Technical Manual 2022: Multiple-Choice Online Causal Comprehension Assessment (MOCCA)-College

MOCCA-College Technical Report (MCTR) 2022

Ben Seipel,

California State University, Chico

Sarah E. Carlson,

Georgia State University

Virginia Clinton-Lisell,

University of North Dakota

Mark L. Davison,

University of Minnesota- Twin Cities

Patrick C. Kennedy,

University of Oregon

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180417 to the California State University, Chico. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

MOCCA-College Technical Manual

Contents

Acknowledgements

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

Special thanks are extended to our core MOCCA-College team of undergraduate students, graduate students, and research project assistants who aided in item writing, item analysis, recruitment, data collection, and other tasks: Heather Ness-Maddox, Amanda Dahl, Xinle Hong, Meghan Tadeo, Ashley Overstreet, Terrill Taylor, Surjya Bajpayee, Aurore Phenow, Qian Zhao, Peter Li, Youfu Yan, Jiayi Deng, Yun Leng Wong, and Hao Jia.

We also have expressed our sincere appreciation to our panel of experts who supported the project from the beginning and challenged us to make MOCCA-College the best it could be: Drs. Gina Biancaosa, Dolores Perin, Joseph Magliano, and John Sabatini.

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

MOCCA-College would not have been possible without the support of our administrative and programming teams. We offer sincere thanks to Rachael Beyer, Joann Rose, Joleen Barnhill, and Blanca Estrada for keeping the MOCCA-College development on task and organized. We also are indebted to our programming team. Without their skills, MOCCA-College would not exist in its current form: Scott McCammon and Emberex. Similarly, we recognize and appreciate the support of the University of Oregon's Center on Teaching and Learning (CTL) for hosting all computer-administered versions of MOCCA (i.e., the original MOCCA, MOCCA-CAT, MOCCA-College) in their organization.

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

The research reported here was supported by the Institute of Education Science, U.S. Department of Education, through Grant R305A180417 to the California State University, Chico. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

MOCCA-College Theoretical and Measurement Foundations

Originally designed for students in Grades 3 through 5, MOCCA (formerly 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, or reading words accurately and fluently. They might have limited vocabulary and background knowledge. But there are some students who don't comprehend well and don't fall into these categories.

Researchers have dubbed this latter group of readers "poor comprehenders" and have found that they struggle to generate inferences that help them maintain a coherent idea of what a text is about. These poor comprehenders are usually trying to make sense of what they read, but they do so primarily by relying on strategies that don't fully do the trick. It turns out, they tend to rely on one of two strategies: paraphrasing or generating elaborations, which include elaborative inferences, personal associations, and self-explanations. Both are great strategies, but neither alone will result in excellent comprehension. What's more, research suggests that students who rely on paraphrasing versus making generating elaborations require somewhat different instruction (McMaster et al., 2012; van den Broek et al., 2017).

MOCCA-College extends this previous research and assessment development to college students. Research has shown that far too many college students are not ready for college-level reading. Thus, the purpose of MOCCA-College is to help identify why postsecondary students struggle with reading comprehension; information that can be used to inform the kinds of instructional support students need.

Theoretical Foundation

Proficient readers at any level must attend to a range of text characteristics (e.g., letters, sounds, words) as well as their understanding of the content by drawing on and integrating explicit text information and background knowledge. Consequently, skill in reading words, although necessary for comprehension, is not sufficient on its own to guarantee comprehension. To comprehend successfully, readers must use comprehension processes to build a coherent mental representation of a text. Mental text representations are idiosyncratic, cognitive structures people create to understand the situation. This coherent mental representation is called a situation model. A *situation model* includes events from a text along dimensions of time, space, characters, character goals, and causality (Graesser & Clark, 1985; Graesser et al. 1994; Kintch, 1998; McNamara et al., 1996).

Research has yielded evidence for poor comprehension among both intermediate grade students (Cain & Oakhill, 1999, 2006; Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007) and older readers where no word reading difficulties exist (Hoachlander et al., 2003; Thurlow & van den Broek, 1997). A preponderance of the evidence for poor comprehenders comes from research with intermediate grades. Although frequency of poor comprehension varies across studies, the occurrence of poor comprehension does not vary.

Types of Poor Comprehenders

Research has demonstrated that poor comprehenders are not a monolithic group. Although they share characterization as failure to engage in necessary comprehension processes, the alternative comprehension processes in which they engage instead can also help distinguish them (Carlson et al., 2012; McMaster et al., 2012). When word reading, other component skills, and knowledge are ruled out as causes of comprehension problems, research has shown that poor comprehenders make fewer *necessary inferences* than do proficient comprehenders do (Thurlow & van den Broek, 1997).

During reading, proficient comprehenders engage in a host of comprehension processes, but only some are truly necessary to comprehension (Graesser et al., 1994; McKoon & Ratcliff, 1992; van den Broek et al., 2005). 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 (Trabasso & van den Broek, 1985). 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 can 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.

Although poor comprehenders do make causally coherent inferences, they tend to make fewer of them than proficient comprehenders do, *and* they instead tend to rely on one of two reading comprehension processes that are good practices but are neither necessary nor sufficient for maintaining *causal* coherence (Coté, 1998; Trabasso & Magliano, 1996a, 1996b; van den Broek et al., 2001; Wolfe & Goldman, 2005). The first of these processes is *paraphrasing*. Paraphrases restate or rephrase prior text, which can support coherence, but are not strictly necessary for maintaining causal coherence. Moreover, they do not strictly rely on background knowledge. The second process that poor comprehenders tend to overuse is elaboration. Although we prefer the more precise term lateral connection, because this category includes more than just elaborations (e.g., self-explanations, evaluations, and associations), we will use the term "elaboration" here, both because it is the more common term in the literature, and because elaboration is the most common response type in this category. Elaborations access background knowledge but are not necessarily *causally coherent* connections (i.e., those that close the gap in a causal manner by drawing on relevant background knowledge to help connect to the text).

Research has shown that both good and poor comprehenders can and do use *many* other comprehension processes during reading; 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 (Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). In other words, what distinguishes poor comprehenders from good comprehenders, holding their word reading and vocabulary constant, is their less consistent and strategic use of causally coherent inferences. And what further distinguishes poor comprehenders *from each other is the comprehension process they tend to overuse instead*: paraphrases or elaboration.

Think alouds and MOCCA-College

The established differences in cognitive processes generated during reading between good and the two types of poor comprehenders comes from previous think-aloud research (Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). During a think-aloud task, a reader reads aloud a unit of text (e.g., a sentence) and verbalizes what they are thinking about while reading. Think alouds are an effective measure of online cognitive processes during reading comprehension. Think alouds have been used to identify specific comprehension processes (e.g., inferences; paraphrases; associations; metacognitive responses) that take place during reading. In fact, think alouds are the prime source of evidence that readers indeed use different types of comprehension processes. Think alouds are also the prime source of evidence that poor comprehenders can be distinguished diagnostically by the processes they overuse—paraphrases or elaborations. Although think alouds are an online, rich, and reliable source of information on comprehension processes, they are also impractical for schools because of the data collection, coding, and analysis burden they pose. The benefit of think alouds, combined with their limitations, however, led to the development of MOCCA as a new assessment tool to identify differences in struggling comprehenders.

Recalls and MOCCA-College

After reading a text, readers are asked to recall what they remember about the text they just read. Recalls are primarily considered a measure of memory for the text also referred to as literal comprehension. Because the amount of information accurately recalled from a text is indicative of reading skill (Cao & Kim, 2020), a positive correlation between recalled information and MOCCA-College causally-correct answers would be validating evidence for the assessment. Recalls can also be further analyzed in relation to the specific information (idea unit) recalled from the original text read. For the purposes of developing MOCCA-College, the connectedness of the idea unit to the other idea units in a text in terms of cause and effects (i.e., causal cohesion of the text) was examined. Readers tend to recall highly connected idea units more than other idea units particularly when overall reading comprehension skills are strong (Pavias et al., 2016; van den Broek et al., 2009; Yeari & Lavie, 2021). This relates to the concept of sensitivity to structural centrality in which readers become more cognizant of what ideas in a text are important for cohesion as readers gain experience and skill (van den Broek et al., 2012).

The Design of MOCCA and MOCCA-College Items

Each original MOCCA and MOCCA-College item consists of a short seven-sentence narrative text. It is a discourse-level cloze task: instead of deleting every n^{th} word as in traditional cloze or maze tasks, however, the sixth *sentence* of each seven-sentence text is deleted. After each text, there are three corresponding multiple-choice responses to fill in the missing sentence (cf. Carlson, Seipel, & McMaster, 2014; Davison, Biancarosa, Carlson, Seipel, & Liu, 2018). Each response type corresponds to one of three types of responses that were developed and inspired by the responses found in students' think-aloud protocols: A *causally coherent inference* which relies on causal information in the text and is necessary for maintaining coherence (Trabasso & van den Broek, 1985); a *paraphrase* which rephrases or restates prior text and does

not maintain causal coherence; and an elaboration which includes elaborations, associations, or evaluations about the text but does not maintain causal coherence.

Original MOCCA items are narrative texts with a causal structure centered *on a main goal and motivated subgoals and events* (e.g., Trabasso & van den Broek, 1985). MOCCA-College items use the same structure; however, the assessment includes expository items in addition to narrative items to better represent the text type and topics of college students. Expository items range in a variety of academic content (e.g., political science, historical accounts) and are further identified as cause/effect, problem/response, chronological, and descriptive. Additionally, MOCCA-College items have Flesch-Kincaid Grade Levels appropriate for college students (Flesch-Kincaid Grade Levels 6-14).

As mentioned, for each test item, the student must choose one of three alternative responses to fill in the deleted sentence. As described in more detail in the literature review, in addition to the correct answer (i.e., a causally coherent inference), the two remaining alternative responses are informative distractors: a *paraphrase* and *an elaboration*.

Recalls and MOCCA-College

After reading a text, readers are asked to recall what they remember about the text they just read. Recalls are primarily considered a measure of memory for the text also referred to as literal comprehension. Because the amount of information accurately recalled from a text is indicative of reading skill (Cao & Kim, 2020), a positive correlation between recalled information and MOCCA-College causally-correct answers would be validating evidence for the assessment. Recalls can also be further analyzed in relation to the specific information (idea unit) recalled from the original text read. For the purposes of developing MOCCA-College, the connectedness of the idea unit to the other idea units in a text in terms of cause and effects (i.e., causal cohesion of the text) was examined. Readers tend to recall highly connected idea units more than other idea units particularly when overall reading comprehension skills are strong (Pavias et al., 2016; van den Broek et al., 2009; Yeari & Lavie, 2021). This relates to the concept of sensitivity to structural centrality in which readers become more cognizant of what ideas in a text are important for cohesion as readers gain experience and skill (van den Broek et al., 2012).

Intended Uses

General Use

MOCCA-College is designed to identify and diagnose postsecondary students who struggle with reading comprehension. MOCCA-College is appropriate for students who are entering or are enrolled in postsecondary institutions. Use beyond these student populations has not been validated and is not supported. More specifically, MOCCA-College is designed to identify the cognitive processes that poor comprehenders overuse while reading (i.e., paraphrasing/repeating text or generating invalid inferences, connections, irrelevant elaborations, associations, or evaluations). The information gained from administration can be used to measure general reading comprehension ability (i.e., good or poor comprehender), identifying types of

poor comprehenders (i.e., paraphraser, elaborator), and determining comprehension efficiency (i.e., fast or slow).

Specific Intended Uses

MOCCA-College has three validated forms which can be used for diagnosis and potentially progress monitoring. Given the slow rate at which reading comprehension changes and the number of forms available, it is recommended that MOCCA-College be administered no more than three times per academic year. MOCCA-College has concurrent and construct validity evidence to indicate that it provides information similar to other more-traditional reading comprehension assessments (see Validity section below). Therefore, it can be appropriately used as a screening measure for all students.

Additionally, MOCCA-College can be used as a cognitive diagnostic tool about individual students who struggle with reading comprehension. That is, MOCCA-College not only identifies those at risk for poor reading comprehension, but also provides instructionally relevant diagnostic information about why a student is struggling with reading comprehension. Specifically, it identifies the cognitive reading comprehension processes that a student who struggles with comprehension overuses.

