Mathematics Performance and Cognition (MPAC) Interview

Measuring First- and Second-Grade Student Achievement in Number, Operations, and Equality in Spring 2014

Robert C. Schoen
Mark LaVenia
Zachary M. Champagne
Kristy Farina

FEBRUARY 2016
PUBLIC VERSION
Research Report No. 2016-01
(updated August 2017)

The research and development reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award No. R305A120781 to Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Suggested citation: Schoen, R. C., LaVenia, M., Champagne, Z. M., & Farina, K. (2016). *Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2014* (Research Report No. 2016–01). Tallahassee, FL: Learning Systems Institute, Florida State University. doi: 10.1725/fsu.1493238156

Detailed information about items are not included in this report. This information was removed in order to release the psychometric report and maintain test security. Requests to view the full report should be directed to Robert Schoen (rschoen@lsi.fsu.edu)

Mathematics Performance and Cognition (MPAC) Interview

Measuring First- and Second-Grade Student Achievement in Number, Operations, and Equality in Spring 2014

Research Report No. 2016–01

Robert C. Schoen

Mark LaVenia

Zachary M. Champagne

Kristy Farina

February 2016

(updated August 31, 2017)

Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM)
Learning Systems Institute
Florida State University
Tallahassee, FL 32306
(850) 644-2570

Acknowledgements

Apart from the critically important support of the Institute of Education Sciences, the successful development and implementation of this interview involved many experts in mathematics education and many more students. Some of the key people involved with the development and field-testing of the interview are listed here along with their roles in the endeavor.

Robert Schoen was integrally involved with designing and implementing the interview, training interviewers, creating the coding system, determining the appropriate structure and specification for the measure models, interpretation of results, and report writing. Mark LaVenia performed the data analysis for the 2-parameter logistic model, item factor analysis, and correlations among various tests and contributed to writing the report. In addition to managing the data, Kristy Farina designed and maintained the data-entry system, assisted with creating the coding system, trained the interviewers on the data-entry system, generated descriptive statistics, and assisted with report writing. Zachary Champagne conducted video coding of interviews and assisted with writing of the report.

We would like to acknowledge the reviewers of early drafts of the interview and express our gratitude for their contributions of expertise. These reviewers include Thomas Carpenter, Victoria Jacobs, Ian Whitacre, Walter Secada, Juli Dixon, Kristopher Childs, and Amanda Tazaz.

Juli Dixon, Kristopher Childs, and Amanda Tazaz made noteworthy contributions to the design and implementation of interviewer training.

The interview team included Kristopher Childs, Amanda Tazaz, Nesrin Sahin, Vernita Glenn-White, Rebecca Gault, Laura Tapp, Pooja Vaswani, Harlan Thrailkill, Ian Whitacre, Wendy Bray, Karon Kerr, Katie Harshman, Gillian Trombley, Erika Moore, Edward Knote, and Luisa Placido. Special thanks are in order for the contributions of Kristopher Childs and Amanda Tazaz for managing the day-to-day activities during those intense weeks of interviews.

Video coding of the interviews was conducted by three of the authors of this report (Robert Schoen, Zachary Champagne, and Kristy Farina) as well as Ian Whitacre and Nesrin Sahin.

Anne Thistle provided valuable assistance with manuscript preparation.

We are especially grateful to the Institute of Education Sciences at the U.S. Department of Education for their support and the students, parents, principals, district leaders, and teachers who agreed to participate in the study and contribute to advancing knowledge in mathematics education. Without them, this work is not possible.

Table of contents

List of appendices

List of tables (Available tables limited to maintain test security.)

List of Figures

Executive Summary

This report provides an overview of the development, implementation, and psychometric properties of a student mathematics interview designed to assess first- and second-grade student achievement and thinking processes. The student interview was conducted with 622 first- or second-grade students in 22 schools located in two public school districts in a single state in the southeastern U.S. during spring 2014. Focused on the domain of number, operations, and equality, the student interview was designed (a) to measure student achievement in mathematics and (b) to gather information about the strategies students use to solve the mathematics problems. Because the interview was designed for both of these purposes, we call it the Mathematics Performance and Achievement (MPAC) interview.

