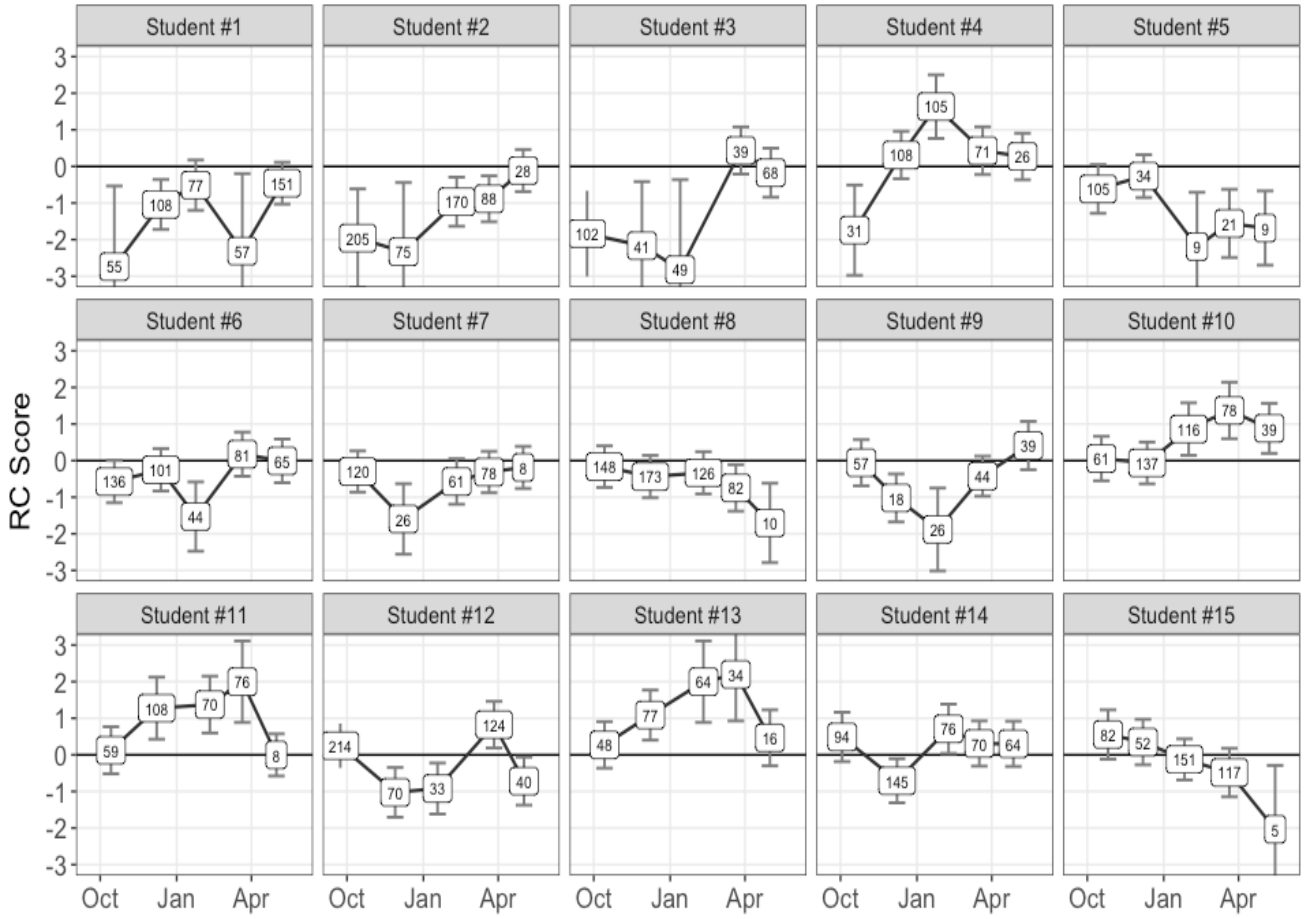
When measuring an achievement variable such as reading comprehension, it can be expected that student scores will increase over time, remain essentially constant, or decrease some due to measurement error. But the results obtained from both the larger sample of students and the two progress monitoring groups demonstrated that there are students whose score patterns over time demonstrated significant negative change. In some cases, negative change might occur for a student if on a given testing day they were ill or becoming ill, but given the frequency with which significant negative change occurred in these data, illness is unlikely to account for the observed degree and magnitude of negative change.

Research by Wise (2023) has hypothesized that negative change in test scores might be the result of inattentive responding. Their research has demonstrated that inattentive or unmotivated responding can be identified by analyzing the response times of a student within a particular test, and that using methods based in IRT a student's score can be recalculated (and

improved) after eliminating items that are responded to more quickly than others. Based on this research, the results for the 15 progress monitoring students who completed five CAT MOCCAs were reanalyzed by calculating average item response time for each of the five tests completed. The results are reported in Figure 7.

Some of the results suggest that unmotivated responding might be responsible for some instances of negative change or spurious growth. For example, Student 1's two lowest test scores were for Tests 1 and 4 (average 55 and 57 seconds per item) while their three higher test scores had substantially higher mean response times. Student 5 had longer response times for their first two tests (means of 108 and 34 seconds), then mean response times dropped to between means of 9 and 21 seconds concurrent with a drop in θ estimates of over a standard deviation for their three lower scores. For Students 6, 7, and 8 their single lowest score also had the lowest average response time, with the largest decrease for Student 8 whose longest average response time was an average of 173 seconds but the shortest (on their fifth test) was only 10 seconds. Finally, three students—Students 11, 13, and 15—had large decreases in θ estimates for Test 5, and in each case their mean response times were very short compared to response times in their first four tests, at 8, 16, and 5 seconds, respectively. It thus appears that some negative change can be attributed to unmotivated responding, as measured by average response times per item, for some students.

Figure 7. Average Response Time in Seconds Per Item for Each Test for 15 Students Who Completed Five Tests, Plotted at Their RC θ Estimate