Inappropriate Uses of MOCCA-College and MOCCA-College Data

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

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

Administration

Administration Qualifications

General Administration

MOCCA-College is generally considered a *Level A* assessment (see Self-Assessment section below). This means that it requires minimal special qualifications to administer and interpret scores. It is recommended that the assessment be administered and interpreted by personnel who understand MOCCA-College and of reading comprehension. Specifically, the assessment should generally be administered by a teacher, paraprofessional, administrator, school psychologist, 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.

Self-Assessment

One unique aspect of MOCCA-College as compared to the other versions of MOCCA is it has the capacity for self-assessment by post-secondary students. Potential test-takers can register via this URL: https://mocca.uoregon.edu/#/admin/register. Registration requires the user to supply a name, supply an email address (to retrieve password/results), and agree to the terms of use. Unlike General Administration, Self-Assessment does not provide a full report of scores or data. If completed, the Self-Assessment will provide a personal classification based on the assessment results rather than a full report of scores which would be more appropriate for an instructor or administrator who uses MOCCA-College with a group of students. Additionally, at the end of the assessment, the system presents an embedded YouTube feedback video based on the results that provide interpretation and suggestions for improving comprehension skills. (See more about the videos under *Scoring and Interpretation* section below).

MOCCA-College is only validated for computerized administration. Although the original MOCCA was developed in a paper-and-pencil format (Carlson et al., 2014), no paper-and-pencil versions are available at this time.

Administration System Requirements

The MOCCA-College 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 network of the school, or testing center, or individual test taker. For self-administration, a secure wi-fi connection will suffice. Once logged-in, examinees have the option to read or listen to the instructions. The auditory instructions are optional. When selecting the pre-recorded instructions, headphones are encouraged but optional. In a group testing environment, headphones are strongly encouraged to minimize distractions.

Development Process

The original MOCCA was a paper-and-pencil assessment for students in Grades 3-5, with a single form consisting of 40 items (Carlson et al., 2014). A measurement grant from the Institute of Education Sciences (IES) R305A140185; PI Dr. Gina Biancarosa) enabled the refinement of MOCCA into its current form, which is a computer-administered assessment with three forms per grade for Grades 3-5 (R305A140185; PI Dr. Gina Biancarosa). An additional measurement grant from the IES (R305A180417) allowed for the current extension to MOCCA-College and application for college-level students (R305A180417; PI Dr. Ben Seipel). There is a computer adaptive version of the child intermediate grade version, MOCCA-CAT, under development with funding from the Institute of Education Sciences (R305A190393; PI Dr. Gina Biancarosa).

The Evolution of the MOCCA Item Format

Original MOCCA Items

The original paper-and-pencil MOCCA included four responses: one correct, two informative incorrect responses, and one uninformative incorrect response (a true distractor) (Carlson et al., 2014). After further development and refinement to a computer-administered version, the final structure of MOCCA uses three responses: one correct and two informative incorrect responses. The main reasons for including four responses originally was to reduce the probability of obtaining the correct answer by random guessing. In theory, including three instead of four responses in the final structure of MOCCA should reduce random error associated with guessing, thereby increasing the reliability and the IRT information function throughout the range of ability (i.e., θ) but especially at the low end of the scale where guessing is more common.

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

To reiterate, the three response types that MOCCA incorporates are:

- **Causally coherent inferences**, which represent the correct responses because they provide information necessary to fill the causal gap between the 5th and 7th sentences, completing the text by stating or implying whether the goal or purpose of the main character (or text) has been met.
- **Paraphrases**, which are incorrect because they paraphrase either the original or updated goal in a text but add no new information, thereby leaving the causal gap unfilled.
- **Elaborations**, which are also incorrect because they build on the 5th sentence of the text by adding extra-textual information through elaboration, association, or explanation, but they do not fill the causal gap. (Note: In some previous literature we refer to elaborations as "lateral connections" e.g., Carlson et al., 2014).

Revising Original MOCCA Items for MOCCA-College

To help meet the goal of including items with a range of readability, the current version of MOCCA-College uses revised versions of some items from the original MOCCA. Because the goal was to include a range of readability across the items, the original, validated MOCCA items from the 5th grade forms were screened by the grant PIs for potential revisions and use in MOCCA-College. Specifically, items were screened for difficulty, potential appropriateness for college-students (i.e., content, topic), and feasibility of revisions. Because the original MOCCA

items were written for younger readers, planned revisions were designed around increasing readability based on Flesch-Kincaid Grade Level. Once selected and revised, potential items were included for review by our content-review panel (see *Phase 4* section below under *New MOCCA-College Items*).

New MOCCA-College Items

In addition to revised, original MOCCA items, MOCCA-College includes new items – using both narrative and expository texts. Because the original MOCCA does not include expository items, all expository items are new. In addition, new narrative items were developed to supplement the original MOCCA narrative items modified for MOCCA-College. To develop new items, an item-writing team was created. The item-writing team consisted of three of the four MOCCA-College PIs and other research personnel. Item writing and revisions proceeded in five phases: 1) ideation and initial text writing, 2) text review and revisions, 3) distractor development and review, 4) content review, and 5) item revisions and selection.

Phase 1: Ideation and Initial Text Writing.

In Phase 1, the item-writing team generated a list of potential topics for new narrative and expository texts. The team aimed to generate ideas and topics that spanned the typical college curriculum and experience. Phase 1 also consisted of a single author writing the complete seven sentence text from on the list of topics and ideas. Flesch-Kincaid (FK) Grade Levels were recorded at the completion of each story to ensure there was a sufficient number of items that ranged from FK 6-14, and the idea list was updated to reflect whether an existing or new idea had been used.

Phase 2: Text Review and Revisions.

In Phase 2, a team of 2 or more authors, not including the original author from Phase 1, reviewed and revised the item/text. Items were reviewed for (a) causal coherence, (b) content accuracy, (c) age appropriateness, (d) grammar and spelling, (e) freedom from bias, and (f) engagingness. As part of the process, the sixth sentence was removed and placed below the text as the causally coherent, or correct, response. As necessary, items were revised based on the above criteria (a-f).

Phase 3: Distractor Development and Review.

In Phase 3, two additional distractor responses were written: one paraphrase and one elaboration. Each response was written to and reviewed against response specifications. Specifically, the response specifications included all responses had to be similar in length and readability. In addition, paraphrases had to be structured in one of two ways: 1) They needed to either paraphrase the main goal or idea of the text; or 2) They had to paraphrase the updated goal of the text. Elaborations had to build on an idea presented in sentence five of the text. The elaboration had to require background knowledge or introduce new information. As such, elaborations could be definitions, clarifications, tangential information, emotions, etc. Distractors were then carefully vetted by the item writing team. Specifically, paraphrases received extra attention to ensure that they did not include implicit inferences about information in the text.

Elaborations received extra attention to ensure that they did not close the causal gap between the 5th and 7th sentences.

Phase 4: Content Review Panel.

In Phase 4, a panel of five local educators with experience with college-bound and college students reviewed the items as intact text (i.e., with the causally correct response inserted into the full text as the "missing" 6th sentence). Additionally, two graduate students, new to the project, reviewed all items.

The panel reviewed the items along five dimensions: (a) causal coherence; (b) accuracy; (c) appropriateness regarding vocabulary, syntax, length, content, and background; (d) freedom from bias (i.e., offensive or privilege); and (e) engagingness (i.e., relevance and attention) (See Appendix A). The panel was trained on the purpose of MOCCA-College, the intent behind the dimensions they were rating, the meaning of the rating scale they used, and training in how to use the Qualtrics online interface for rating. The review panel used a formal scale to rate items on these dimensions. Any dimension per item marked as a concern prompted the reviewer to provide an explanation. Interrater reliability was not a goal in this training because judgments of the dimensions were necessarily subjective. Instead, the goal was for consistent and appropriate use of the rating scales and Qualtrics system. Teacher reviews were completed in four waves of about approximately 65 items each. Each and every MOCCA-College item was reviewed by at least three members of the content review panel.

Phase 5: Item Revision and Selection.

In Phase 5, the author team reviewed the review panel ratings and written feedback in depth. In general, the feedback on the items was favorable. In some cases, revisions were made to stories texts based on the feedback.

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

In all, nearly 300 items were written, including revisions to the original MOCCA items. Of these, 200 items made it to the pilot phase with 50 appearing on each of 4 forms for the Regional Pilot study (see more *MOCCA-College Regional Pilot* section below). Several items were "retired" based on feedback from the review panel because they were deemed as too problematic to revise based on the timeline and item specifications. Others were determined to be too easy or childish for college students. Final exclusions were based on efforts to create

forms that were as nearly equal as possible in terms of average and distribution of Flesch-Kincaid Grade Levels and counts of items featuring (a) narrative and expository items; (b) narrative items with male, female, and indeterminate gender characters; (c) expository items with content that crossed disciplines; (d) narrative endings considered to be happy, sad, and neutral emotionally; (e) whether the emotion at the end of narrative was explicitly stated or needed to be inferred; and (f) whether an expository item addressed a sensitive or controversial topic such as death.

MOCCA-College Regional Pilot: Fall 2018-Spring 2019

In Year 1 of the funded project, the MOCCA-College regional pilot occurred in California, the Upper Midwest (primarily North Dakota), and the Southeast (primarily Georgia) between December 2018 and June 2019. A total of 1,220 students consented to take the assessment, and 1,170 satisfied our requirement of having completing at least 10 items (i.e., enough items to ensure an accurate calculation of response type propensity). MOCCA-College personnel oversaw administration of the assessment during the pilot. Three raw scores were computed for each student. The first was the traditional number correct (i.e., the number of items for which the student selected the Causally Coherent Inference response). The remaining two scores were (a) the Paraphrase score, the number of times the student picked the Paraphrase response, and (b) the Elaboration score, the number of times the student selected an Elaboration response. Additionally, reading comprehension efficiency scores were calculated for each student, determined by dividing the minutes of total testing time divided by the number of correct responses.

Results from the pilot study indicated that the internal consistency reliability (i.e., Cronbach's alpha) for the correct score was excellent across all forms (.92 - .95). The Paraphrase scores had high reliability (.85 - .88) as well. The reliabilities for the Elaboration score were more variable, but still acceptable to good (.77 - .88). Across all forms, the expository items were more difficult (.71 - .76) than were narrative items (.85 - .90).

These results informed revisions for the field test (Year 2) version of MOCCA-College. Several item statistics were used to inform the item revision process. For the correct response, these included the item difficulty (proportion correct), the item-total correlation, and the two-parameter logistic (2PL) item response theory parameters (difficulty and discrimination parameters.) For the informative, incorrect response alternatives (i.e., Paraphrase and Elaboration), the item statistics included the "difficulty" (i.e., the proportion of respondents who selected the response) and the item-total correlation, where the total score was the number of items for which the student had selected that incorrect alternative. Items with excellent statistics were examined and compared to those with particularly poor statistics in order to refine item specifications for the informative incorrect responses. Items with poor item-total correlations for any of the three informative scores (i.e., correct, paraphrase, elaboration) were targeted for revision. Items were dropped during revisions if the item was deemed too difficult or complicated to revise. Finally, items where all item-total correlations, including for the correct response type, were poor, were also dropped outright. Thus, the final stage of the pilot was to

create three new forms (instead of four forms that were used in the pilot) of 50 items each. Forms were again balanced for FK Grade Level readability and the other features as were for the pilot to prepare for the National Field Test.

MOCCA-College National Field Test: Fall 2019-Fall 2020.

In Year 2 of the funded project, the MOCCA-College field test occurred nationwide. Students and organizations were recruited from participating universities and programs in the pilot. Additionally, TRIO and similar programs were recruited via the Council for Opportunities in Education annual conference. Similarly, students and programs were recruited via listservs and social media. A total of 1,728 students consented to take the assessment, and 1,683 satisfied our requirement of having completing at least 10 items. Notably, the onset and duration of the COVID-19 pandemic interfered with both recruitment and data collection for this field test.

Results indicated that the three new forms were slightly more difficult, on average, than the four forms used in the Regional Pilot. Nonetheless, the data indicated that the internal consistency reliability (i.e., Cronbach's alpha) for the correct score was still excellent across all forms (.91 - .92), and the Paraphrase scores still had high reliability (.84 - .85) as well. The reliabilities for the Elaboration score were again more variable, but still acceptable to good (.75 - . 80). Similar to the Regional Pilot, across all forms expository items were more difficult (.73 - .77) than were the narrative items. Across all forms, the expository items were more difficult (.73 - .77) than the narrative items (.79 - .86), similar to the Regional Pilot.