The MPAC interview consists of a series of mathematics problems that the students are asked to solve. It is similar to a mathematics test, except that the interviewer poses the problems and has the opportunity to observe how the student solves the problems and to ask the students to report the strategies they used. The MPAC interview uses a semistructured format. The sequence and wording of the general instructions and the mathematics problems are designed to be presented in the same order and spoken exactly from the interviewer's script. Subsequent follow-up questions varied and depended upon the interviewer's ability to perceive and understand the student's strategy as well as the student's ability to demonstrate or articulate how he or she arrived at the given answer.

Our primary motivation in writing the current report is to create a reference document that detailed the development/validation process that we undertook and archive the results of that work for our own reference. The work was so complex, we wanted to create a document that we could use to remind ourselves what happened and what we learned from the experience. A secondary purpose is to provide transparency to our research, so that scrutiny could be duly applied by the research community and allow the opportunity for critical feedback to be provided by peers and colleagues. We hope there is a tertiary benefit to those undergoing similar investigations so their work may benefit from the findings and lessons we learned through the worked reported in this document.

Content and Construct Validity

The development process for the interview protocol consisted of several phases. A review of extant research literature influenced the first draft. The first draft and target blueprint were reviewed by internal members of the evaluation team and experts in mathematics and student cognition serving on the project advisory board. On the basis of feedback from these experts, the items were revised and recombined through several rounds of revision before a completed draft of the protocol was pilot tested in February and March 2014 with 34 first- or second-grade students from schools not included in the analytic sample of 622 students for the efficacy study. These pilot tests resulted in additional edits to the set of items, the verbal script for the interview, the instructions for pacing of the interview, and the data-recording system.

As an important step in refining the MPAC interview protocol, the 34 pilot interviews informed final development of the mathematics problems on the interview, the interview implementation protocol, and the data-coding scheme. The team of interviewers participated in the pilot interviews to gain familiarity with the protocol and to practice interviewing students as one part of the multifaceted strategy to train interviewers and maximize consistency among them.

Each interview was video recorded by a webcam attached to a laptop computer. The interviewer captured notes on paper during the interview and then entered these data through a Microsoft SharePoint team site using InfoPath software. The video recordings of a stratified random sample of 79 interviews were also coded by a separate reviewer as a check for consistency among interviewers of the implementation of the protocol and coding of data. The overall rate of interrater agreement for whether students provided correct or incorrect answers was .96.

Factorial Validity

For each grade level, MPAC student interview items were screened for outlier parameter estimates. Items determined to have extreme or inadmissible parameter estimates were removed from subsequent analyses. Remaining items were fitted to a correlated-trait item factor analysis (IFA), specified in accordance with an *a priori* five-factor blueprint. Items that did not demonstrate adequate item salience were also dropped. Both empirical and theoretical considerations were applied, maximizing optimal psychometrics for the scales while retaining the intended content validity for the interviews.

The resulting final set of items for each grade were fitted to an IFA model with a higher-order factor structure, wherein the five first-order factors were regressed onto a single second-order factor. The higher-order factor score is intended to be used as the overall achievement score on the interview. The root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis Index (TLI) goodness-of-fit statistics indicated that the models provided a close fit to the data. The Grade 1 higher-order model-fit statistics were $\chi^2(204) = 281.69$, $p < .001$; RMSEA = .03, 90% CI [.02, .04]; CFI = .98; and TLI = .98. The Grade 2 higher-order model fit statistics were $\chi^2(225) = 301.75$, $p < .001$; RMSEA = .04, 90% CI [.02, .04]; CFI = .98; and TLI = .98.

Reliability

The reliabilities of the final MPAC scales were determined from a composite reliability estimate for the higher-order factor and ordinal forms of Cronbach's alpha (α) for the subscales. The Grade 1 higher-order factor composite reliability was .92; that for Grade 2 was also .92. Grade 1 α subscale estimates all exceeded the conventional minimum value of .7 and four exceeded the target value of.8 (range .78 to .90). On the Grade 2 interview, the α estimate for one subscale was below the .7 conventional minimum (.64); the other four subscales met or exceeded the target value of .80 (range .80 to .91). The full research report presents diagnostic and supplementary analyses of scale reliability, including ordinal forms of Revelle's beta (β) and McDonald's omega hierarchical (ω_h) coefficients and IRT information-based reliability estimates.