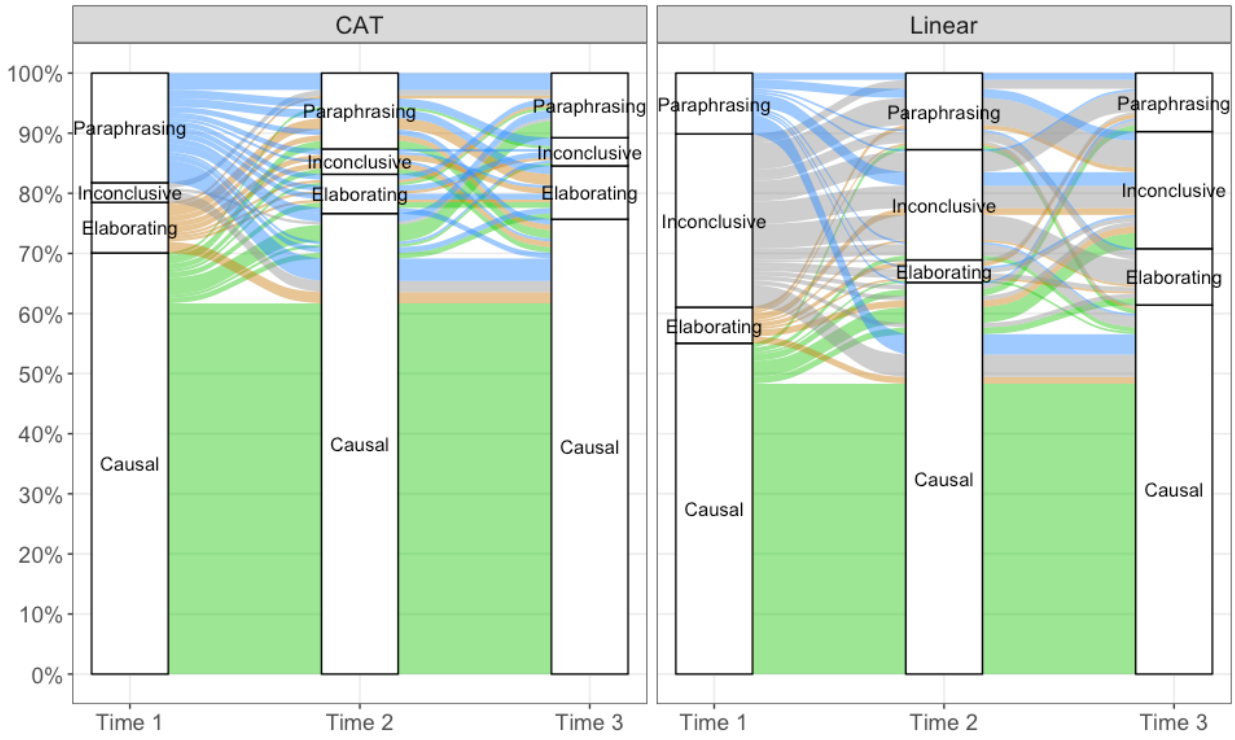


Process Propensity Classifications over Time

As shown in Figures 5 and 6, the PP classifications of students in the progress monitoring groups shifted over time in various patterns. To better understand how those classifications were changing over time, changes in the PP classifications were analyzed for a subset of the larger change sample over three testing occasions. Results are shown in Figure 8. This figure displays individual trajectories of change (or lack thereof) in PP classifications across the three testing occasions, comparing how the classifications changed using CATs and FITs. In these figures, the horizontal lines delineate the frequency areas for each type of classification. The colored lines reflect change (or lack of change) from an initial classification over time, and the thickness of the lines reflect the number of examinees with that change pattern.

Figure 8 shows that at all testing occasions, CAT identified more students as Causal than did FIT. Similarly, CAT had considerably fewer Inconclusives than FIT, and as a consequence more Elaborators and Paraphrasers than FIT. With regard to change patterns over time, CAT had more Causal students that did not change over time, with a smaller number of students whose classifications changed to other classifications. CAT also identified more students who were initially Elaborators who changed from Paraphrasing to Elaborating or Causal across the three tests, whereas for FIT the major group of Time 1 Paraphrasers changed to Inconclusive and remained in that category through Test 3. The majority of Time 1 Elaborators from CATs became Causal over time whereas that group for FITs generally became classified as Inconclusive.

Figure 8. Process Propensity Classifications for Students Across Three Testing Occasions for CAT and FIT Groups



Conclusions

References

Cooperman, A. W., Weiss, D. J., & Wang, C. (2021). Robustness of adaptive measurement of change to item parameter estimation error. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211033902

Davison, M. L., Biancarosa, G., Seipel, B., Carlson, S. E., Liu, B., & Kennedy, P. C. *Administration, Interpretation, and Technical Manual 2019: MOCCA Technical Report MTR-2019*. Eugene OR: University of Oregon.

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*(4), 238-254. https://doi.org/10.1177/0146621609344844

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica*, *26*(3), 297–307. https://doi.org/10.11613/BM.2016.034

Kim-Kang, G., & Weiss, D. J. (2007). Comparison of computerized adaptive testing and classical methods for measuring individual change. *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.iacat.org/biblio

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift Für Psychologie/Journal of Psychology, 216*(1), 49-58. Retrieved from www.iacat.org/biblio

Lee, J. E. (2015). *Hypothesis testing for adaptive measurement of individual change* (Doctoral dissertation). University of Minnesota, Minneapolis, MN.

National Center on Intensive Intervention. (n.d.). *Tools chart overview*. Intervention & Assessment Tools: Implementation Tools | NCII. https://intensiveintervention.org/tools-charts/overview

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods* (Doctoral dissertation). University of Minnesota, Minneapolis, MN.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2. ed). SAGE.

Tai, M. H., Cooperman, A. W., DeWeese, J. N., & Weiss, D. J. (2023). How do trait change patterns affect the performance of adaptive measurement of change? Journal of Computerized Adaptive Testing? *10(3)*. DOI 10.7333/2307-1003032

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*, *42*(3), 221–239. https://doi.org/10.1177/0146621617726787

Wang, C., Weiss, D. J., & Suen, K. Y. (2021). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research, 56*(3), 459-475. https://doi.org/10.1080/00273171.2020.1730739

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weiss, D., Wang, C., Cheville, A., & Basford, J. & DeWeese, J., (2021). Adaptive measurement of change: A novel method to reduce respondent burden and detect significant individual-

level change in patient-reported outcome measures. *Archives of Physical Medicine and Rehabilitation.* DOI: 10.1016/j.apmr.2021.07.814 (a)

Wise, S. (2023). Expanding the meaning of adaptive testing to enhance validity. *Journal of Computerized Adaptive Testing, 10*(2) ,22-31. DOI 10.7333/2305-1002022

Appendix A: Norm Tables

Table A.1. Percentile Ranks for Reading Comprehension Scale Scores for Grades 2 – 6

	2nd Grade		3rd Grade		4th Grade		5th Grade		6th Grade	
Percentile Rank	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit
1	50	89	50	89	50	108	50	115	50	152
2	90	120	90	120	109	137	116	137	153	180
3	121	128	121	136	138	144	138	158	181	235
4	129	134	137	154	145	176	159	183	236	255
5	135	140	155	172	177	183	184	229	256	267
6	141	149	173	181	184	194	230	240	268	274
7	150	154	182	185	195	206	241	263	275	290
8	155	167	186	192	207	212	264	272	291	291
9	168	172	193	199	213	221	273	279	292	297
10	173	179	200	206	222	229	280	283	298	311
11	180	185	207	208	230	233	284	303	312	329
12	186	189	209	215	234	239	304	309	330	350
13	190	193	216	217	240	254	310	317	351	360
14	194	196	218	226	255	257	318	324	361	380
15	197	201	227	229	258	259	325	344	381	395
16	202	203	230	234	260	266	345	349	396	414
17	204	206	235	238	267	277	350	352	415	422
18	207	209	239	242	278	281	353	359	423	425
19	210	211	243	248	282	297	360	364	426	430
20	212	219	249	256	298	309	365	374	431	437
21	220	223	257	261	310	320	375	382	438	446
22	224	235	262	266	321	327	383	388	447	462
23	236	238	267	270	328	335	389	395	463	475
24	239	240	271	275	336	342	396	410	476	480
25	241	241	276	281	343	347	411	417	481	488
26	242	244	282	286	348	351	418	422	489	513
27	245	245	287	291	352	357	423	424	514	519
28	246	247	292	300	358	364	425	427	520	525
29	248	249	301	303	365	376	428	429	526	531
30	250	252	304	309	377	383	430	432	532	532
31	253	253	310	315	384	391	433	436	533	533
32	254	258	316	321	392	400	437	440	534	535
33	259	260	322	324	401	402	441	441	536	538
34	261	262	325	334	403	412	442	446	539	541
35	263	264	335	340	413	417	447	452	542	543
36	265	269	341	344	418	423	453	456	544	560