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

MOCCA National Calibration Study: Spring 2021-Spring 2022

For the final year of the funded project (note: similar to the 2019-2020 National Field Test, COVID-19 complicated recruitment and testing; thus, the final year was from Spring 2021 through Spring 2022), we recruited a national sample of participants. We built upon our sampling mechanisms in 2019-2020 through continued outreach. Additionally, we added the use of the third-party participant program Prolific to expand our reach. In the end, there were 1,675 test-takers, including 1,6141 who completed at least 10 items.

To reduce overall testing time, we reduced the form length from 50 to 40 items for each of the three forms. Additionally, to equate forms, we selected 10 of the best performing items from the national field test to serve as anchor items on all three forms. To accommodate these

changes, items with relatively weak item statistics in the 2019-2020 field test were dropped. First, we dropped those items identified in the DIF analysis described above, as well as items that did not seem to distinguish well between Paraphrasers and Elaborators. That is, we prioritized the inclusion of items that discriminated as well as possible between students' propensities to choose a paraphrase response versus an elaboration response.

Scoring and Interpretation

MOCCA-College score reports offer a great deal of information not provided by other reading comprehension assessments. Most unique are the Process Propensity and Comprehension Efficiency classifications. These classifications, along with other MOCCA scores are described below. With respect to Paraphrase and Elaboration, we classify students based on an IRT dimension called the Process Propensity (PP) dimension (see the IRT section below). Dimension PP is scaled so that the 0 point is a point of indifference at which a student has a conditional probability of .5 of choosing a paraphrase over an elaborative response on an item of average difficulty given that the student does not select the correct answer. This means that if a respondent has a positive score on Dimension PP, they tend to favor paraphrase responses over elaborative responses, at least by some small amount. Conversely, if the PP score is negative, respondents tend to favor elaborative responses over paraphrase responses by at least a small amount.

Form: Indicates which of the three forms the student took.

Process Propensity: Indicates whether the student has a dominant incorrect response type, and if so, which type of response (Paraphrase or Elaboration) is dominant in the student's incorrect responses if there is a dominant type.

Elaborator: Indicates a test-taker who tends to choose more elaborations (i.e. generate elaborative inferences, personal associations, or self-explanations) when choosing an incorrect response.

Paraphraser: Indicates a test-taker who tends to choose more paraphrases (i.e., restate the text in their own words without adding new information) when choosing an incorrect response.

Indeterminate: Indicates a poor comprehender who could not be confidently classified with one of the two distractor response types. The student shows no clear response preference.

Not Applicable: Indicates a student who performs very well on MOCCA (an above average Scale Score) and who chooses few elaborations or paraphrases.

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

Fast and accurate: Comprehension rate is ≤ 1:24 with 80% or greater accuracy.

Fast and inaccurate: Comprehension rate is ≤ 1:24 with less than 80% accuracy.
Slow and accurate: Comprehension rate is ≥ 1:24 with 80% or greater accuracy.
Slow and inaccurate: Comprehension rate is ≥ 1:24 with less than 80% accuracy.

Scale Score: This is a score between 150 and 950 that reflects overall reading comprehension ability. It is scaled to have a mean of 500 and a standard deviation of 150.

Percentile Rank: Indicates the percentage of students in our national calibration sample who had a Scale Score equal to or lower than that of the test-taker. This norm group is a user group that is not necessarily representative of all entering college students. Descriptive statistics are provided in Table 9 below.

Process Propensity Interpretation and Recommendations

One recommendation that the collective MOCCA team across versions has continuously stated is to always be sure to coordinate MOCCA-College results with other data sources. Only students who have a Scale Score below the mean receives a Process Propensity classification. The Process Propensity classification is used to inform intervention support for poor comprehenders, but above-average comprehenders are unlikely to need such an intervention. Furthermore, above-average comprehenders do not make enough incorrect responses to permit confident Process Propensity classification.

In addition, students who receive **Indeterminate** as their Process Propensity may have decoding or fluency problems, may be using guessing as a test-taking strategy, or may have some other issue underlying poor performance on comprehension measures. Thus, MOCCA-College may not be the best assessment to use to identify areas of focus for improving reading or reading comprehension.

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

Students identified as having an **Elaborator Propensity** appear to be relying on making elaborative inferences, personal associations, and/or self-explanations. They are making inferences, but these inferences do not make the most meaning of what they read. *Although these are all good strategies for reading comprehension, these students need to be encouraged to prioritize maintaining the coherence of the message of what they read. Coherence in reading comprehension narratives often depends on causal relations and how one event or character influences another.*

Comprehension Efficiency Interpretation and Recommendations

As mentioned, always be sure to interpret MOCCA-College results in conjunction with other data sources. Only students who answer at least one item correct (and therefore have a comprehension rate) will receive a comprehension efficiency indicator.

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

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

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

Students who are *slow and inaccurate* do not comprehend well and proceed at a slow pace. Depending on their comprehension rate (also found on MOCCA-College reports), they may just be a bit slow or very slow. A number of issues may be underlying their performance, including, but not limited to, poor decoding and/or fluency. For students who are slow and accurate, the reading process itself may distract attention and working memory in ways that interfere with learning new material through reading (LaBerge & Samuels, 1974; Samuels & Flor, 1997). Other data is necessary to determine their needs.

Feedback Videos

Because of the nature of college instruction, classroom interventions are unlikely. To provide test takers with meaningful, yet accessible feedback, our panel of expert consultants recommended that we develop short (i.e., < 5 minutes) videos that would provide simple interpretation of results and suggest potential next steps. Upon completion of MOCCA-College, the platform provides a link to a YouTube video with feedback tailored to each student's results. The four videos are hyperlinked below:

- Good Comprehenders
- Paraphrasers
- Elaborators
- Indeterminate Results

Equating

Using the 2021-2022 National Calibration sample, we have calibrated items along an IRT based Reading Comprehension (RC) dimension and a Process Propensity (PP) dimension. Dimension PP is used to generate the Process Propensity classification described above. The raw Process Propensity dimension score is not reported, but only the classification based on that dimension is reported. The text below describes the process for deriving a test-taker's Process Propensity classification from their Process Propensity dimension score. Both dimensions were equated across forms using linking items and concurrent calibration.

Equating Item Design

In the 2021-2022 National Calibration sample, all forms contained ten common, anchor items. There were five expository and five narrative anchor items. These items were used as anchor items in the concurrent calibration process. Anchor items were distributed throughout the form. A given anchor item had the same location on all forms. Forms were randomly assigned to test-takers.

Comprehension Dimension Calibration

In calibrating items along the comprehension dimension, each item was scored as 1 if correct and 0 if incorrect. A three-parameter logistic model was fitted to all items using the ten common items as anchor items and with the pseudo-guessing parameter having a tight prior (mean = .25, standard deviation = .025) for all items. The item parameters of each anchor item were held equal across all forms to calculate θ on a common scale with person parameters having mean 0.0 and variance 1.0 in the aggregated population comprised of persons taking any of the three forms. A scale score (SS) was then calculated for each person via the formula:

$$SS = 150 * \theta 150\theta + 500 \quad (1)$$

In the total population aggregated across all forms, the SS has a mean of 500 and a variance of 1.0. Appendix B contains the norm table for the Scale Score with the National Calibration sample as the norming group. The National Calibration sample is a user group and may not be representative of the population of entering college students. Descriptive statistics for the sample are provided in Table 9.

Process Propensity Dimension Calibration and Classification

In the calibration of items along the Process Propensity dimension, a response was scored as 1 for a paraphrase response, 0 for an elaboration response, and as missing for a causal coherent response. Responses were fitted using a two-parameter logistic model. Process propensity scores are on a θ scale with item parameters calibrated to have a mean of 0.0 and a variance of 1.0 in the aggregated population of persons taking any of the three forms.

Scores on the Process Propensity dimension are then used to derive a process propensity classification for each test-taker. Given that the Process Propensity has item parameters scaled to

have a mean of 0, a person with θ = 0.0 will have a .5 probability of choosing either a Paraphrase or Elaboration response for an item of average difficulty. That is, if θ = 0.0, the test-taker has no propensity toward either the Paraphrase or the Elaboration response. For persons with θ > 0, their probability of choosing the Paraphrase response is greater than .5 for an item of average difficulty, meaning that they have a paraphrase propensity, although it would be a small propensity toward the Paraphrase response if θ is only slightly greater than 0. Conversely, for persons with θ < 0, their probability of choosing an Elaboration response is greater than .5, meaning that they have an Elaboration propensity, although again, the Elaboration propensity would be small if θ is only slightly below 0.0.

Our goal is to classify a student as having a Paraphrase Propensity if their response pattern shows a clear tendency in their responses toward the Paraphrase responses over the Elaboration response when the test-taker responds incorrectly. Conversely, our goal is to classify a student as having an Elaboration Propensity if their response pattern shows a clear tendency toward the Elaboration response when the test-taker makes an incorrect response. To identify a clear tendency toward one or the other incorrect response types, we use the generalized likelihood ratio (GLR).

The GLR procedure begins by first establishing an indifference region along Dimension PP around the cut-off separating elaborating from paraphrasing propensities, θ = 0.0 in the present case. Let UB be the upper bound for the indifference region and let LB be the lower bound: LB < 0 < UB. We chose an indifference region covering one standard deviation from $LB = \hat{\theta}_{i.PP}$= -0.5 to $UB = \hat{\theta}_{i.PP} = 0.5$ where $\hat{\theta}_{i.PP}$ is the weighted maximum likelihood estimate (Warm, 1989) for the location of person *i* on Dimension PP.

The GLR is a ratio of two likelihoods that correspond to two regions along Dimension PP. The first region is the region below the lower bound: $\theta_{i.PP} \leq$ LB = -0.5. The second is the region above the upper bound: $\theta_{i.PP} \geq$ UB = 0.5. The numerator is the value of $\theta_{i.PP}$ in the region above the upper bound with the highest likelihood given the test taker's response pattern. The denominator is the value of $\theta_{i.PP}$ in the region below the lower bound with the highest likelihood. The exact computation of the GLR depends on where the estimate $\hat{\theta}_{i.PP}$ is located.

If $\hat{\theta}_{i.PP}$ is in the indifference region (LB $\leq \hat{\theta}_{i.PP} \leq UB$), then

$$GLR = \frac{L(\text{UB}|\boldsymbol{X}_i)}{L(\text{LB}|\boldsymbol{X}_i)} \quad (2)$$

where $L(\text{UB}|\boldsymbol{X}_i)$ is the likelihood of the person's response vector $\boldsymbol{X}_i$ at the upper bound $\theta_{i.PP} =$ UB, 0.5 in our case, and $L(\text{LB}|\boldsymbol{X}_i)$ is the likelihood of the person's response vector $\boldsymbol{X}_i$ at the lower bound $\theta_{i.PP} = LB$, -0.5 in our case.

If $\hat{\theta}_{i.PP} \geq$ UB, then the algorithm computes

$$\text{GLR} = \frac{L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)}{L(\text{LB}|\boldsymbol{X}_i),}. \quad (3)$$

where $L(\text{LB}|\boldsymbol{X}_i)$ is defined as before and $L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)$ equals the likelihood of the person's response vector at $\theta_{i.PP} = \hat{\theta}_{i.PP}$. If $\hat{\theta}_{i.PP} \leq$ LB, then

$$\text{GLR} = \frac{L(\text{UB}|\boldsymbol{X}_i)}{L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)}. \quad (4)$$

where $L(\text{UB}|\boldsymbol{X}_i)$ is defined as before, and $L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)$ is the probability of the student's response vector at the point $\theta_{i.PP} = \hat{\theta}_{i.PP}$.