Concurrent Validity

We examined the concurrent validity of the Grade 1 and Grade 2 interviews by correlating the MPAC factor scores with scores generated from the Discovery Education Assessment (DEA; 2010) and Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008). The correlations between the MPAC higher-order factor score and the DEA Total were .72 for Grade 1 and .69 for Grade 2. The correlations between the MPAC higher-order factor score in the two different grade levels and the ITBS Mathematics Problems and Mathematics Computation tests were .74 and .65 at Grade 1 for each ITBS test, respectively; and .79 and .63 at Grade 2 for each ITBS test, respectively. All correlations between the MPAC higher-order factor score and the DEA Total score and ITBS tests were statistically significant with *p*-values less than .001. MPAC subscale correlations with DEA Total ranged from .60 to .71 in Grade 1 and from.46 to.71 in Grade 2. MPAC subscale correlations with ITBS ranged from.53 to.75 in Grade 1 and from .55 to .78 in Grade 2.

Vertical Scaling

The large number of items that are common to both the Grade 1 and Grade 2 MPAC forms allows for the vertical scaling the two forms, opening the possibility for analyses that pool across grade level. The execution of the measurement invariance analyses and subsequent vertical scaling of the Grade 1 and Grade 2 MPAC forms is not covered in this technical report but will be reported on in a forthcoming addendum.

Summary

The MPAC interview

- Measures the mathematical thinking and achievement of first- and second-grade students
- Focuses on the domain of number, operations, and equality
- Was conducted with a diverse sample of 622 students in spring 2014 in 22 schools located in two school districts that were implementing a standards-based curriculum very similar to the Common Core State Standards for Mathematics, and
- Has strong psychometric properties and meets standards for educational and psychological measurement

The development process for this interview involved expert review that verified the alignment of the content of the interview with current research and with fundamentally important ideas in first- and second-grade mathematics that are consistent with the content of the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Analysis of interviewer coding agreement indicates high coding reliability and adherence to the interview protocol. The high reliability and close model fit are probably the result of the iterative process of development and feedback from a variety of experts, pilot testing with students, and extensive training of interviewers.

Factor analytic models involving five lower-order factors and a single higher-order factor indicate good model fit and sufficiently high reliability across the typical distribution of person ability for first- and second-grade students. The interview results are highly correlated with other instruments currently in use by school districts that have been judged as valid for use to measure student achievement in first and second grades.

1. Introduction and Overview

The dual purpose of the MPAC interview is to measure student achievement in the domain of number, operations, and equality and to gather information on the strategies students use in the process of solving problems in this domain. We therefore developed a semistructured interview protocol wherein the interviewers follow an initial script to introduce each problem and then to improvise with follow-up questions appropriate to the individual student's strategy choice and explanation. These follow up questions focus on gathering information about how students arrive at their answers. The MPAC interview is carefully designed to avoid asking students to prove their answers, solve the problem in more than one way, or justify the use of a particular strategy.

The MPAC interview consists of 29 items in Grade 1 and 30 items in Grade 2. These items are grouped into three categories for the implementation of the interview: Counting, Word Problems, and Equations and Calculations.[1] Table 1 provides a blueprint of the categories and numbers of items asked of Grade 1 and Grade 2 students.

Table 1. Blueprint for the Grade 1 and Grade 2 MPAC Student Interviews Used Spring 2014

	Number of items	
Section	Grade 1	Grade 2
Counting	6	6
Word Problems	7	8
Equations and Calculations	16	16
Total	29	30

Approximately 80% of the questions on the Grade 1 and Grade 2 interviews were identical. When they were not, the questions in the Grade 2 interview were similar in nature but involved higher numbers, which both increase the difficulty proportionally with age and help to access information about how these older students are making sense of operations on multidigit whole numbers. The questions that are identical are presented in the same order in the two grades.