	2nd Grade		3rd Grade		4th Grade		5th Grade		6th Grade	
Percentile Rank	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit
37	270	276	345	347	424	429	457	462	561	577
38	277	278	348	353	430	431	463	469	578	578
39	279	281	354	357	432	435	470	470	579	581
40	282	282	358	360	436	440	471	480	582	586
41	283	283	361	363	441	450	481	482	587	588
42	284	285	364	377	451	453	483	484	589	590
43	286	286	378	383	454	463	485	489	591	595
44	287	288	384	389	464	470	490	493	596	602
45	289	291	390	394	471	476	494	494	603	605
46	292	301	395	396	477	480	495	501	606	608
47	302	303	397	406	481	481	502	503	609	612
48	304	307	407	408	482	484	504	506	613	615
49	308	308	409	412	485	488	507	510	616	619
50	309	310	413	418	489	492	511	512	620	622
51	311	314	419	424	493	496	513	514	623	627
52	315	316	425	425	497	499	515	523	628	631
53	317	318	426	430	500	500	524	525	632	633
54	319	320	431	435	501	501	526	531	634	634
55	321	322	436	436	502	504	532	533	635	635
56	323	328	437	439	505	508	534	535	636	636
57	329	340	440	444	509	511	536	541	637	637
58	341	345	445	449	512	515	542	543	638	639
59	346	351	450	453	516	517	544	545	640	645
60	352	356	454	458	518	518	546	547	646	648
61	357	361	459	460	519	523	548	560	649	651
62	362	367	461	462	524	528	561	566	652	652
63	368	372	463	465	529	531	567	571	653	653
64	373	373	466	468	532	535	572	575	654	656
65	374	374	469	470	536	538	576	577	657	657
66	375	380	471	471	539	540	578	581	658	659
67	381	384	472	475	541	550	582	583	660	661
68	385	387	476	481	551	558	584	585	662	662
69	388	388	482	484	559	564	586	595	663	668
70	389	390	485	487	565	568	596	603	669	669
71	391	398	488	494	569	572	604	605	670	673
72	399	420	495	498	573	577	606	608	674	680
73	421	424	499	502	578	580	609	622	681	685
74	425	428	503	506	581	588	623	627	686	687
75	429	439	507	507	589	592	628	631	688	688
76	440	443	508	508	593	603	632	633	689	693

	2nd Grade		3rd Grade		4th Grade		5th Grade		6th Grade	
Percentile Rank	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit	L. Limit	U. Limit
77	444	450	509	510	604	607	634	635	694	699
78	451	462	511	512	608	611	636	637	700	701
79	463	465	513	519	612	621	638	642	702	709
80	466	469	520	527	622	627	643	645	710	710
81	470	476	528	531	628	631	646	646	711	711
82	477	482	532	533	632	634	647	654	712	712
83	483	485	534	537	635	635	655	657	713	714
84	486	492	538	546	636	639	658	659	715	719
85	493	499	547	555	640	649	660	670	720	722
86	500	506	556	565	650	656	671	671	723	723
87	507	512	566	578	657	661	672	675	724	727
88	513	517	579	584	662	666	676	684	728	732
89	518	523	585	590	667	674	685	694	733	733
90	524	539	591	609	675	679	695	697	734	734
91	540	546	610	620	680	696	698	715	735	735
92	547	549	621	630	697	709	716	717	736	736
93	550	555	631	634	710	717	718	734	737	747
94	556	570	635	643	718	720	735	745	748	749
95	571	577	644	655	721	721	746	760	750	760
96	578	583	656	666	722	751	761	781	761	781
97	584	612	667	696	752	757	782	796	782	796
98	613	635	697	719	758	775	797	822	797	822
99	636	950	720	950	776	950	823	950	823	950

Appendix B: Item Parameters: Reading Comprehension Dimension

Table B.1.

Three Parameter Logistic Item Parameters for the Reading Comprehensions Dimension with All Lower Asymptote Parameters Constrained to 0.24

Item	Discrimination	Difficulty	Asymptote
RG028	2.8495	-0.1485	0.24
21	1.945	-0.0635	0.24
417	2.6568	-0.4049	0.24
495	2.1583	-0.1065	0.24
GB023	1.5699	0.4305	0.24
272	2.5877	-0.7485	0.24
KH018	2.0988	0.91	0.24
RG089	1.936	0.1859	0.24
GB010	1.8433	-0.5224	0.24
57	1.9443	-0.2864	0.24
255	1.8423	0.3554	0.24
42	2.2431	-0.1656	0.24
98	2.159	-0.733	0.24
77	2.0429	-0.9663	0.24
GB015	2.4534	-0.5557	0.24
JB034	2.0665	-0.3553	0.24
58	2.0356	-0.6015	0.24
RG005	1.6439	-0.434	0.24
GB048	1.5623	0.6085	0.24
275	1.7177	-0.5014	0.24
217	2.5257	-0.48	0.24
60	1.9242	-0.1032	0.24
GB009	1.8841	0.0006	0.24
GB021	2.6216	-0.1653	0.24
451	2.7797	-0.4423	0.24
BS003	1.4683	0.6726	0.24
BS002	1.823	-0.2348	0.24
BS024	2.2563	-0.4945	0.24
BS001	1.7435	0.0231	0.24
BS016	2.0245	0.183	0.24
468	2.9672	-0.2676	0.24
81	1.8039	0.1423	0.24
6	2.0901	-0.9054	0.24
316	1.7389	-0.6339	0.24

Item	Discrimination	Difficulty	Asymptote
87	1.9669	-0.4587	0.24
GB039	2.6075	-0.5131	0.24
152	2.1534	0.2408	0.24
302	1.8066	-0.0386	0.24
GB019	1.7959	-0.1979	0.24
JB033	2.0864	0.2946	0.24
235	1.587	-0.4641	0.24
142	2.4898	0.0888	0.24
IE025	2.234	-0.4212	0.24
JB036	2.0347	-0.1563	0.24
IE003	1.6415	0.6825	0.24
151	1.8758	-0.6726	0.24
GB028	1.9536	0.9086	0.24
71	1.9968	-0.3727	0.24
GB018	1.944	-0.3497	0.24
54	2.4623	-0.2384	0.24
33	2.6722	-0.6145	0.24
452	1.9504	-0.4825	0.24
RG075	2.3815	-0.361	0.24
GB017	2.5236	0.0413	0.24
487	3.1834	-0.2713	0.24
BS008	1.566	-0.2036	0.24
BS005	2.1924	-0.616	0.24
BS019	2.1138	-0.0802	0.24
BS020	1.2828	0.8852	0.24
BS029	1.458	0.4008	0.24
253	2.8449	-0.2401	0.24
GB011	1.6796	-0.6447	0.24
241	1.7274	-0.0134	0.24
343	2.3142	0.2264	0.24
GB043	2.082	-0.4509	0.24
414	2.0652	-0.2028	0.24
177	2.3919	-1.0151	0.24
435	1.6332	-0.4127	0.24
KH003	1.6747	-0.3755	0.24
GB013	2.0266	-0.6503	0.24
319	2.3582	0.5288	0.24
140	2.1326	-0.4501	0.24
236	1.5649	0.6274	0.24
GB024	2.3474	-0.2583	0.24
RG008	2.0646	0.8136	0.24
110	2.4541	-0.2995	0.24