Recall that Dimension PP is a bipolar dimension such that respondents who always choose an elaborator response when incorrect will be at the lower end, and respondents who always select a paraphrase response when incorrect will be at the upper end. Therefore, the lower region $\theta_{i.PP} \leq LB$ will be an elaborator region in that test-takers in this region are more likely to choose elaboration responses over paraphrase responses. Conversely, those in the upper region $\theta_{i.PP} \geq UB$ are more likely to choose a paraphrase response. As we have implemented the GLR decision rule, a student is classified as a paraphraser if the likelihood of their response vector is at least 9.0 times greater if $\theta_{i.PP}$ is in the upper region $\theta_{i.PP} \geq UB$ than if it is in the lower region $\theta_{i.PP} \leq LB$. Conversely, the decision rule will classify a person as an elaborator if the likelihood of the student's response vector is nine times greater if $\theta_{i.PP} \leq LB$ than if $\theta_{i.PP} \geq UB$. Thus, we use the GLR statistic to assign process propensity classifications (paraphraser or elaborator) to respondents whom we can classify with reasonable confidence. If we cannot classify a respondent with confidence, they receive a classification of Indeterminate or Not Applicable. Struggling readers are more likely to receive a classification (Paraphrase or Elaborating), because they make more incorrect responses, and they are also more likely to need a classification for the purpose of designing supplementary instruction.

Reliability and Precision

Reliability and Precision

This section reports reliability data for the current, 40 item MOCCA-College forms using data collected during the National Calibration sample 2021-2022 data collection.

Internal Consistency Reliability: Raw Scores and IRT Scores

Table 1 reports the mean and standard deviation (in parentheses) of raw scores for each response type: number of Number Correct responses, number of Paraphrase responses, and number of Elaboration responses by form. Each of these scores has a possible range of 0 – 40 because each form contained 40 items. Across the three forms, the means for each response type are similar. For the Number Correct variable, means ranged from 30.85 – 31.41. In other words, on average, students answered an average of about 75% of the items correctly. For the Paraphrase response, the means ranged from 3.19 – 3.64. On average, across the forms, 8-9% of the test-takers were classified as having a propensity toward Paraphrase responses. Similarly, for the Elaboration response, means ranged from 2.80 to 3.36, and 7-8% of the test-takers were classified as having a propensity toward Elaboration responses. The average number of items completed ranged from 38.71 – 38.85 across the three forms. Although not everyone completed all of the items, most respondents answered nearly all items.

Table 1

Sample Size (N), Means, and Standard Deviations (in Parentheses) of Raw Scores for MOCCA-College by Form

	N	Number Correct	Paraphrase	Elaboration
Form 1	579	30.85 (9.11)	3.64 (4.07)	3.36 (3.38)
Form 2	521	30.91 (9.70)	3.19 (3.66)	3.23 (3.31)
Form 3	514	31.41 (9.47)	3.50 (4.15)	2.80 (3.28)

Table 2 reports the internal consistency reliability (i.e., alpha) for each of the raw scores. The Reliabilities for the Number Correct scores are all at or slightly above .90, whereas those for the Paraphrase response exceed .80. and those for the Elaboration response all exceed .75. All three of the raw scores have internal consistency reliabilities that are good to excellent. This pattern is the same as that found in earlier studies of the MOCCA test for elementary school students: the highest reliabilities were for the Number Correct score followed by the Paraphrase score. (Davison, Biancarosa, Seipel, Carlson, Liu, & Kennedy, 2019). These reliabilities describe the 40 item versions of the test administered in the National Calibration sample 2021 – 2022.

Table 2

Internal Consistency Estimates of Reliability (Coefficient Alpha) by Form for Raw Scores

	Number Correct	Paraphrase	Elaboration
Form 1	.91	.83	.77
Form 2	.90	.81	.77
Form 3	.91	.84	.78

Table 3 reports the marginal reliabilities for the Reading Comprehension Scale Score, which range from .80 - .83 across the three forms. For the sample as a whole, the marginal

reliabilities for the Process Propensity score were low, ranging from .46 - .50. There is a simple reason for these low reliabilities. Each student's Process Propensity score for each respondent is based on a limited number of responses, the number of responses were items they answered incorrectly. As shown in Table 1, on average respondents had provided approximately an average of six to seven incorrect responses, depending on the form.

Table 3

Marginal Reliability Estimate by Form for the IRT Reading Comprehension and Process Propensity Dimensions

	Reading Comprehension	Process Propensity
Form 1	.83	.48
Form 2	.81	.46
Form 3	.80	.50

Appendices C & D shows the histograms of IRT scores for the Reading Comprehension and Process Propensity dimensions, respectively. It also shows, as well as the information functions for each of these dimensions. Table 4 reports the test-retest and alternate forms reliabilities for both the raw scores and the IRT dimensions.

In the test-retest sample, the reliability was .70 for the Number Correct score, and .79 for the Paraphrase score. It was somewhat smaller for the Elaboration score, .62. The test-retest reliability for the Reading Comprehension Dimension was .676, whereas the test-retest corresponding correlation with the Process Propensity dimension was low in the total test-retest sample (.64).

In the alternate-forms sample, the reliability estimates were .70, .80, and .46 for the Number Correct, Paraphrase, and Elaboration raw scores. The alternate forms correlation for the Reading Comprehension dimension was .69. The alternate forms correlation for the Process Propensity dimension was a nonsignificant .14.

Table 4

Test-retest and Alternate Forms Reliability for Raw Scores and IRT Dimensions

Sample	Number Correct	Paraphrase	Elaboration	Reading Comprehension Dimension	Process Propensity Dimension
Test-retest	.70**	.79**	.62**	.76**	.64**
Alternate form	.70**	.80**	.46**	.69**	.14

*$p < .05$, **$p < .01$*

Validity

In this section, we begin by reporting on a multidimensional IRT analysis of the reading comprehension responses to address a central construct validity question: do the narrative and expository items function on the same reading comprehension dimension or two separate comprehension dimensions? In the second portion, we present data addressing the criterion-related and construct validity of MOCCA-College, using data on its correlations with college admissions tests and first-year grade point average (GPA).

MOCCA-College contains both narrative and expository passages. The question naturally arises as to whether responses to both narrative and expository passages are functions of a single reading comprehension dimension. To answer this question, we turned to a confirmatory IRT analysis. For each of the three forms, we fit two confirmatory IRT models. The first was a unidimensional model with all items discriminating along the same dimension. The second was a two-dimensional, simple structure model with all narrative items discriminating along one dimension and with all expository items discriminating along the other. Both were modeled as three-parameter logistic IRT models, but with the guessing parameter having a tight prior (mean = .25, standard deviation = .025). Table 6 shows the statistics used to compare the one- and two-dimensional model fit for each form. These multidimensional models were fitted with the IRTPRO software (Vector Psychometric Group, 2011). Table 5 shows the statistics used to compare the one- and two-dimensional model fit for each form.

Table 5

Dimension Correlations and Fit Measures Comparing 1-Dimensional and 2-Dimensional Solutions

Parameter	Form 1		Form 2		Form 3	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2

ρ		.98		1.00		.97
-2LL	15904.89	15901.29	14155.54	14155.61	13666.76	13658.72
AIC	16066.89	16065.29	14317.54	14319.61	13828.76	13822.72
BIC	16422.50	16425.29	14666.05	14672.43	14175.32	14173.56
RMSEA	.09	.09	.07	.07	.07	.07

Note: ρ = correlation parameter for Reading Comprehension and Process Propensity Dimension, *LR* = likelihood ratio test of null hypothesis that both models fit equally well, *AIC* = Akaike Information Criterion, *BIC* = Bayesian Information Criterion

The first row shows the dimension correlation parameter estimate for each form. These are the correlation parameters from the two-dimensional solutions. These correlations were .98, 1.00, and .97 for Forms 1 – 3 respectively. These estimates suggest that the narrative and expository dimensions in the two-dimensional solutions are virtually identical for all three forms. Using the -2LL statistic in row 2, we computed a likelihood ratio statistic for each form to test the null hypothesis that the one- and two-dimensional solutions fit the data equally well. The statistic was significant ($p < .05$) only for Form 3. The differences in the AIC are extremely small except for Form 3, where the AIC is somewhat smaller for the two-dimensional solution. Differences in the BIC are also small. For forms 1 & 2, the BIC is smaller for the one-dimensional solution, whereas for Form 3 it is smaller for the two-dimensional solution. The RMSEA is identical for the one- and two-dimensional solutions for all forms. While there are some differences in fit that would favor the two-dimensional model for Form 3, the correlation parameters for all Forms (even for Form 3) suggest that the two dimensions of the two-dimensional solution are virtually, if not completely, identical. Furthermore, the RMSEAs are identical for the one- and two-dimensional solutions. These data lead us to conclude that the one-dimensional model is the best model to retain considering both parsimony and fit.

For the reasons cited above, there are large amounts of missing data on the response variables for the Process Propensity items. We have not included an analysis like that in Table 5 for one- and two-dimensional Process Propensity items, because we are unsure how multidimensional IRT algorithms and fit measures perform when there is such a large amount of missing data.

Given the data in Table 5, we came to the conclusion that there is not a distinct difference between the dimensions accounting for reading comprehension success on the narrative and expository items. However, there still may still be differences in the item parameters for the two

item types. Table 6 reports the mean item discrimination and difficulty parameters for expository and narrative items on both the Reading Comprehension Items and the Process Propensity items.

The mean differences between the mean item discrimination parameters were significant for the expository and narrative items along the Reading Comprehension Dimension, but not for the Process Propensity Dimension. This difference between the expository and narrative item discriminations remained significant even after controlling for form and year in which the item was written. Narrative items had a higher mean discrimination on every form.

Table 6

Mean IRT Discrimination (a) and Difficulty (b) parameters for the Reading Comprehension and Process Propensity Dimensions by Form

	Expository Items			Narrative Items		
	Mean	*SD*	*n*	Mean	*SD*	*n*
			Reading Comprehension Dimension			
Form 1						
a						
	1.7029	0.3560	16	1.9404	0.3425	17
b						
	-0.4896	0.6492	16	-1.3368	0.5522	17
Form 2						
a						
	1.5224	0.2623	16	1.7758	0.2528	17
b						
	-0.6595	0.7538	16	-1.3810	0.3650	17
Form 3						

a						
	1.6579	0.3170	18	1.8461	0.2103	16
b						
	-0.7402	0.6077	18	-1.3641	0.3593	16

Process Propensity Dimension

Form 1	Expository Items			Narrative Items		
	Mean	*SD*	*n*	Mean	*SD*	*n*
Form 1						
a						
	0.7415	0.1546	16	0.7667	0.1535	17
b						
	-0.0415	1.3166	16	0.4322	0.6294	17
Form 2						
a						
	0.7385	0.1457	16	0.7940	0.1201	17
b						
	-0.0866	1.2731	16	0.0782	0.9469	17
Form 3						
a						
	0.7539	0.1428	18	0.7901	0.1260	16
b						
	-0.3500	0.7352	18	-0.0204	0.9370	16

These findings have important implications for high school and college students. Most of the course readings for these students contain expository, not narrative material. If there were two distinct dimensions, one for narrative and one for expository, this would raise serious questions about measuring reading comprehension for college students using narrative material. However, these data suggest that both narrative and expository items measure the same reading comprehension ability and, although narrative material is less common in course readings, narrative items actually have higher discrimination for the measurement of the ability used in the comprehension of both narrative and expository material. Along the Process Propensity Dimension, we found no significant differences in the item parameters for expository and narrative items. Thus, both types of items are valid indicators for the measurement of the reading comprehension ability underlying the expository material most common in the course materials of high school and college students.

Construct and Concurrent Validity

Our goal in the validity analyses reported below was to evaluate the criterion-related and construct-related validity of MOCCA-College. In evaluating construct validity, we studied the convergent validity of MOCCA-College by correlating scores with college admissions reading tests or composite scores that include a reading component. We predicted that MOCCA-College would demonstrate discriminant validity by correlating more highly with reading tests than with mathematics tests. We also correlated MOCCA-College with first year college grade-point (GPA). Test scores and GPA were provided by the participating universities. Correlations are reported separately for the two universities that provided data.

Construct Validity

Using admissions test data and grade point averages from two universities, we computed concurrent validity coefficients. Results are reported in Table 7. For the ACT test, we have reported correlations with individual subtests, but for the SAT, we were only able to report correlations with the SAT composite for the one university that provided SAT Total data. One university provided us with SAT Math and a Reading/Writing total (but not Reading alone), neither of which bear directly on the convergent validity of MOCCA-College as a test of reading.