Interviewers were instructed to explain to students at the beginning of the interview that they were conducting the interview because they are interested in how students solve math problems. After a student solved each word problem, the interviewer asked, “How did you get that answer?” The interviewer could make modifications to the exact wording, such as asking “How did you get 43?” or “I think I see what you did, but can you explain to me how you were using the cubes to find out your answer?”

The purpose of the interviewer's follow-up question was not to find out whether students could prove their answers. Rather, it was to make the thinking process students actually used more salient. When

[1] Although these categories were used for the purpose of conducting the interview, note that these were not the categories for the psychometric model used to analyze the data. See the Data Analysis and Results sections for information about the facets of knowledge used for the purpose of data analysis and reporting of achievement outcomes.

the student's response was something like "I did it in my head," the interviewer asked a probing follow-up question such as, "Can you tell me what you did in your head?" When the strategy is readily apparent, and the interviewer has very high confidence in how the student solved the problem, the interviewer might simply say "I see just how you got that answer" and proceed to the next problem.

The interviewers were instructed specifically not to ask students to prove their answers or to show how they might solve it in a different way. For example, as a subtle but important variant of the standard follow-up question, the interviewers did *not* ask questions such as, "How do you know that is the answer?," "Why did you solve it that way?," or "Why did you use cubes to solve this problem?"

Sometimes a student's explanation of how the problem was solved and what the interviewer observed the student to do appeared to be inconsistent. Unless the interviewer had indisputable, positive evidence to the contrary, the way the student explained arriving at the answer was accepted as accurate, even when the interviewer retained some doubt whether that was exactly how the answer was generated. In attempt to minimize the instances of revisionist explanations, the tempo of the interview was kept fairly rapid. (But the fast tempo did not apply to the period between posing of the problem and the student's providing the final answer.)

Students sometimes changed their answers while explaining how they arrived at their answers. Ultimately, the student's final answer was accepted and recorded as the official response. To avoid introducing bias, interviewers were advised to take caution to respond in the same way—in words, facial expressions, and voice inflection—regardless of whether the student generated a correct answer. A more complete list of the instructions for interviewing is presented in Appendix A.

In general, the problems within the Word Problems section of the interview were ordered from easier to more difficult, where the difficulty was largely determined by the problem type and the numbers involved in the problem. When a student was unsuccessful at correctly solving three consecutive problems in this section, the interviewer had the option to terminate it and to move to the Equations and Calculations section of the interview. This *mercy rule* is based on the assumption that the student would not correctly solve the later problems after several failed attempts at easier problems. The rule was an attempt to avoid causing undue stress to children who were not performing well (and knew it). Interviewers were instructed to use their own clinical judgment to decide when to terminate a section or an item and to move on to the next to avoid causing undue stress. In addition, interviews lasting more than one hour were politely terminated after the student finished the current problem.

1.1. Section 0: Introductions and Question about Student Attitudes

The interviewer began the interview by introducing him or herself and verified the name and grade level of the student (through cordial introductions). The interviewer explained that the focus of the interview was on *how* students solve mathematics problems. The interviewer asked the student's assent to be interviewed and to be video recorded. The student's assent was recorded on the metadata sheet. If the student did not assent, the interview was politely terminated without prejudice.

1.2. Section 1: Counting

This section of the interview was intended to gather information about student abilities in several key aspects of verbal counting: counting forward and counting backward (by ones), counting when starting at a number other than one, crossing decade numbers while counting (including one hundred), and counting forward and counting backward by tens.

These items were intended to be easy for most students and served several purposes. Some of the more difficult items were used to measure knowledge. In general, the items were designed to be fairly easy for students, and the section was placed first in the interview to build students' comfort level by allowing them to solve a few tasks successfully.

One additional item was used only with students identified by the teacher or the school as English Language Learners or as having limited English proficiency. This counting item was used as a screening tool to determine whether the student was sufficiently comfortable engaging in the assessment in English and, in turn, whether interviewing the student in English was appropriate.

Students did not have access to tools in the Counting section (other than their mind and their fingers). Four types of tools (paper, markers, snap cubes, and base-ten blocks) were presented to the student at the end of the Counting section.