Item	Discrimination	Difficulty	Asymptote
GB026	2.1061	0.4065	0.24
174	1.9358	-0.353	0.24
39	2.0126	-0.3096	0.24
NM002	1.8935	-0.484	0.24
393	1.6456	-0.7988	0.24
184	2.1366	-0.0018	0.24
JB004	2.32	-0.4172	0.24
IE024	1.7364	0.3816	0.24
244	2.7817	-0.2744	0.24
BS004	1.836	-0.8659	0.24
BS007	1.8563	-0.1643	0.24
BS012	1.9426	-0.3654	0.24
BS009	1.5263	0.0899	0.24
BS028	1.9039	-0.5792	0.24
201	1.7343	0.4259	0.24
NM001	2.2528	-0.3831	0.24
361	2.1491	-0.0822	0.24
432	2.3812	-0.5481	0.24
301	1.9273	0.5264	0.24
294	2.229	-0.0007	0.24
97	2.1097	0.094	0.24
387	2.3589	-0.2374	0.24
NM005	2.3979	-0.1238	0.24
131	2.159	-0.6533	0.24
506	1.8316	-0.7797	0.24
193	2.0269	0.6626	0.24
IE055	1.6835	1.24	0.24
12	2.726	-0.7866	0.24
GB004	2.4189	0.1839	0.24
380	2.0784	-0.6922	0.24
IE022	2.3402	-0.0451	0.24
GB040	2.4104	0.0678	0.24
RG017	1.5708	0.9221	0.24
104	2.0072	0.3411	0.24
263	1.7464	-0.4482	0.24
GB047	1.9134	-0.1974	0.24
346	2.1423	-0.4241	0.24
BS031	1.4662	0.1468	0.24
BS021	1.8286	0.4671	0.24
BS014	1.6556	0.2893	0.24
370	2.7783	-0.2223	0.24
251	1.6526	-0.083	0.24

Item	Discrimination	Difficulty	Asymptote
211	2.1014	-0.6809	0.24
IE016	1.948	0.5248	0.24
IE075	2.3642	0.0337	0.24
JB043	1.836	1.3544	0.24
406	2.1949	-0.5599	0.24
JB030	1.9994	0.0598	0.24
284	1.665	0.7588	0.24
411	2.0094	0.2039	0.24
KH031	2.0505	-0.1835	0.24
225	1.6051	0.3882	0.24
467	1.98	-0.3065	0.24
396	1.6636	0.6049	0.24
408	1.6242	-0.0773	0.24
323	1.9147	0.2091	0.24
GB046	2.591	-0.736	0.24
416	1.8792	-0.7467	0.24
KH073	1.8913	-0.3627	0.24
250	1.9718	-0.0993	0.24
GB041	2.28	-0.1484	0.24
50	2.3948	-0.3994	0.24
JB042	1.7544	-0.0552	0.24
35	2.1162	-0.4789	0.24
BS025	1.3583	0.7164	0.24
BS033	1.913	-0.0515	0.24
BS017	1.295	0.9308	0.24
KH064	2.8222	-0.1276	0.24
107	1.6487	0.2894	0.24
182	2.2115	-0.4471	0.24
KH078	1.6832	0.8148	0.24
RG068	2.1068	-0.0014	0.24
369	2.277	-0.825	0.24
326	1.9292	-0.2654	0.24
27	2.08	-0.6473	0.24
332	2.0007	0.441	0.24
256	1.5033	0.1858	0.24
132	2.3416	0.005	0.24
260	2.787	-0.798	0.24
JB037	1.6997	0.2891	0.24
130	2.0364	-0.1141	0.24
126	2.0332	-0.0618	0.24
321	1.8675	-0.2908	0.24
JB028	2.1131	0.0754	0.24

Item	Discrimination	Difficulty	Asymptote
378	1.8865	1.0944	0.24
RG061	2.3256	-0.0614	0.24
262	2.2816	0.6166	0.24
NM019	2.1869	-0.3031	0.24
KH001	1.6349	0.1366	0.24
395	2.0733	-0.5353	0.24
GB030	2.5009	-0.66	0.24
BS022	2.07	0.2692	0.24
BS030	1.9014	0.1725	0.24
BS010	1.5762	-0.3584	0.24
GB002	2.7023	0.6119	0.24
171	1.8513	-0.4955	0.24
198	1.9929	-0.9785	0.24
463	1.7575	0.3478	0.24
128	2.1155	-0.2871	0.24
310	1.9625	-0.134	0.24
164	2.1967	-0.2094	0.24
403	1.904	-0.6057	0.24
218	1.9507	0.2607	0.24
136	2.0303	-0.4179	0.24
JB014	2.1891	-0.1047	0.24
NM010	2.2316	-0.0002	0.24
205	1.8047	0.0681	0.24
498	1.6929	0.1794	0.24
GB049	1.929	0.8896	0.24
324	2.2374	-0.3567	0.24
GB034	1.9404	1.5327	0.24
IE041	1.7204	0.6481	0.24
499	1.7092	0.3692	0.24
RG003	2.0524	0.3001	0.24
JB005	2.1306	0.0886	0.24
494	1.9482	0.0689	0.24
464	1.735	-0.366	0.24
BS013	1.6301	-0.2589	0.24
GB222	1.3518	0.4286	0.24
293	2.5945	0.0893	0.24
283	1.948	0.0655	0.24
IE048	1.6833	0.7322	0.24
480	1.767	0.0457	0.24
485	2.2387	-0.5249	0.24
GB050	1.8195	0.9699	0.24
RG063	2.3582	0.3588	0.24

Item	Discrimination	Difficulty	Asymptote
195	1.7797	0.3579	0.24
376	1.5154	-1.0716	0.24
352	1.615	0.0284	0.24
KH062	2.2848	0.1945	0.24
309	1.7102	-0.7947	0.24
38	2.1651	-0.0753	0.24
265	2.0565	-0.1592	0.24
IE058	1.81	1.0142	0.24
147	1.777	-0.2512	0.24
JB076	2.1731	-0.3913	0.24
NM015	2.4482	-0.1096	0.24
KH023	2.21	0.7956	0.24
JB010	1.9021	-0.1745	0.24
374	2.446	-0.173	0.24
GB042	1.9514	0.7284	0.24
231	1.5545	0.1968	0.24
JB013	1.7464	-0.4418	0.24
NM011	2.8041	-0.1748	0.24
BS026	1.4774	-0.6565	0.24
BS023	1.7362	0.8275	0.24
113	2.6729	-0.0696	0.24
46	2.2787	-0.7152	0.24
IE034	2.191	0.5655	0.24
RG064	2.0422	0.2338	0.24
270	1.9922	0.4242	0.24
504	2.0407	0.0562	0.24
183	1.5558	-0.3412	0.24
124	1.7323	-0.3636	0.24
427	2.0553	-0.4181	0.24
331	1.8669	0.1288	0.24
105	3.23	0.8915	0.24
GB005	1.7746	0.7728	0.24
82	2.3666	-0.8085	0.24
192	1.5072	-0.0122	0.24
446	2.1808	-0.4766	0.24
420	1.8265	-0.401	0.24
NM009	1.9264	0.1766	0.24
230	1.7211	-0.1782	0.24
GB033	1.7989	1.1823	0.24
440	1.8992	0.0581	0.24
240	1.5606	0.1769	0.24
RG010	2.1012	0.2558	0.24

Item	Discrimination	Difficulty	Asymptote
JB041	1.9482	-0.015	0.24
RG070	1.8188	0.3946	0.24
GB221	2.097	-0.1897	0.24
GB223	1.8522	0.3345	0.24
GB032	2.602	0.3406	0.24
425	1.7858	-0.497	0.24
GB035	1.9917	1.6611	0.24
86	1.9502	0.0721	0.24
KH066	1.7984	0.5264	0.24
IE028	2.1383	0.7145	0.24
RG032	2.7569	-0.323	0.24
354	1.6709	-1.1633	0.24
IE026	2.3826	-0.0013	0.24
308	1.8977	0.2232	0.24
94	1.6937	-0.3276	0.24
13	2.3251	0.1739	0.24
160	2.0633	0.135	0.24
20	1.5997	-0.475	0.24
297	2.3939	0.185	0.24
GB036	1.6827	0.4249	0.24
349	1.5006	-0.115	0.24
500	1.632	-0.3455	0.24
KH054	2.0691	1.1974	0.24
190	1.8889	-0.1875	0.24
JB008	2.0274	0.4901	0.24
281	2.0686	0.966	0.24
288	2.0206	-0.6177	0.24
220	2.2253	0.0272	0.24
GB022	2.5879	0.1192	0.24
GB003	2.5323	0.2971	0.24
KH002	2.2391	-0.0123	0.24
GB025	1.7393	0.304	0.24
JB045	1.8071	0.429	0.24
188	2.2212	-0.2149	0.24
189	2.472	-0.3594	0.24
209	2.3164	-0.0707	0.24
RG024	1.8056	1.6065	0.24
41	2.2549	-0.383	0.24
17	2.2173	-0.9869	0.24
RG014	2.3161	0.3952	0.24
466	2.5493	-0.6199	0.24
GB007	1.9256	1.0572	0.24