The Reading Comprehension Scale Score was significantly correlated with the ACT Composite score at both universities. It was also significantly correlated with the ACT English test of grammar and rhetorical skills in the second university but not the first. It was significantly correlated with ACT Math and Reading in both universities. The Reading correlations are higher than the Math correlations, providing evidence for the discriminant validity of the test. First term college GPA was significantly correlated with the Reading Comprehension Scale Score in both universities ($p < .05$), but the correlations are small, particularly in the second university. Not shown in Table 7, we had data on the ACT Science Reading test at the second university. The correlation with the Reading Comprehension Scale Score was .44 ($p < .01$). Given that the Science Reading assessment involves substantial passage reading and inference, one would expect a correlation between the ACT Science Reading test and the Reading Comprehension Scale Score based on inferential reading comprehension items.

Table 7

Correlations of MOCCA IRT-based Reading Comprehension (RC) and Process Propensity (PP) scores with College Admissions Assessments and Grade Point Average Data

	ACT Admissions Test				1st Term GPA	High School GPA	SAT Total
	Composite	English	Math	Reading			
University 1							
RC Scale Score	.59**	.08	.43**	.54**	.23**	.35**	.47**
N	81	83	82	82	438	446	215
PP Theta	-.14	.06	-.12	-.13	-.06	.05	-.10
N	79	81	80	80	426	434	209
University 2							
RC Scale Score	.53**	.48**	.37*	.47**	.14*	NA	NA
N	251	240	240	228	289		
PP Theta	-.13	-.17**	-.05	-.15**	.03	NA	NA
N	229	219	219	206	265		

*$p < .05$, **$p < .01$, NA = no data available

The Process Propensity dimension was not significantly correlated with any GPA or admissions test scores in University 1. However, it did correlate significantly ($p < .05$) and negatively with the ACT English and ACT Reading scores in University 2. Given the bipolar nature of the Process Propensity dimension with Elaboration at the negative end, the negative correlation suggests that test-takers in University 2 with a propensity toward Elaboration rather than Paraphrases tended to have somewhat higher scores on the ACT English and ACT Reading assessments than did those with a propensity towards Paraphrases in University 2.

Table 8 shows the correlations of the college admissions tests and GPA with MOCCA-College raw scores. Because raw scores are not equated across forms, results are given by form. In Table 8, sample sizes vary by form and criterion variable. Readers are encouraged to look at consistent trends across forms as much as statistical tests.

Table 8

Correlations of MOCCA Raw Scores with College Admissions Tests and GPA by University and Form

	ACT						
	ACT Admissions Test						SAT Total

	Composite	English	Math	Reading	1st Term GPA	High School GPA	
			University 1				
Form 1							
CCI	.58**	.54**	.54**	.55**	.16	.33**	.62**
PAR	-.62**	-.47*	-.60**	-.58**	-.18*	-.28**	-.53**
ELA	-.54**	-.48*	-.45*	-.45*	-.07	-.19*	-.59**
N	25	25	25	25	147	147	71
Form 2							
CCI	.51**	.51**	.27	.51**	.29*	.,36*	.28*
PAR	-.58**	-.55**	-.38**	-.53**	-.20*	-.46**	-.25*
ELA	-.55**	-.41*	-.37*	-.60**	-.24**	-.27**	-.30*
N	32	33	33	33	144	147	69
Form 3							
CCI	.30	.14	.16	.33	.21*	.31**	.39**
PAR	-.11	-.06	-.06	-.05	-.26**	-.29**	-.39**
ELA	-.18	-.18	-.13	-.18	-.22**	-.34**	-.30*
N	24	25	24	24	147	152	75
			University 2				
Form 1							
CCI	.42**	.41**	.42**	.21	.34**	NA	NA
PAR	-..57**	-.51**	-.48**	-.42**	-.25**		
ELA	-.46**	-.40**	-.43**	-.37**	-.25**		
N	80	77	77	71	94		
Form 2							
CCI	.33**	.31**	.22	.32**	-.11	NA	NA
PAR	-.40**	-.39**	-.21	-.42**	.10		
ELA	-.47*	-.40*	-.30	-.46*	.11		
N	80	77	77	75	94		
Form 3							
CCI	.30**	.31**	.11	.32**	.24*		
PAR	-.40**	-.42**	-.15	-.46**	-.25**		
ELA	-.25*	-.24*	-.08	-.29*	-.22*		
N	81	78	78	76	92		

Although there are some exceptions to these trends, the MOCCA-College scores are usually significantly related to the college admissions tests. The CCI scorer is consistently

positively correlated with the admissions tests while the Paraphrase and Elaboration scores are negatively related. The MOCCA-College raw scores tend to be as or more highly correlated (in absolute value) with the ACT composite and the SAT total than with the other variables. The MOCCA-College variables tend to be more highly correlated with the ACT Reading Assessment than the ACT Math which supports the discriminant validity of the MOCCA test. However, Form 1 is an exception to this trend in both universities as it is as correlated with the Math subtest as it is with the Reading subtest. In University 1, the sample sizes for the ACT tests are small (24 - 33) for all forms and should be viewed with some skepticism.

The MOCCA-College raw scores variables tend to be more highly correlated with high school GPA than with first term GPA. In the one university for which high school GPA data were available, all of the raw score variables are significantly ($p < .05$) related to high school GPA for all three forms. First term GPA was significantly related to MOCCA-College raw scores first term GPA for two of the three forms in both universities, but the MOCCA-College variables generally accounted for less than 10% of the variance in first term GPA. Generally, these results tend to confirm the significant but small association between college success and reading comprehension as found by Clinton-Lisell, Taylor, Carlson, Davison, & Seipel (2022) in their meta-analysis.

Fairness

The fairness of MOCCA was examined in two ways. First, all items were subjected to a sensitivity review by teachers as part of the content review described above. In this step, items were evaluated by a panel of teachers as to bias, whether the content would be offensive or would provide an advantage to one demographic group over others. Second, items were examined for differential item functioning by gender and by race/ethnicity (Blacks vs. Whites and Hispanics vs. Whites). Blacks and Hispanics were chosen because they are the largest of the underrepresented race/ethnicity groups in higher education.

Average Score Differences by Gender and Race/Ethnicity

To test for demographic differences, we performed analyses of variance that included gender or race/ethnicity. Table 9 shows means for Dimension RC Scale Score and Dimension PP θ scores, along with the percentage of test-takers classified as having a Paraphrase or Elaboration propensity, by form, gender, and race/ethnicity. Table 9 also shows the percentage of test-takers classified as having a Paraphrase or Elaboration propensity by Form, Sex, and Race/ethnicity.

Males and females did not differ significantly on Dimension RC Scale Score ($t = .269$, $p = .258$) or Dimension PP θ scores ($t = 1.159$, $p = .247$). There were, however, significant differences by race/ethnicity for the two largest underrepresented race/ethnicity groups. For Blacks and Whites, the mean differences shown in Table 9 were significant on both the Reading Comprehension Dimension ($t = 4.71$, $p < .001$) and the Process Propensity Dimension ($t = 4.65$, $p < .001$). For Hispanics and Whites, the mean difference was significant on the Reading Comprehension Dimension ($t = 4.49$. $p < .001$) but the difference on the Process Propensity Dimension was not significant ($t = 0.39$, $p = .70$). American Indians are a third underrepresented group, but our sample size for this group was small (20).

Table 9

Descriptive Data for Students Taking at Least Ten Items by Form, Gender, and Race/ethnicity

	N	% of Sample	SS RC Mean (SD)	Theta PP Mean (SD)	Paraphraser Propensity %	Elaborator Propensity %
Overall	1614	100	502.11 (151.85)	0.02 (1.40)	9%	6%
Form						
Form 1	579	36%	502.28 (149.69)	0.02 (1.50)	10%	7%
Form 2	521	32%	501.61 (154.98)	0.05 (1.37)	8%	5%
Form 3	514	32%	502.43 (151.34)	0.00 (1.31)	9%	6%
Overall	1614	100%	502.11 (151.85)	0.02 (1.40)	9%	6%
			Gender			
Female	934	67%	505.66 (142.17)	-0.03 (1.41)	8%	6%
Male	460	33%	507.93 (160.52)	0.06 (1.39)	9%	5%
			Race/ethnicity			
Am.A Indian	20	2%	485.64	-0.45		
Asian	140	12%	487.94	-0.22		
Black	148	13%%	476.64 (151.01)	-0.31 (1.48)	7%	8%
Hispanic	169	14%	485.67 (142.25)	0.07 (1.31)	12%	6%
White	698	59%	544.00 (148.78)	0.07 (1.40)	8%	5%

Males and females did not differ significantly on either the Dimension RC Scale Score (t = .269, p = .258) or Dimension PP θ scores (t = 1.159, p = .247). There were, however, significant differences by race/ethnicity for the two largest underrepresented race/ethnicity groups. For Blacks and Whites, mean differences were significant on both the Reading Comprehension Dimension (t = 4.71, p < .001) and the Process Propensity Dimension (t = 4.65, p < .001). For Hispanics and Whites, the mean difference was significant on the Reading Comprehension

Dimension (t = 4.49. p < .001) but not on the Process Propensity Dimension (t = 0.39, p = .70) American Indians are a third underrepresented group, but our sample size for this group was insufficient (n =20) for formal statistical tests of average group differences.

Differential Item Functioning (DIF)

DIF analyses were completed for all items in the calibration sample using the Mantel-Haenszel (M-H) statistic to test the hypothesis of no DIF for each item, using alpha = .05. Analyses were conducted by gender, for Black vs. White students, and for Hispanic vs. White students in the calibration sample. These two comparisons were chosen because the two comparison groups (Blacks and Hispanic students) are underrepresented groups in higher education and their means differed significantly from that of White students. American Indians are also an underrepresented group in higher education, but their numbers in our sample (n = 20) were too small for a DIF analysis.

In the gender analysis, only two items yielded significant M-H statistics, leading to rejection of the null hypothesis with an alpha level of .05. One item favored females and one item favored males. Given the large number of items tested, two significant statistics results is less than expected from sampling error given alpha of .05. Inspection of the items' content uncovered no offensiveness or clear advantages for one group over the other. Thus, given the small proportion of significant results, the equal number of items favoring males and females, and no evidence of content bias, both items have been retained in the item pool.

In the Black vs. White and Hispanic vs. White analyses by race/ethnicity, there were no significant M-H statistics at an alpha of .05. Therefore, no items were rejected for reasons of race/ethnicity bias. It should be noted, however that the samples of Blacks and Hispanic students were not large, and therefore the race/ethnicity DIF tests were only modestly powered.

Summary and Conclusions

Reliability and validity evidence for MOCCA-College is promising. All three raw scores show good internal consistency, with reliability data, the raw scores all had good reliabilities of .70 or above. Reliabilities were highest for the Number Correct (i.e., the Causally Coherent Inference) score (.9 or above), followed by the Paraphrase score (.8 or above), and the Elaboration score (upper .70s). The marginal reliabilities for the Reading Comprehension Scale Score were in the low .80s, while those for the Process Propensity Score were below .6. The test-retest and alternate forms reliabilities for the Number Correct and Paraphrase score exceeded .70, but not the Elaboration Score. The test–retest correlation for the Reading Comprehension Scale Score was also above .70, but the alternate-forms reliability was slightly below this benchmark at .69. The test-retest reliability of the Process Propensity Dimension was above .6 but below .7, while the alternate-forms reliability was very low. One implication of the low reliabilities for the Process Propensity Dimension is that very few students can are classified confidently as having a Paraphrase or Elaboration propensity, as shown in Table 9.

From a validly perspective, we began by investigating the question of whether the reading dimension underlying narrative items was the same as the dimension underlying the expository items. The various fit measures provided little evidence for the two-dimensional

structure, with the exception of Form 3. However, even for Form 3, the dimension correlation parameter of the two–dimensional solution suggested that the dimension underlying the narrative items was nearly identical to that underlying the expository dimension. Thus, the data gave little support to the hypothesis that the dimensions underlying narrative and expository items were distinct. Examination of the IRT parameters revealed that the narrative items had higher average discrimination parameters than the expository items and lower average difficulty parameters.

Next, we examined the construct and criterion-related validity of MOCCA-College by correlating scores with college admissions test scores, first term college GPA and high school GPA. The MOCCA-College Reading Comprehension Dimension was consistently correlated with the ACT composite, SAT composite, and ACT reading, results that support the convergent validity of the test. With some exceptions, it was generally more highly correlated with the ACT reading score than the ACT math score, results that provide some support for the discriminant validity. In what might at first seem to challenge its discriminant validity, in the one university for which we had data, the MOCCA-College Reading Comprehension Scale Score was significantly correlated with the ACT Science Score, but the ACT Science test is a science reading test involving inferencing from reading passages. Therefore, the correlation with the science test actually supports the concurrent validity of the test.