Table 2 (available in the full report) shows the six corresponding items in each grade level in the Counting section.

1.3. Section 2: Word Problems

This section contained problems representing a range of difficulty and consisting of two subtypes: (1) standard addition and subtraction and (2) standard multiplication and division (grouping-type problems). The more difficult problems were presented later in the section. Table 3 (available in the full report) provides a list of the sequence of word problems by showing the type of problem and the numbers presented in the problem for the sake of brief comparison.

Interviewers were instructed to be mindful of the time elapsed during the interview. If a student had not completed the word-problem section with 35 minutes of the start of the interview, the interviewer allowed the student to finish the current problem and then proceeded to the Equations and Calculations section.

1.4. Section 3: Equations and Calculations

This section contains questions about students' understanding of equations and their ability to perform calculations involving addition or subtraction. Three types of problems are included in this section: computation, true/false questions about equations, and solving equations for an unknown quantity. Table 4 in the full report (redacted from this report) shows the sequence of problems in this section. Note that the items in this section were not modeled as part of the same lower-order factor; they were grouped in a single section for the purpose of implementation of the interview. More information about the modeling can be found in the Data Analysis and Results sections of this report.

2. Procedures

2.1. Instrument Development

The development process for the student interview protocol consisted of several phases:

1. Review of literature and evaluation of the goals of the Cognitively Guided Instruction program
2. Development of first written draft of the interview items and protocol
3. Review of draft protocol by internal members of the evaluation team and several members of the project advisory board
4. Revision of protocol based on feedback
5. Pilot testing of interview protocol and training of interviewers
6. Revision of protocol and development of electronic data-entry system

Because the interview was used in spring 2014 for the purpose of evaluating the impact of a teacher professional-development program based around a program related to Cognitively Guided Instruction (CGI), the corpus of literature related specifically to CGI (e.g., Carpenter et al., 1989; 1999; 2003; Falkner et al. , 1999; Jacobs et al., 2007) was also reviewed. In addition to review and analysis of these published sources, CGI experts on staff and on the project advisory board were consulted about those aspects of student thinking likely to be affected by a teacher's involvement in the program. To avoid overalignment of the interview with the CGI program, we took abundant caution to avoid using problems that were encountered by teachers in the CGI program. In addition, the workshop leaders and coordinators did not have access to the items on the interview.

Conceptual categories were determined on the basis of a review of scholarly literature related to student thinking in the domain of number, operations, and equality. From these sources, the major categories of Counting, Word Problems, and Equations and Calculations were determined to be likely to provide important information about the effect of the CGI program on student thinking.

The original draft protocol was shared with senior project personnel and revised according to internal feedback. A draft interview protocol was written and shared with several advisory board members (including Victoria Jacobs, Ian Whitacre, and Thomas Carpenter). Feedback from these experts resulted in substantive changes to items, including types of problems included, numbers used in the problems, administration instructions, and the number of items in each category.

The content of the interview was designed to align with central topics in number, operations, and equality in the general first- and second-grade curriculum. It was designed to be valid for use as a mathematics achievement measure for use with students in first- and second-grade mathematics classrooms. The topics are consistent with the framework of the Common Core State Standards for Mathematics (NGA & CCSSO, 2010) and with the standards in the accountability system in place in the schools where the field study was conducted.

2.2. Interviewer Training

Gathering data for a semistructured interview in a way that permits a fair comparison from interview to interview requires considerable skill and coordination on the part of the interviewers. Almost all of the personnel involved in interviewing were faculty or graduate students in mathematics education or elementary education. All had some experience teaching mathematics and studying how students learn mathematics.

In accordance with state regulations, a rigorous, formal background check (including fingerprinting and FBI screening) was performed on all prospective interviewers. Fourteen individuals completed the following training procedures and conducted interviews in spring 2014.

2.2.1. Phase 1 of Interviewer Training

The first phase involved a classroom-style orientation and introduction to the interview and related research on student thinking, as well as a discussion and guidelines for how the interviewers were expected to behave in schools.