Item	Discrimination	Difficulty	Asymptote
78	1.9848	0.2258	0.24
155	2.0082	0.1154	0.24
44	2.4456	-0.3472	0.24
429	1.5878	-0.826	0.24
481	2.3316	0.2409	0.24
371	1.9761	0.0938	0.24
RG004	1.8234	1.2081	0.24
471	1.6106	0.7521	0.24
IE037	2.4839	0.8844	0.24
138	2.33	-0.019	0.24
31	2.0397	-0.5315	0.24
IE032	2.5344	0.2885	0.24
RG054	2.6246	0.2702	0.24
26	2.0918	-0.3291	0.24
88	1.6665	0.461	0.24
479	1.7974	-0.3725	0.24
RG022	2.0381	0.8739	0.24
JB049	1.5288	0.5601	0.24
KH004	2.0648	0.7619	0.24
IE044	1.6895	1.4544	0.24
356	2.2969	-0.3523	0.24
259	1.7221	-0.2013	0.24
GB051	2.0635	0.416	0.24
IE089	1.5101	1.5751	0.24
493	2.0941	-0.8017	0.24
311	1.6935	-0.0223	0.24
419	1.7167	0.1759	0.24
327	1.5672	-0.032	0.24
290	2.1551	-1.0665	0.24
159	1.6085	-0.6622	0.24
KH027	2.3594	-0.4937	0.24
109	2.1779	0.1058	0.24
GB031	2.168	0.3688	0.24
GB006	2.2265	1.2777	0.24
51	2.3832	0.158	0.24
333	2.067	-0.5549	0.24
GB016	2.501	0.275	0.24

Appendix C: Item Parameters: Process Propensity Dimension

Table C.1. Two Parameter Logistic Item Parameters for the Process Propensity Dimension

Item	Discrimination	Difficulty	Asymptote
RG028	1.293	-0.0105	0
21	1.0707	0.7361	0
417	1.3347	-0.2905	0
495	0.9586	-0.1716	0
GB023	1.4452	0.3499	0
272	1.3865	0.0237	0
KH018	1.2641	-0.9064	0
RG089	1.471	0.0641	0
GB010	1.481	-0.3036	0
57	1.2124	-0.8955	0
255	1.0394	0.3286	0
42	1.7685	0.2943	0
98	1.3367	-0.1017	0
77	1.471	-0.0875	0
GB015	1.4139	0.0558	0
JB034	1.3022	-0.0501	0
58	1.216	0.0685	0
RG005	1.6767	-0.4579	0
GB048	1.6533	0.6728	0
275	1.2301	-0.2372	0
217	1.2573	-0.5158	0
60	1.1399	-0.7593	0
GB009	1.3391	0.0412	0
GB021	1.2815	0.024	0
451	1.2033	0.6264	0
BS003	0.9687	-0.1666	0
BS002	0.8609	0.0705	0
BS024	0.8122	-0.1746	0
BS001	0.9541	-0.5764	0
BS016	0.8738	0.2642	0
468	1.1871	-0.0257	0
81	1.1722	-0.433	0
6	1.217	0.2162	0
316	1.3037	0.1895	0
87	1.1195	0.0937	0
GB039	1.495	0.177	0
152	1.651	-0.0109	0
302	1.0134	-0.6387	0

Item	Discrimination	Difficulty	Asymptote
GB019	1.3274	-0.3696	0
JB033	1.4006	0.2666	0
235	1.1456	-0.096	0
142	1.113	-0.3086	0
IE025	1.6733	-0.0021	0
JB036	1.5006	-0.5186	0
IE003	1.5628	0.0051	0
151	1.2192	0.644	0
GB028	1.206	0.0614	0
71	1.0414	0.2398	0
GB018	1.5407	-0.051	0
54	1.2758	0.615	0
33	1.3195	-0.6492	0
452	1.0817	-0.2164	0
RG075	1.42	0.4685	0
GB017	1.2755	-0.1465	0
487	1.0635	0.27	0
BS008	0.8699	-0.5637	0
BS005	0.9586	-0.6628	0
BS019	0.8388	0.1528	0
BS020	0.933	0.2495	0
BS029	0.7019	-0.0346	0
253	1.3282	-0.7858	0
GB011	1.5586	0.0769	0
241	1.2835	0.0933	0
343	1.0163	0.0515	0
GB043	1.5057	-0.0683	0
414	1.0936	-0.1854	0
177	1.2929	0.3413	0
435	1.1152	0.4008	0
KH003	1.5314	-0.393	0
GB013	1.3896	-0.1058	0
319	0.9285	0.8571	0
140	1.3529	-0.5373	0
236	1.1767	-0.3023	0
GB024	1.5847	0.6069	0
RG008	1.3906	0.7555	0
110	1.1699	-0.2439	0
GB026	1.3503	-0.644	0
174	1.0887	0.2929	0
39	1.7051	0.1536	0
NM002	1.564	0.1237	0

Item	Discrimination	Difficulty	Asymptote
393	1.2514	0.0963	0
184	1.4724	-0.3521	0
JB004	1.7376	0.0075	0
IE024	1.2796	-0.5732	0
244	0.9476	0.0517	0
BS004	0.9697	-0.2597	0
BS007	0.8931	-0.0415	0
BS012	0.8234	-0.3842	0
BS009	0.8594	0.5579	0
BS028	0.9076	-0.0177	0
201	0.9052	0.4395	0
NM001	1.6878	-0.0545	0
361	1.1116	0.2154	0
432	1.1696	-0.276	0
301	1.0382	0.2694	0
294	0.9462	0.1363	0
97	1.0062	-0.1907	0
387	0.978	-0.448	0
NM005	1.3534	-0.1294	0
131	1.6816	0.0563	0
506	1.1604	-0.1309	0
193	1.2551	0.2723	0
IE055	1.19	0.8924	0
12	1.2327	-0.0993	0
GB004	1.3476	-0.1843	0
380	1.7556	-0.1885	0
IE022	1.4169	0.026	0
GB040	1.2796	0.0783	0
RG017	1.3466	-0.5619	0
104	1.1298	-0.1749	0
263	1.1723	0.6772	0
GB047	1.6787	0.1877	0
346	1.2855	-0.2915	0
BS031	0.9004	0.7477	0
BS021	0.7497	0.4291	0
BS014	0.872	-0.2006	0
370	1.283	0.7092	0
251	1.2934	0.2852	0
211	1.3418	-0.2489	0
IE016	1.2811	0.2653	0
IE075	1.3981	-0.198	0
JB043	1.2772	-0.2421	0