Analysis of mean scores on the Reading Comprehension Dimension for the two largest underrepresented race/ethnicity groups revealed significant differences in mean scores between Blacks and Whites and Hispanics and Whites. However, in the DIF analyses for these two groups, no items displayed DIF ($p < .05$). The sample sizes for Blacks and Hispanics were not large, so the tests for race/ethnicity DIF would have had only modest power. DIF analyses were also conducted for females and males. Only two of 100 items displayed significant DIF with one item favoring males and one item favoring females. Content analyses of these two items did not reveal any offensive content or any content aspect that might give an advantage.

Overall, these findings support MOCCA-College as a valid and reliable assessment of the cognitive processes of reading comprehension. Further, three forms containing narrative and expository items that range from FK 6-14 align with the types of topics that students in postsecondary education would read. Thus, we feel confident that the college version of MOCCA-College can be used to identify causal comprehension difficulties for postsecondary students in postsecondary education.

Although our data support the use of this new version of MOCCA-College, we encourage instructors, counselors, administrators, and other staff to use other reading assessments to corroborate MOCCA-College results and ensure they have sufficient information to accurately understand the students' needs. Nonetheless, MOCCA-College results, along with the instructional recommendations, can be used to help postsecondary students find ways to improve their reading comprehension skills and strive in education and life.

References

August, D., Francis, D. J., Hsu, H. Y. A., & Snow, C. E. (2006). Assessing reading comprehension in bilinguals. *The Elementary School Journal*, *107*(2), 221-238.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at-risk. *Educational and Psychological Measurement*, 79(1), 65-84. doi: 10.1177/0013164418763255Yoon, H.-J., Seipel, B., Carlson, S. E., Liu, B., & Davison, M. L. (in press). Constructing subscores that add validity: A case study identifying students at risk. *Educational and Psychological Measurement*.

Cao, Y., & Kim, Y. S. G. (2021). Is retell a valid measure of reading comprehension?. *Educational research review*, *32*, 100375.

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

Clinton-Lisell, V., Taylor, T., Carlson, S. E., Davison, M. L., & Seipel, B. (2022). Performance on reading comprehension assessments and college achievement: A meta-analysis, J*ournal of College Reading and Learning, 52:3*, 191-211, DOI: 10.1080/10790195.2022.2062626

Davis, B. J., Johnston, A. M., Barnes, M. A., & Desrochers, A. (2007). *Bridging inferences in children from grades three to eight*. Poster presented at the Annual Convention of Canadian Psychological Association. Halifax, Nova Scotia, Canada.

Davison, M. L., Biancarosa, G., Carlson, S. E., & Seipel, B. (2018). Preliminary findings on the computer-administered Multiple-choice Online Causal Comprehension Assessment, a diagnostic reading comprehension test. *Assessment for Effective Intervention*, *43(3)*, 169 – 181.

Davison, M. L., Davenport, E. C. Jr., Chang, Y.-F., Vue, K., & Su, S. (2015). Criterion-related Validity: Assessing the Value of Subscores. *Journal of Educational Measurement*, *52*, 263 – 279.

Davison, M. L. & Davenport, E. C. Jr. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods*, *7*(4), 468-484.

Dorans, N. & Holland, P. W. (1988). Differential item performance and the Mantel-Haenszel procedure. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35 – 66). Hillsdale NJ, Erlbaum.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*(2), 627-635.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Liu, B., Kennedy, P. C., Seipel, B., Carlson, S. E., Biancarosa, G., & Davison, M. L. (under review). Can we learn from student mistakes in a formative, reading comprehension assessment?

McMaster, K. L., den Broek, P. van, A. Espin, C., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*(1), 100–111. https://doi.org/10.1016/j.lindif.2011.11.017

Pavias, M., van den Broek, P., Hickendorff, M., Beker, K., & Van Leijenhorst, L. (2016). Effects of social-cognitive processing demands and structural importance on narrative recall: Differences between children, adolescents, and adults. Discourse Processes, 53(5-6), 488-512. https://doi.org/10.1080/0163853X.2016.1171070

Pike, M. M., Barnes, M. A., & Barron, R. W. (2010). The role of illustrations in children's inferential comprehension. *Journal of experimental child psychology*, *105*(3), 243-255.

Samuels, S. J., & Flor, R.F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly*, *13(2)*, p. 107 — 121

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, *23*(4), 304-325.

Smarter Balanced Assessment Consortium (2015). *English language arts & literacy computer adaptive test (CAT) and performance task (PT) stimulus specifications*. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2015/08/ELA-Stimulus-Specifications.pdf.

Smarter Balanced Assessment Consortium (October 6, 2016). *Smarter Balanced Assessment Consortium: 2014-2015 Technical Report*. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better DECISIONS through science. *Scientific American*, *283*, 82-87.

Tennessee Department of Education (2018). *Overview of testing in Tennessee.* Retrieved from https://www.tn.gov/education/assessment/testing-overview.html, August 30, 2018.

Thorndike, R. M., & Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education* (8th ed.). New York, NY: Pearson

van den Broek, P., Espin, C., McMaster, K., & Helder, A. (2017). Developing reading

comprehension interventions: Perspectives from theory and practice. In E. Segers & P. W. van den Broek (Eds.), *Developmental Perspectives in Written Language and Literacy* (pp. 85–102). John Benjamins Publishing Company.

van den Broek, P., White, M. J., Kendeou, P., & Carlson, S. (2009). Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. In R. Wagner, C. Schatschneider, & C. Phythian-Sence (Eds.), Beyond Decoding: The Behavioral and Biological Foundations of Reading Comprehension, (pp.107-123), Guilford Publications.

Vector Psychometric Group. IRTPRO guide, 2011. Retrieved from http://vpgcentral.com, January 15, 2021.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427–50. http://dx.doi.org/10.1007/bf02294627.

Yeari, M., & Lavie, A. (2021). The role of surface text processing in centrality deficit and poor text comprehension of adolescents with Attention Deficit Hyperactivity Disorder: A think-aloud study. Learning Disabilities Research & Practice, 36(1), 40-55. https://doi.org/10.1111/ldrp.12237

Appendix A: Content Review Panel Details

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

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

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

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

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

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

Appendix B: Norm Table for Reading Comprehension Scale Scores

Percentile Rank	Lower Bound	Upper Bound
1	50	147
2	148	170
3	171	192
4	193	213
5	214	235
6	236	251
7	252	267
8	268	283
9	284	296
10	297	306
11	307	317
12	318	324
13	325	331
14	332	340

Percentile Rank	Lower Bound	Upper Bound
15	341	345
16	346	353
17	354	361
18	362	367
19	368	375
20	376	379
21	380	384
22	385	388
23	389	396
24	397	403
25	404	409
26	410	415
27	416	417
28	418	420
29	421	425
30	426	429

Percentile Rank	Lower Bound	Upper Bound
31	430	435
32	436	438
33	439	442
34	443	447
35	448	451
36	452	455
37	456	460
38	461	464
39	465	468
40	469	473
41	474	478
42	479	482
43	483	486
44	487	490
45	491	494
46	495	497

Percentile Rank	Lower Bound	Upper Bound
47	498	499
48	500	503
49	504	506
50	507	509
51	510	512
52	513	515
53	516	518
54	519	521
55	522	523
56	524	527
57	528	530
58	531	533
59	534	535
60	536	540
61	541	545
62	546	548

Percentile Rank	Lower Bound	Upper Bound
63	549	550
64	551	555
65	556	559
66	560	563
67	564	566
68	567	570
69	571	576
70	577	581
71	582	585
72	586	587
73	588	591
74	592	596
75	597	600
76	601	604
77	605	608
78	609	612

Percentile Rank	Lower Bound	Upper Bound
79	613	618
80	619	619
81	620	627
82	628	634
83	635	636
84	637	643
85	644	648
86	649	659
87	660	666
88	667	673
89	674	677
90	678	685
91	686	688
92	689	699
93	700	703
94	704	739

Percentile Rank	Lower Bound	Upper Bound
95	740	771
96	772	780
97	781	799
98	800	818
99	819	950

Appendix C: Item Response Theory Statistics for the Reading Comprehension Dimension

Table C.1

Item Statistics for Dimension 1

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CN132	0.886	0.485	2.318	-1.267	0.196
CE091	0.855	0.442	2.468	-0.995	0.212
OM349	0.884	0.375	2.351	-1.064	0.346
CN108	0.922	0.442	2.191	-1.537	0.247
CE048	0.857	0.377	1.710	-1.160	0.248
CN121	0.915	0.425	1.885	-1.581	0.244
CE042	0.846	0.503	2.343	-0.979	0.233
CN131	0.802	0.431	1.772	-0.822	0.244

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE098	0.906	0.484	2.073	-1.437	0.245
OM308	0.873	0.476	1.994	-1.204	0.244
OM312	0.865	0.461	1.940	-1.169	0.245
OM420	0.925	0.414	1.760	-1.718	0.249
CE012	0.717	0.322	1.355	-0.436	0.252
CN103	0.488	0.329	1.862	0.482	0.225
CE084	0.842	0.452	1.882	-1.024	0.248
OM335	0.912	0.503	2.025	-1.504	0.245
CE025	0.684	0.359	1.464	-0.270	0.247
CE090	0.661	0.322	1.502	-0.125	0.254

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE088	0.855	0.503	2.287	-0.999	0.253
CE024	0.722	0.475	2.062	-0.384	0.243
CN104	0.950	0.514	2.114	-1.886	0.246
CN114	0.943	0.520	2.877	-1.658	0.239
CE080	0.620	0.287	1.275	0.037	0.248
CE043	0.863	0.440	1.746	-1.186	0.249
CN112	0.796	0.362	1.462	-0.886	0.244
CE094	0.643	0.383	1.467	-0.110	0.241
CE033	0.775	0.424	1.552	-0.733	0.246
CN111	0.918	0.425	1.819	-1.633	0.247

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE061	0.899	0.418	1.993	-1.393	0.251
OM276	0.801	0.391	1.399	-0.932	0.249
CE032	0.394	0.210	1.351	1.199	0.239
OM259	0.896	0.538	2.035	-1.381	0.243
CE054	0.911	0.441	2.271	-1.426	0.247
CE073	0.761	0.409	1.689	-0.623	0.247
OM136	0.903	0.247	1.360	-1.721	0.251
CE063	0.590	0.331	1.544	0.127	0.239
OM204	0.875	0.393	2.038	-1.180	0.253
CN109	0.931	0.352	1.835	-1.739	0.250

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE056	0.785	0.387	1.645	-0.758	0.246
OM188	0.898	0.400	1.802	-1.459	0.246
CN126	0.886	0.380	1.618	-1.433	0.247
OM036	0.874	0.326	1.488	-1.378	0.250
CE050	0.751	0.314	1.516	-0.589	0.251
CN124	0.924	0.493	2.188	-1.580	0.245
CE053	0.893	0.241	1.293	-1.651	0.254
OM022	0.828	0.365	1.664	-0.985	0.250
CE100	0.850	0.435	1.693	-1.140	0.245
OM407	0.928	0.477	1.829	-1.737	0.246

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE039	0.349	0.143	1.624	1.259	0.229
CN106	0.746	0.388	1.631	-0.552	0.248
OM159	0.889	0.347	1.463	-1.524	0.250
CE008	0.598	0.158	1.064	0.267	0.266
CE020	0.782	0.341	1.449	-0.766	0.254
CE068	0.878	0.413	1.717	-1.303	0.251
CN116	0.935	0.466	1.892	-1.770	0.249
CE005	0.623	0.301	1.387	0.017	0.249
OM156	0.903	0.534	1.916	-1.492	0.243
CE002	0.618	0.200	1.234	0.149	0.268

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
OM252	0.939	0.541	2.194	-1.753	0.245
CE075	0.780	0.440	1.543	-0.774	0.244
CN133	0.921	0.424	1.672	-1.705	0.252
CE003	0.791	0.389	1.496	-0.851	0.246
OM359	0.852	0.481	1.729	-1.163	0.243
CE078	0.844	0.418	1.539	-1.162	0.246
OM470	0.933	0.568	2.173	-1.684	0.244
CN110	0.919	0.421	1.699	-1.688	0.248
CE059	0.851	0.360	1.473	-1.226	0.249
CE093	0.781	0.286	1.427	-0.768	0.254