During the first two four-hour training sessions, the prospective interviewers discussed typologies for word problems, classes and definitions of archetypical strategies students use to solve for single- and multidigit numbers. The training also included an introduction and examples of relational thinking with respect to the equal sign. The training included a discussion of general principles concerning interviewing children, including guidelines for behaviors. Several ideas from the chapter titled *Guidelines for Clinical Interviews* from the book *Entering the Child's Mind: The Clinical Interview in Psychological Research and Practice* (Ginsburg, 1997) were used to frame the discussion. The prospective interviewers viewed videos of students in an interview setting and discussed the strategies that students used in the video recorded interviews. Each interviewer received a copy of *Children's Mathematics: Cognitively Guided Instruction* (Carpenter et al., 1999) and was assigned to read chapters concerning how students solve addition, subtraction, multiplication, and division problems involving single- and multi-digit numbers.

2.2.2. Phase 2 of Interviewer Training

The second phase of interviewer training involved an iterative process of piloting the interview with students and then discussing and reflecting on the purpose of the interview, interviewer techniques, student thinking, the interview protocol, and the data categories resulting from the interview.

In the first wave of pilot interviews, one of the more experienced interviewers conducted the interviews (with students in three private schools and one charter school) while the less experienced interviewers observed. Subsequent days conducting pilot interviews provided all prospective interviewers with opportunities to practice the role of interviewer. These pilot interviews provided opportunity for the interviewers to practice simultaneously conducting the interview, recording data, and using the video recording devices. Phase 2 of the training provided opportunities for the interviewers to reflect and discuss the protocol with the goal of attaining high internal consistency in implementation and a common understanding of the goals and procedures. It also provided opportunities to relieve some of the anxiety the interviewers were feeling about conducting interviews before the real data were collected.

2.2.3. Phase 3 of Interviewer Training

The third phase of training occurred during the first two weeks of real data collection. During this period, interviews were conducted in pairs by an interviewer and an observer. Both of these individuals were trained members of the interview team. The interviewers conducted the interviews while the observers sat next to them and observed the interview (and interviewee). Both members of the pair recorded data according to the standard protocol, and they compared and discussed their notes and recollections with respect to adherence to the protocol as well as the coding of the data they recorded. The video recordings of a stratified sample of these first interviews were coded by the project principal investigator. The data he coded for the interview as well as a written analysis of adherence to the interview protocol was sent to each of the interviewers during this period for them to compare and consider.

The purpose of this third phase was to provide adequate learning opportunities to continue to strive toward high consistency in implementation of the protocol and also to provide an opportunity for the less experienced interviewers to gain more practice and comfort before working on their own. These occasional checks for consistency continued throughout the data-collection period as a guard against drifting procedures for implementation of the interview or coding the student strategies.

After interviews were conducted, the video recordings of the interviews were also coded from June through September 2014. A random sample of the interview videos was selected and coded by trained interviewers. Video coding procedures were identical to those used by the interviewers with one exception. The video coders had the option to code items as *interviewer bias*, which indicated that the interview strayed from the protocol in a way that invalidated the item. Percent agreement between video coders and interviewers was calculated, and those results and the rate of incidence of items flagged as interviewer bias are available in the Results section of this report.

Table 5 provides an overview of the training period and the major activities during that period.

Table 5. Schedule, Duration, and Type of Activity in the Interviewer Training Period

Date	Duration	Activity
Feb 27	4 hours	Introduction to problem types, strategies, interviewing guidelines
Feb 28	4 hours	Introduction to problem types, strategies, interviewing guidelines
Feb 29–Mar 10	3 hours	Reading assigned chapters in resource books
Mar 10	6 hours	Practice interviews and debrief/reflection session
Mar 11	6 hours	Practice interviews and debrief/reflection session
Mar 17	4 hours	Practice interviews and debrief/reflection session
Mar 24	4 hours	Practice interviews and debrief/reflection session
Apr 1–Apr 11	6 hours/day	Paired interviews with interviewer and observer collecting real data and debriefing after each interview