Item	Discrimination	Difficulty	Asymptote
406	1.4384	-0.4566	0
JB030	1.3245	-0.2118	0
284	1.196	0.1294	0
411	1.4151	0.4322	0
KH031	1.2152	-0.1835	0
225	0.987	0.1735	0
467	1.345	0.1877	0
396	1.0469	-0.7332	0
408	1.3263	-0.6627	0
323	1.4928	-0.2325	0
GB046	1.6116	0.4555	0
416	1.3641	-0.2361	0
KH073	1.6947	0.2619	0
250	1.0739	-0.1349	0
GB041	1.4826	-0.2719	0
50	1.2084	-0.5746	0
JB042	1.7552	0.109	0
35	1.1254	-0.1607	0
BS025	0.7575	-0.5548	0
BS033	0.8463	0.0044	0
BS017	0.9851	-0.5087	0
KH064	1.2833	0.5583	0
107	0.9129	-0.5805	0
182	1.2765	-0.0058	0
KH078	1.2017	0.4812	0
RG068	1.5164	0.1584	0
369	1.3129	0.1395	0
326	1.1121	0.1865	0
27	1.3133	0.0001	0
332	0.8255	0.3041	0
256	1.4132	-0.0835	0
132	0.9608	-0.2083	0
260	1.0994	-0.3327	0
JB037	1.749	0.1302	0
130	1.172	0.0173	0
126	1.0996	-0.3654	0
321	1.188	-0.0358	0
JB028	1.4664	-0.1859	0
378	0.9976	-0.7439	0
RG061	1.6221	-0.2372	0
262	0.8665	-0.1777	0
NM019	1.3102	0.3104	0

Item	Discrimination	Difficulty	Asymptote
KH001	1.2925	-0.717	0
395	1.4855	-0.1705	0
GB030	1.4748	-0.3458	0
BS022	0.8291	0.4772	0
BS030	0.9272	0.1062	0
BS010	0.8867	0.4576	0
GB002	1.4098	-0.0748	0
171	1.0914	0.8468	0
198	1.4951	0.1173	0
463	0.9399	0.3401	0
128	1.1305	-0.4819	0
310	1.2284	0.1268	0
164	1.4878	0.0381	0
403	1.2412	0.3719	0
218	0.8988	-0.606	0
136	1.1487	-0.025	0
JB014	1.5747	-0.2401	0
NM010	1.3212	0.2793	0
205	1.1376	0.1088	0
498	1.1208	-0.2125	0
GB049	1.3627	-0.2204	0
324	1.1973	-0.3255	0
GB034	1.0462	-0.3758	0
IE041	1.3887	0.3237	0
499	0.8882	0.2449	0
RG003	1.4042	-0.6922	0
JB005	1.4756	0.0438	0
494	0.9347	0.0006	0
464	1.6597	0.3595	0
BS013	0.8931	0.1474	0
GB222	0.8513	-0.049	0
293	1.0846	0.3518	0
283	1.3739	0.1096	0
IE048	1.2719	-0.2037	0
480	0.9231	0.1472	0
485	1.4102	-0.2562	0
GB050	1.5856	0.0682	0
RG063	1.3364	0.3353	0
195	0.8925	-0.4404	0
376	1.5524	0.0737	0
352	1.0253	-0.6893	0
KH062	1.3046	0.0475	0

Item	Discrimination	Difficulty	Asymptote
309	1.1278	0.4019	0
38	1.1174	0.2305	0
265	1.0537	-0.1796	0
IE058	1.2216	0.2898	0
147	1.1178	-0.6572	0
JB076	1.5222	0.0233	0
NM015	1.3246	0.2918	0
KH023	1.396	-0.1441	0
JB010	1.3022	-0.6645	0
374	1.1596	-0.2078	0
GB042	1.5116	0.3405	0
231	1.1026	-0.4573	0
JB013	1.7872	-0.0344	0
NM011	1.5783	0.3603	0
BS026	0.8623	-0.2294	0
BS023	0.9364	-0.1298	0
113	0.46	-0.5994	0
46	1.3121	-0.0562	0
IE034	1.3642	-0.2532	0
RG064	1.5773	0.2851	0
270	0.952	0.2039	0
504	1.352	0.2975	0
183	1.2402	-0.5367	0
124	1.4576	0.1943	0
427	1.2629	0.541	0
331	1.2577	0.2804	0
105	0.3646	1.499	0
GB005	1.1429	0.2802	0
82	1.1769	-0.6477	0
192	1.3071	0.6143	0
446	1.0661	0.1065	0
420	1.4144	0.1603	0
NM009	1.2929	0.7563	0
230	1.0537	0.6324	0
GB033	1.2325	0.2398	0
440	1.0378	-0.3827	0
240	1.0603	-0.1915	0
RG010	1.2597	0.1399	0
JB041	1.453	0.1643	0
RG070	1.4081	0.151	0
GB221	1.022	0.2737	0
GB223	0.8378	0.234	0

Item	Discrimination	Difficulty	Asymptote
GB032	1.1359	-0.0649	0
425	1.1521	0.0594	0
GB035	1.268	-0.6306	0
86	1.0989	0.3058	0
KH066	1.2235	0.0743	0
IE028	1.1987	-0.2487	0
RG032	1.4426	-0.0287	0
354	1.5397	-0.3416	0
IE026	1.4618	-0.108	0
308	1.1171	0.6298	0
94	1.0134	0.4094	0
13	1.6425	-0.3209	0
160	1.0576	0.3259	0
20	1.335	0.3976	0
297	0.9503	0.1703	0
GB036	1.1915	-0.1089	0
349	1.2735	0.5566	0
500	1.1978	0.2313	0
KH054	1.1183	0.0285	0
190	1.0737	0.02	0
JB008	1.8243	0.1408	0
281	1.4489	0.1072	0
288	1.3121	0.2427	0
220	1.2539	-0.5209	0
GB022	1.395	-0.002	0
GB003	1.4707	-0.4474	0
KH002	1.5439	0.1348	0
GB025	1.2704	-0.6828	0
JB045	1.0853	-0.2366	0
188	1.2238	0.2874	0
189	1.3767	-0.0541	0
209	1.1563	0.0102	0
RG024	1.2271	0.2946	0
41	1.1757	-0.0154	0
17	1.394	-0.289	0
RG014	1.1786	0.0035	0
466	1.0829	-0.0499	0
GB007	1.2752	0.4912	0
78	1.3076	-0.1691	0
155	1.5178	-0.6524	0
44	1.0166	0.5931	0
429	0.884	-2.2302	0

Item	Discrimination	Difficulty	Asymptote
481	1.3202	-0.4387	0
371	0.9063	0.568	0
RG004	1.5885	0.0373	0
471	1.2128	-0.7505	0
IE037	1.3185	-0.4778	0
138	1.196	-0.4485	0
31	1.1873	0.3753	0
IE032	1.1475	0.0705	0
RG054	1.4396	0.1089	0
26	1.114	-0.2909	0
88	1.3413	0.4301	0
479	1.1137	0.3074	0
RG022	1.3478	0.0461	0
JB049	1.2923	0.1631	0
KH004	1.342	-0.4095	0
IE044	1.2213	0.4227	0
356	1.2656	-0.1565	0
259	1.0895	0.3055	0
GB051	1.1908	-0.5955	0
IE089	1.0758	-0.6637	0
493	1.5455	-0.0884	0
311	1.0836	0.6262	0
419	0.969	-0.4652	0
327	1.096	0.1459	0
290	1.3189	0.0895	0
159	1.3302	-0.1453	0
KH027	1.5526	0.1728	0
109	0.9452	-0.1185	0
GB031	1.3852	-0.2038	0
GB006	1.3858	0.4899	0
51	1.2034	-0.1447	0
333	1.2283	0.0878	0
GB016	1.464	-0.2153	0