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE046	0.753	0.364	1.630	-0.586	0.250
OM346	0.813	0.421	1.533	-0.966	0.246
OM376	0.914	0.300	1.503	-1.741	0.251
CE026	0.825	0.390	1.694	-0.985	0.245
OM016	0.951	0.354	1.800	-2.020	0.250
OM080	0.911	0.489	1.970	-1.529	0.246
OM026	0.935	0.468	2.006	-1.754	0.247
CE041	0.646	0.352	1.561	-0.093	0.245
OM367	0.937	0.397	1.747	-1.873	0.248
CE011	0.844	0.359	1.569	-1.115	0.252

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE049	0.690	0.180	1.041	-0.289	0.264
OM479	0.840	0.471	1.673	-1.092	0.244
CE099	0.693	0.365	1.398	-0.347	0.244
CN102	0.920	0.507	1.912	-1.631	0.246
CE055	0.854	0.530	1.891	-1.115	0.245
CN113	0.808	0.486	1.675	-0.915	0.240
CE092	0.704	0.398	1.779	-0.322	0.248
OM334	0.858	0.551	1.866	-1.172	0.241
CE004	0.511	0.157	1.316	0.686	0.262
CE027	0.602	0.256	1.454	0.149	0.254

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CE069	0.865	0.496	1.752	-1.226	0.246
CE013	0.898	0.542	2.050	-1.401	0.245
OM049	0.819	0.487	1.768	-0.945	0.242
OM157	0.869	0.466	1.672	-1.287	0.246
CE001	0.763	0.391	1.497	-0.672	0.247
CN107	0.846	0.519	2.121	-1.013	0.241
CN134	0.852	0.437	1.632	-1.173	0.246
CE060	0.882	0.420	1.779	-1.329	0.248
CE081	0.709	0.256	1.159	-0.426	0.254
CE035	0.893	0.452	1.740	-1.423	0.249

Item ID	Proportion Correct	Item Total Correlation	IRT Discrimi-nation	IRT Difficulty	IRT Asymptote
CN128	0.913	0.487	1.966	-1.537	0.246
OM425	0.846	0.405	1.875	-1.079	0.242

Note. *P* = proportion correct, *R* = correlation of item with theta, *a* = discrimination, *b* = difficulty, and *c* = lower asymptote parameter.

Figure C.1

Histogram for the Person Parameters of the Reading Comprehension Dimension

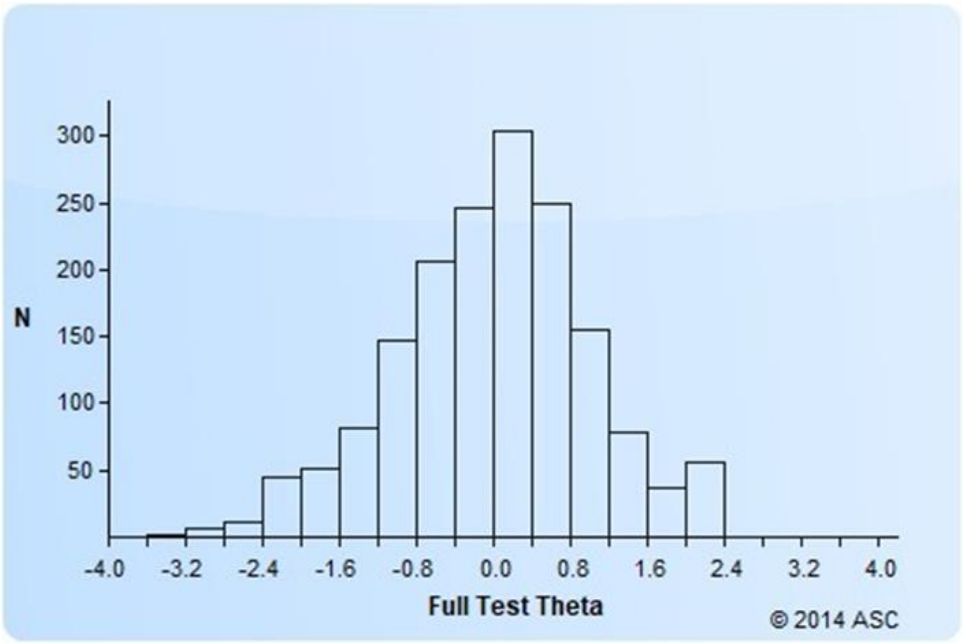


Figure C.2

Histogram for the Item Discrimination Parameters

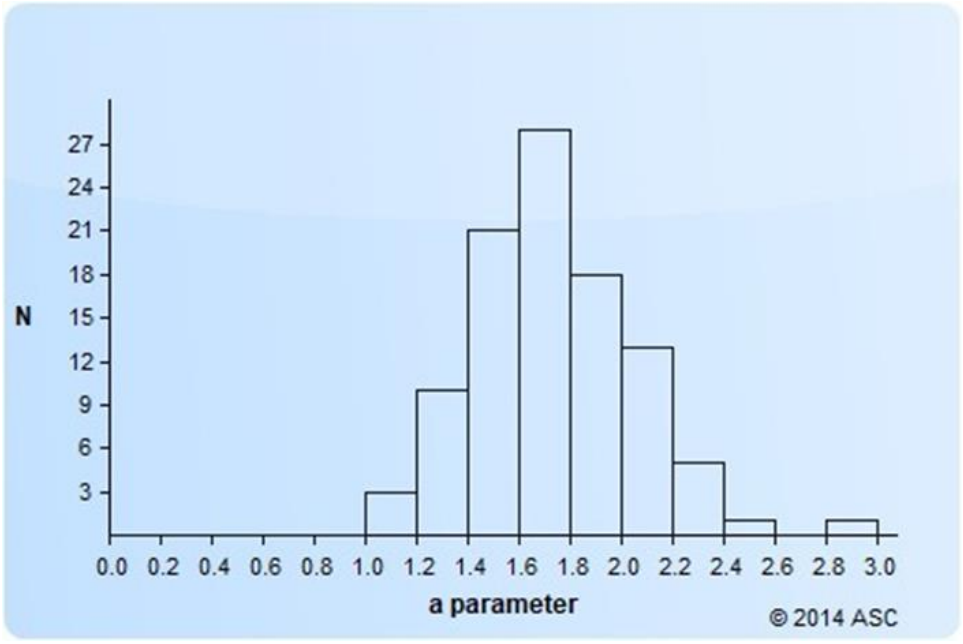


Figure C.3

Histogram for the Item Difficulty Parameters

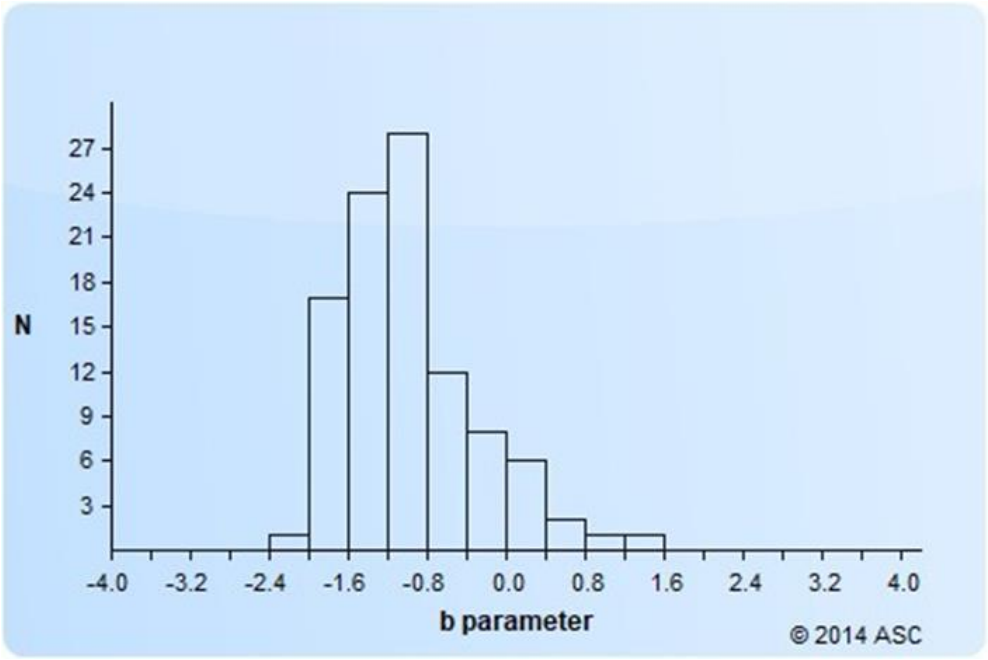


Figure C.4

Histogram for the Item Lower Asymptote Parameters

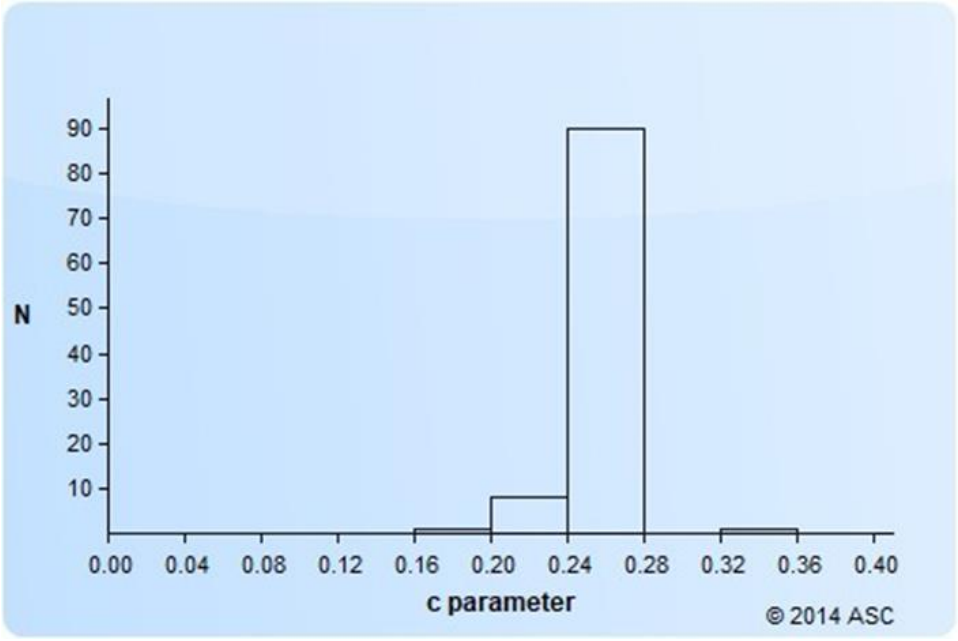
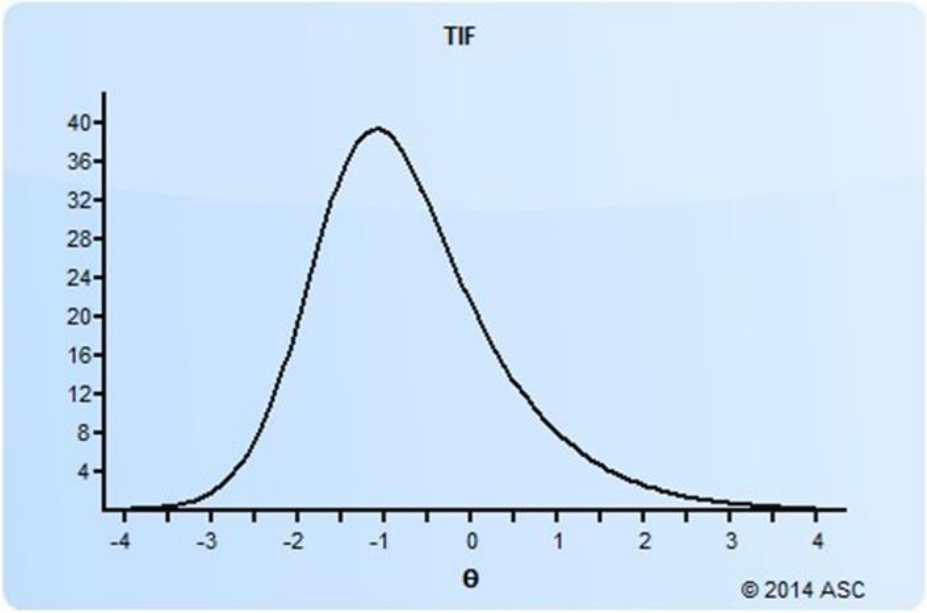