2.2.4. Digression From Protocol

The expectation of each interviewer was to adhere to the script and interviewer guidelines (i.e., Appendix A in full report) at all times. The video coders were instructed to flag items when interviewers digressed from the script dramatically enough that the digressions affected the student's response, either positively or negatively; we coded those digressions as *interviewer bias*. These digressions were infrequent, but they did occur, and the resulting data were recoded as missing. Below are two examples of the more common digressions from the protocol:

1. For the True or False questions in the Equations and Calculations section, EC10–EC 13 (see Appendices B and C), if students read the equation in a manner that was not exactly as it was written, and the interviewer did not prompt the student to reread the equation, we

considered this a digression from the protocol. For example, if the student read the equation *a* = *b* + *c* as *c* + *b* = *a*, and the interviewer did not prompt the student to reread it as it was written, we coded the item as digression from protocol.

2. We considered instances when an interviewer read the incorrect number on the Word Problem items as digressions from the protocol that clearly affected he student's final response and did not include the item.

Out of the 281 video-coded interviews, including approximately 30 items per interview, a total of 65 items were coded for digressions from protocol, an incidence rate of 8 items per 1,000 items. Out of 622 interviews, 44 were known to have been affected by digressions, and the interviews that were affected contained between one and four instances.

2.3. Coding Scheme

The interview was designed to be coded in real time by the interviewer. Strategies that students use to solve problems can be sorted into two broad categories: invented and instructed.[2] In either case, particular attention was given to recording information about strategies and behaviors that might be used to infer student understanding of place value ideas, properties of operations and equality, number fact recall, and relational thinking.

The full interview was pilot tested with 34 students who did not attend schools included in the analytic sample for the efficacy study. These pilot tests resulted in several rounds of incremental edits to the set of items, the verbal script for the interview, the instructions for pacing of the interview, and the data recording system. The details in the data recording and coding system were also further refined during this pilot testing with input from the interviewers.

Data categories included the answer given as well as descriptive codes for the observed strategies, which included named strategies such as join all, separate from, incrementing, compensation, standard algorithm, etc. A more detailed description of each strategy and its substrategies is given in the following section. Although the body of literature surrounding many of these types of strategies defines the strategies as resulting in correct solutions, we encountered many students attempting to use these strategies in the pilot testing phase of the development and generating incorrect answers. As a result, strategies were coded on the basis of the strategy used by the student regardless of whether the answer was correct.

For the Counting items on the interview, we collected data on:

- The answer the student provided
- Whether the student used correct verbal counting

[2] The term "invented" is used here on the basis of decades-long history of use in scholarly literature. The term was coined during a time when these particular strategies were not commonly known by teachers or included in textbooks. Over the past few decades, these strategies have percolated into textbooks and are becoming part of the teaching lexicon, and the boundary between invented and instructed strategies may no longer be clear. On the data coding sheet, the term ad hoc was used in place of invented as the category to describe numerically specific strategies used by students in the interview.

- Whether the student counted by ones or used place value to determine the "what number is ___ less than ___?" items (3–5)

For the items in the Word Problems and Equations and Calculations sections, we collected data on:

- The answer the student provided
- The major strategy used by the student (i.e., Objects Representing All Quantities in the Sets and Subsets , Counting, Ad Hoc, Recalled Fact, Standard Algorithm, Other)
- Selected substrategies by item (where applicable)
- Any physical tools used by the student (when applicable)
- Whether an additive or subtractive strategy was used (where applicable)

2.4. Strategy Type Descriptions

Objects Representing All Quantities in the Sets and Subsets (ORQSS)—We used the ORQSS code when the students used manipulatives or drawings to model all quantities within the problem. Our definition of an ORQSS strategy aligns closely with the definition of direct modeling (Carpenter et al., 1999) with one exception. If a student's model physically represented each quantity in the problem (including the set and subsets), we classified that strategy as an ORQSS strategy and then record the action that we observed. The ORQSS code does not require the student's construction of a model that directly parallels the action occurring in the story problem. For example, if a student used manipulatives to solve a *Join Change Unknown* problem and used them in a manner consistent with a *Separate from* strategy, we coded that strategy under the major strategy of ORQSS, and we coded *Separate from* as the substrategy.