Appendix D: Item Parameters: Process Propensity Dimension

Table D1. Descriptive Statistics for Change in RC Score by Wave, Time Interval, Test Type, and Grade

Change in RC Score from Time 1 to 2 for Wave 1 by Test Type

wave	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	CAT	Not Significant	421	0.18	0.48	0.23	-1.97	1.39	-0.07	0.51
Wave 1	CAT	Significant Change	125	0.61	1.18	0.99	-1.99	3.38	-0.20	1.41
Wave 1	FIT	Not Significant	376	0.07	0.51	0.08	-1.37	1.37	-0.27	0.46
Wave 1	FIT	Significant Change	107	0.79	0.90	1.07	-1.36	2.67	0.42	1.34

Change in RC Score from Time 1 to 3 for Wave 1 by Test Type

wave	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	CAT	Not Significant	239	0.22	0.57	0.29	-1.64	1.66	-0.14	0.64
Wave 1	CAT	Significant Change	86	0.33	1.29	0.45	-2.96	2.69	-0.76	1.29
Wave 1	FIT	Not Significant	212	0.06	0.53	0.06	-1.37	1.17	-0.37	0.45
Wave 1	FIT	Significant Change	73	0.64	1.15	0.97	-1.79	2.81	0.22	1.37

Change in RC Score from Time 2 to 3 for Wave 1 by Test Type

wave	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	CAT	Not Significant	229	0.05	0.54	0.04	-2.25	2.03	-0.24	0.37
Wave 1	CAT	Significant Change	81	-0.16	1.19	-0.24	-2.14	2.65	-1.17	0.78
Wave 1	FIT	Not Significant	196	0.03	0.53	-0.01	-1.27	1.26	-0.35	0.43
Wave 1	FIT	Significant Change	71	0.01	0.89	0.05	-2.45	1.68	-0.66	0.67
Wave 2	CAT	Not Significant	161	0.04	0.64	0.10	-2.20	1.74	-0.34	0.48
Wave 2	CAT	Significant Change	45	0.50	1.46	1.18	-2.73	2.49	-1.15	1.50
Wave 2	FIT	Not Significant	143	0.04	0.47	0.06	-1.05	1.08	-0.27	0.41
Wave 2	FIT	Significant Change	23	0.12	1.33	0.84	-2.37	1.63	-1.08	1.29

Change in RC Score from Time 1 to 2 for Wave 1 by Grade

wave	grade	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	Not Significant	198	0.13	0.54	0.13	-1.97	1.39	-0.18	0.48
Wave 1	Grade 3	Significant Change	67	0.94	1.01	1.10	-1.57	3.38	0.56	1.57
Wave 1	Grade 4	Not Significant	226	0.19	0.47	0.24	-1.11	1.16	-0.10	0.53
Wave 1	Grade 4	Significant Change	64	0.65	0.96	0.99	-1.99	1.68	0.43	1.22
Wave 1	Grade 5	Not Significant	217	0.09	0.50	0.12	-1.39	1.17	-0.24	0.44
Wave 1	Grade 5	Significant Change	59	0.89	1.03	1.14	-1.82	2.69	0.80	1.43

wave	grade	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 6	Not Significant	156	0.10	0.47	0.11	-1.13	1.11	-0.23	0.47
Wave 1	Grade 6	Significant Change	42	0.10	1.13	0.02	-1.53	1.94	-0.96	1.16

Change in RC Score from Time 1 to 3 for Wave 1 by Grade

wave	grade	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	Not Significant	144	0.08	0.58	0.17	-1.37	1.66	-0.36	0.52
Wave 1	Grade 3	Significant Change	51	0.98	1.08	1.22	-1.54	2.81	0.51	1.76
Wave 1	Grade 4	Not Significant	110	0.21	0.55	0.23	-1.64	1.23	-0.21	0.64
Wave 1	Grade 4	Significant Change	39	0.42	1.17	0.89	-2.64	1.93	-0.29	1.27
Wave 1	Grade 5	Not Significant	87	0.10	0.54	0.17	-1.29	1.01	-0.28	0.53
Wave 1	Grade 5	Significant Change	35	0.30	1.23	0.47	-2.59	2.69	-0.37	1.22
Wave 1	Grade 6	Not Significant	110	0.20	0.55	0.24	-1.29	1.63	-0.17	0.54
Wave 1	Grade 6	Significant Change	34	-0.06	1.29	0.23	-2.96	2.52	-1.20	1.01

Change in RC Score from Time 2 to 3 for Wave 1 and Wave 2 by Grade

wave	grade	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	Not Significant	132	0.00	0.62	0.01	-2.25	2.03	-0.44	0.47
Wave 1	Grade 3	Significant Change	50	0.07	1.01	0.25	-2.45	2.02	-0.66	0.78
Wave 1	Grade 4	Not Significant	108	0.02	0.51	0.00	-1.35	1.50	-0.34	0.33
Wave 1	Grade 4	Significant Change	37	-0.03	1.09	0.12	-1.93	2.49	-0.79	0.67
Wave 1	Grade 5	Not Significant	80	0.03	0.45	0.01	-1.05	1.22	-0.28	0.26
Wave 1	Grade 5	Significant Change	31	-0.42	1.02	-0.64	-1.91	2.29	-1.18	0.00
Wave 1	Grade 6	Not Significant	105	0.12	0.50	0.14	-1.19	1.81	-0.15	0.39
Wave 1	Grade 6	Significant Change	34	-0.03	1.11	-0.29	-2.14	2.65	-0.87	0.90
Wave 2	Grade 3	Not Significant	115	0.09	0.59	0.04	-1.62	1.74	-0.31	0.50
Wave 2	Grade 3	Significant Change	23	0.44	1.34	1.05	-1.65	2.49	-1.02	1.28
Wave 2	Grade 4	Not Significant	93	-0.03	0.60	-0.02	-2.20	1.61	-0.39	0.40
Wave 2	Grade 4	Significant Change	25	0.55	1.36	1.23	-2.28	1.91	-1.21	1.39
Wave 2	Grade 5	Not Significant	92	0.06	0.51	0.15	-1.84	0.92	-0.16	0.43
Wave 2	Grade 5	Significant Change	18	0.05	1.61	0.93	-2.73	2.02	-1.61	1.19
Wave 2	Grade 6	Not Significant	4	-0.06	0.47	0.02	-0.69	0.42	-0.23	0.20
Wave 2	Grade 6	Significant Change	2	0.43	2.04	0.43	-1.02	1.87	-0.29	1.15

Change in RC Score from Time 1 to 2 for Wave 1 by Grade and Test Type

wave	grade	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	CAT	Not Significant	100	0.18	0.56	0.25	-1.97	1.39	-0.14	0.57
Wave 1	Grade 3	CAT	Significant Change	39	0.87	1.11	1.06	-1.57	3.38	0.17	1.60
Wave 1	Grade 3	FIT	Not Significant	98	0.09	0.52	0.05	-1.37	1.37	-0.24	0.44
Wave 1	Grade 3	FIT	Significant Change	28	1.03	0.86	1.19	-1.25	2.67	0.84	1.46
Wave 1	Grade 4	CAT	Not Significant	130	0.26	0.45	0.31	-1.11	1.16	0.03	0.55
Wave 1	Grade 4	CAT	Significant Change	32	0.43	1.12	0.93	-1.99	1.64	-0.36	1.19
Wave 1	Grade 4	FIT	Not Significant	96	0.09	0.48	0.07	-0.98	1.05	-0.20	0.47
Wave 1	Grade 4	FIT	Significant Change	32	0.87	0.73	1.04	-1.30	1.68	0.88	1.24
Wave 1	Grade 5	CAT	Not Significant	108	0.12	0.48	0.12	-1.39	1.03	-0.07	0.46
Wave 1	Grade 5	CAT	Significant Change	29	1.05	1.13	1.22	-1.82	2.69	0.99	1.65
Wave 1	Grade 5	FIT	Not Significant	109	0.06	0.53	0.12	-1.19	1.17	-0.31	0.41
Wave 1	Grade 5	FIT	Significant Change	30	0.73	0.91	1.06	-1.10	2.03	0.19	1.39
Wave 1	Grade 6	CAT	Not Significant	83	0.15	0.43	0.19	-0.99	1.11	-0.09	0.41
Wave 1	Grade 6	CAT	Significant Change	25	-0.05	1.15	-0.11	-1.53	1.94	-1.05	1.11
Wave 1	Grade 6	FIT	Not Significant	73	0.05	0.50	0.06	-1.13	1.03	-0.36	0.50
Wave 1	Grade 6	FIT	Significant Change	17	0.32	1.08	0.73	-1.36	1.80	-0.83	1.19