Figure C.5

Test Information Function for the Reading Comprehension Dimensions



Appendix D: Item Statistics for the Process Propensity Dimension

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
CN132	0.515	0.321	0.607	0.035
CE091	0.614	0.125	0.668	-0.555
OM349	0.448	0.223	0.566	0.520
CN108	0.545	0.237	0.761	0.032
CE048	0.713	-0.119	0.871	-0.866
CN121	0.245	0.227	0.995	1.347
CE042	0.253	0.263	0.524	2.239
CN131	0.593	-0.038	0.705	-0.285
CE098	0.434	0.248	0.746	0.560

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM308	0.370	0.188	0.764	0.910
OM312	0.464	0.191	0.549	0.426
OM420	0.381	0.440	1.042	0.735
CE012	0.868	-0.235	0.808	-2.248
CN103	0.284	0.041	0.652	1.734
CE084	0.692	-0.123	0.614	-1.124
OM335	0.510	0.172	0.830	0.118
CE025	0.158	0.302	0.742	2.552
CE090	0.683	0.037	0.900	-0.823
CE088	0.597	-0.001	0.661	-0.447

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
CE024	0.542	0.030	0.572	-0.077
CN104	0.724	0.158	0.994	-0.843
CN114	0.522	0.044	0.615	-0.003
CE080	0.808	0.047	0.920	-1.549
CE043	0.390	0.172	0.817	0.706
CN112	0.402	0.191	0.608	0.849
CE094	0.608	0.061	0.715	-0.472
CE033	0.540	0.190	0.874	0.014
CN111	0.500	0.000	0.682	0.256
CE061	0.660	-0.013	0.837	-0.580

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM276	0.393	0.114	0.671	0.844
CE032	0.402	0.081	0.485	0.968
OM259	0.500	-0.228	0.621	0.218
CE054	0.557	0.017	0.658	-0.206
CE073	0.681	-0.082	0.916	-0.643
OM136	0.364	0.398	0.918	0.654
CE063	0.405	0.255	0.475	0.952
OM204	0.603	-0.108	0.712	-0.411
CN109	0.590	0.340	0.736	-0.236
CE056	0.793	-0.233	0.831	-1.423

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM188	0.448	0.321	0.897	0.555
CN126	0.356	0.360	0.728	0.914
OM036	0.394	0.251	0.816	0.731
CE050	0.430	-0.097	0.527	0.659
CN124	0.308	0.152	0.893	0.963
CE053	0.764	0.009	0.960	-1.048
OM022	0.716	-0.253	0.767	-1.036
CE100	0.636	-0.108	0.740	-0.593
OM407	0.216	0.185	0.813	1.541
CE039	0.084	0.459	0.786	3.290

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
CN106	0.338	0.235	0.588	1.252
OM159	0.895	-0.222	1.023	-1.999
CE008	0.796	-0.304	0.618	-2.115
CE020	0.288	0.282	0.732	1.329
CE068	0.355	0.182	0.844	0.821
CN116	0.455	0.189	0.859	0.282
CE005	0.513	0.164	0.699	-0.021
OM156	0.592	-0.136	0.640	-0.418
CE002	0.655	-0.021	0.676	-0.890
OM252	0.516	0.256	0.781	-0.078

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
CE075	0.571	-0.038	0.740	-0.269
CN133	0.600	-0.038	0.821	-0.380
CE003	0.670	-0.387	0.588	-0.997
OM359	0.387	-0.048	0.724	0.805
CE078	0.544	-0.012	0.650	-0.026
OM470	0.382	0.125	0.849	0.733
CN110	0.756	-0.027	0.988	-1.061
CE059	0.880	-0.296	1.131	-1.684
CE093	0.477	0.049	0.728	0.215
CE046	0.484	0.039	0.741	0.149

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM346	0.625	-0.010	0.882	-0.506
OM376	0.523	0.337	0.755	-0.082
CE026	0.371	-0.119	0.527	1.092
OM016	0.400	0.273	0.853	0.586
OM080	0.511	-0.134	0.964	0.041
OM026	0.636	-0.177	0.653	-0.443
CE041	0.654	-0.183	0.556	-0.981
OM367	0.375	0.073	0.746	0.771
CE011	0.709	-0.085	0.866	-0.800
CE049	0.523	0.319	1.023	-0.006

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM479	0.512	0.250	0.762	0.206
CE099	0.601	0.136	0.764	-0.389
CN102	0.425	0.363	0.831	0.458
CE055	0.767	-0.103	0.920	-1.182
CN113	0.792	-0.045	1.012	-1.197
CE092	0.784	0.028	0.914	-1.320
OM334	0.394	0.241	0.631	0.847
CE004	0.624	-0.084	0.662	-0.639
CE027	0.312	0.318	0.863	1.189
CE069	0.765	-0.150	0.849	-1.232

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
CE013	0.529	0.226	0.871	-0.020
OM049	0.209	0.331	0.916	1.779
OM157	0.515	0.114	0.842	0.148
CE001	0.559	0.131	0.715	-0.076
CN107	0.494	0.046	0.669	0.225
CN134	0.851	-0.246	0.890	-1.728
CE060	0.508	-0.029	0.634	0.271
CE081	0.493	0.203	0.621	0.179
CE035	0.556	0.145	0.672	-0.235
CN128	0.795	0.098	0.851	-1.471

Item ID	Proportion Paraphrase	Item Total Correlation	IRT Discrimination	IRT Difficulty
OM425	0.705	-0.005	0.702	-0.986

Note. *P* = proportion choosing Paraphrase rather than Elaboration when responding incorrectly, *R* = correlation of theta with item, *a* = item discrimination, -*b* = item difficulty

Figure D.1

Histogram for the Person Parameters of the Process Propensity Dimension

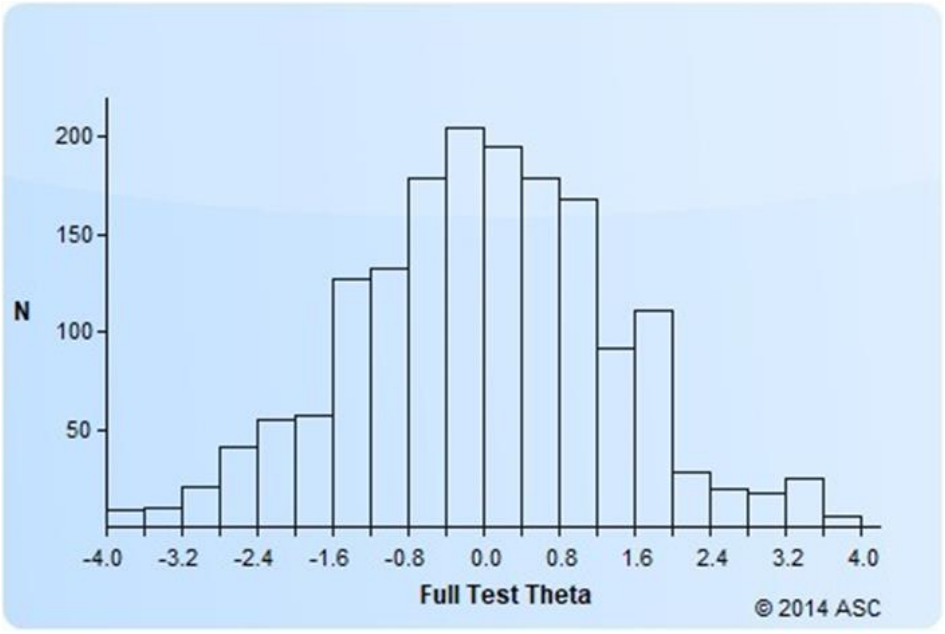


Figure D.2

Histogram of the Item Discrimination Parameters for the Process Propensity Dimension

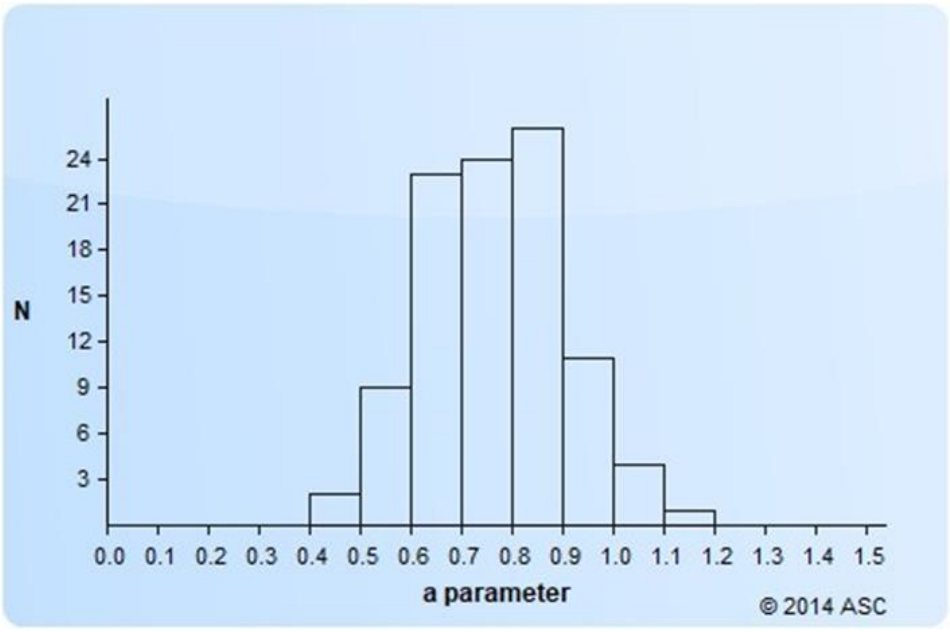


Figure D.3

Histogram of the Item Difficulty Parameters for the Process Propensity Dimensions

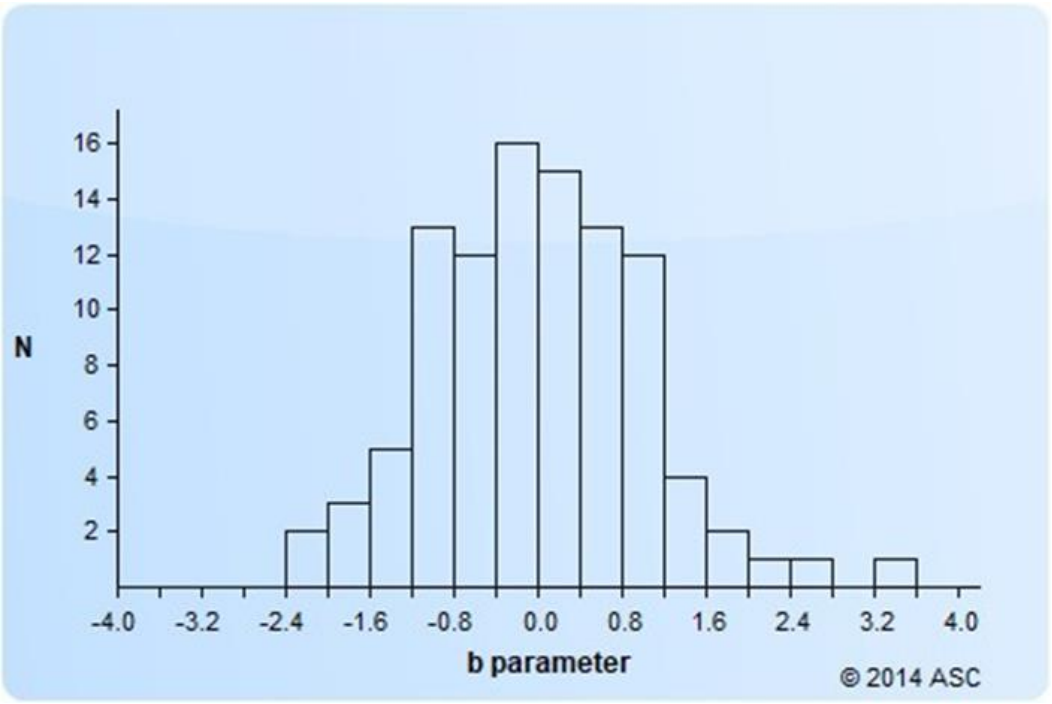


Figure D.4

Histogram of the Test Information Function for the Process Propensity Dimension