Change in RC Score from Time 1 to 3 for Wave 1 by Grade and Test Type

wave	grade	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	CAT	Not Significant	59	0.07	0.60	0.20	-1.23	1.66	-0.35	0.52
Wave 1	Grade 3	CAT	Significant Change	25	0.68	1.18	1.06	-1.54	2.28	-0.13	1.62
Wave 1	Grade 3	FIT	Not Significant	85	0.08	0.57	0.12	-1.37	1.07	-0.36	0.52
Wave 1	Grade 3	FIT	Significant Change	26	1.27	0.90	1.30	-1.36	2.81	1.03	1.81
Wave 1	Grade 4	CAT	Not Significant	71	0.33	0.55	0.38	-1.64	1.23	-0.01	0.71
Wave 1	Grade 4	CAT	Significant Change	25	0.30	1.25	0.84	-2.64	1.93	-0.31	1.26
Wave 1	Grade 4	FIT	Not Significant	39	0.00	0.50	-0.10	-0.98	1.17	-0.38	0.29
Wave 1	Grade 4	FIT	Significant Change	14	0.63	1.03	0.91	-1.25	1.89	0.25	1.24
Wave 1	Grade 5	CAT	Not Significant	40	0.09	0.58	0.13	-1.29	1.00	-0.44	0.61
Wave 1	Grade 5	CAT	Significant Change	14	0.30	1.40	0.14	-2.59	2.69	-0.32	1.18
Wave 1	Grade 5	FIT	Not Significant	47	0.10	0.52	0.17	-0.92	1.01	-0.28	0.51
Wave 1	Grade 5	FIT	Significant Change	21	0.30	1.13	0.64	-1.79	1.59	-0.41	1.19
Wave 1	Grade 6	CAT	Not Significant	69	0.32	0.53	0.33	-1.29	1.63	0.06	0.68
Wave 1	Grade 6	CAT	Significant Change	22	-0.03	1.37	0.16	-2.96	2.52	-1.01	1.01
Wave 1	Grade 6	FIT	Not Significant	41	0.01	0.51	-0.01	-0.92	1.16	-0.38	0.33
Wave 1	Grade 6	FIT	Significant Change	12	-0.12	1.18	0.25	-1.57	1.53	-1.28	0.70

Change in RC Score from Time 2 to 3 for Wave 1 and Wave 2 by Grade and Test Type

wave	grade	type	AMC_decision	n	mean	sd	median	min	max	q1	q3
Wave 1	Grade 3	CAT	Not Significant	54	-0.06	0.66	0.01	-2.25	2.03	-0.46	0.40
Wave 1	Grade 3	CAT	Significant Change	24	-0.16	1.14	-0.28	-1.81	2.02	-1.29	0.58
Wave 1	Grade 3	FIT	Not Significant	78	0.04	0.60	0.02	-1.27	1.26	-0.42	0.57
Wave 1	Grade 3	FIT	Significant Change	26	0.28	0.84	0.54	-2.45	1.32	-0.01	0.78
Wave 1	Grade 4	CAT	Not Significant	71	0.07	0.51	0.03	-1.35	1.50	-0.23	0.31
Wave 1	Grade 4	CAT	Significant Change	23	-0.05	1.20	0.02	-1.93	2.49	-1.24	0.82
Wave 1	Grade 4	FIT	Not Significant	37	-0.06	0.49	-0.12	-1.04	1.02	-0.40	0.34
Wave 1	Grade 4	FIT	Significant Change	14	0.00	0.92	0.23	-1.72	1.54	-0.39	0.54
Wave 1	Grade 5	CAT	Not Significant	37	-0.01	0.46	0.00	-1.05	0.86	-0.34	0.19
Wave 1	Grade 5	CAT	Significant Change	12	-0.61	1.16	-0.89	-1.91	2.29	-1.27	-0.14
Wave 1	Grade 5	FIT	Not Significant	43	0.06	0.45	0.02	-0.94	1.22	-0.25	0.33
Wave 1	Grade 5	FIT	Significant Change	19	-0.31	0.93	-0.56	-1.51	1.68	-0.98	0.18
Wave 1	Grade 6	CAT	Not Significant	67	0.15	0.50	0.14	-1.19	1.81	-0.05	0.39
Wave 1	Grade 6	CAT	Significant Change	22	-0.02	1.27	-0.03	-2.14	2.65	-0.93	0.94
Wave 1	Grade 6	FIT	Not Significant	38	0.06	0.49	0.02	-0.76	1.21	-0.32	0.39
Wave 1	Grade 6	FIT	Significant Change	12	-0.04	0.81	-0.35	-1.00	1.45	-0.63	0.61
Wave 2	Grade 3	CAT	Not Significant	58	0.14	0.68	0.28	-1.62	1.74	-0.28	0.60
Wave 2	Grade 3	CAT	Significant Change	13	0.67	1.43	1.20	-1.65	2.49	-1.09	1.50
Wave 2	Grade 3	FIT	Not Significant	57	0.03	0.48	-0.01	-1.03	1.08	-0.34	0.38
Wave 2	Grade 3	FIT	Significant Change	10	0.13	1.20	0.06	-1.24	1.63	-0.92	1.21
Wave 2	Grade 4	CAT	Not Significant	51	-0.07	0.67	-0.07	-2.20	1.61	-0.41	0.27
Wave 2	Grade 4	CAT	Significant Change	18	0.56	1.42	1.22	-2.28	1.91	-0.66	1.45
Wave 2	Grade 4	FIT	Not Significant	42	0.02	0.49	0.03	-1.03	0.95	-0.29	0.43
Wave 2	Grade 4	FIT	Significant Change	7	0.51	1.30	1.29	-1.40	1.55	-0.28	1.34
Wave 2	Grade 5	CAT	Not Significant	48	0.05	0.56	0.14	-1.84	0.92	-0.38	0.43
Wave 2	Grade 5	CAT	Significant Change	13	0.15	1.60	1.01	-2.73	2.02	-1.34	1.20
Wave 2	Grade 5	FIT	Not Significant	44	0.08	0.45	0.15	-1.05	0.79	-0.16	0.42
Wave 2	Grade 5	FIT	Significant Change	5	-0.20	1.82	0.84	-2.37	1.40	-1.97	1.12
Wave 2	Grade 6	CAT	Not Significant	4	-0.06	0.47	0.02	-0.69	0.42	-0.23	0.20
Wave 2	Grade 6	CAT	Significant Change	1	1.87	NA	1.87	1.87	1.87	1.87	1.87
Wave 2	Grade 6	FIT	Significant Change	1	-1.02	NA	-1.02	-1.02	-1.02	-1.02	-1.02