When the student used an ORQSS-type strategy, we used the following names of substrategies when applicable to specific problems:

- Join/Count All
- Join/Add To
- Separate/Take From
- Separate To
- Matching
- Trial and Error
- Grouping
- Measurement
- Partitive
- Other (explain)

The descriptions and classifications for these strategies and substrategies were informed by the definitions provided by Carpenter et al. (1999). Additional information on how the student counted the set representing the answer was also recorded.

Counting—We used the Counting code when the student employed a strategy in which at least one of the quantities in the problem was not represented physically. For these items, we coded the direction of the count (forward or backward), the number name that began the count, the number name that ended the count, and how the student counted (e.g., by ones, twos, tens and ones).

Recalled Fact—When the student stated that the answer to the problem was recalled from memory, we code it Recalled Fact. These could include the fact presented or use an application of the commutative property. In addition, we coded for those students who recalled an addition fact to solve a subtraction problem, such as using the knowledge that $a + b = c$ to solve $c – a = b$.

Derived Fact—When the student stated that the answer was derived from another known fact, we code these as a Derived Fact. Derived facts were used when the student combined known quantities when a specific fact was not known at a recall level. An example would be a case where student first decomposed one of the addends to determine a sum of ten and then added the remaining amount to the intermediate sum.

Ad Hoc–When the student employed a numerically-specific strategy, we classified it as Ad Hoc. We deliberately avoided the term "invented" here because many strategies historically called "invented" are now being taught, together with their names. This is certainly true of the textbook series in use in the schools in our analytic sample.

Within the Ad Hoc strategy, we coded (where applicable) use by students of an incrementing, compensation, or combining-tens-and-ones (Carpenter et al., 1999) substrategy. We also observed and coded for place value and repeated addition or subtraction substrategies. Some items included a finer level of detail in the coding scheme than others. See Appendices B and C for the interview protocols with the coding schemes for each item. In general, Ad Hoc strategies were consistent with numerically specific strategies (for a discussion of these types of strategies, see Smith, 1995).

Standard Algorithm—When students used the standard United States algorithm for addition or subtraction, we coded for the following items:

- The student's final response
- Whether the student used counting or fact recall to determine the values in individual places
- When the student used an incorrect variation of the algorithm, the following so-called buggy algorithm applications
  - Subtracted "up"
  - Wrote 2-digit partial sum without regrouping
  - Regrouped, did not add regrouped ten
  - Regrouped across zero—skipped zero place
  - "Borrowed" from zero as if ten
  - Considered zero minus to be zero
  - "Borrowed" without subtracting adjacent ten

For the True or False items on the interview, we coded for the following:

- The student's response (True or False)
- How the student decided whether the equation was true or false (common responses were included for each item and are presented in Appendices B and C)

3. Data Analysis

3.1. Description of the Sample

The sample was composed of 2,631 students (1,442 Grade 1 and 1,414 Grade 2) for whom signed parental consent was obtained. The student sample comes from 22 schools in two diverse public school districts (7 schools in one district; 15 in the other) in a single state located in the southeastern United States. First- and second-grade teachers in these schools were participating in a large-scale, cluster-randomized controlled trial evaluating the efficacy of a teacher professional development program in mathematics. Half of the schools in the sample were assigned at random to the treatment condition; the other half to the control condition.

Students in the sample completed three measurement instruments as part of their participation in the study: a whole-group-administered, written pretest at the beginning of the 2013–2014 school year; the Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008), also administered in a whole-class setting at the end of the 2013–2014 school year; and a student interview, which was administered in an individual, one-on-one setting at the end of the 2013–2014 school year. That interview serves as the primary subject of the current research report.

In addition to those three instruments completed as part of the research study, students in one of the participating districts (comprising 7 schools) completed the Discovery Education Assessment (DEA) Common Core Edition (DEA, 2010) during the 2013–2014 school year. Results from the administration of the DEA were provided by the school district and are herein used as part of our concurrent validity analyses.

Table 6 reports the sample sizes for each of the four measurement instruments. Table 7 reports the demographics for the sample of participating students.

Table 6. Student Sample Size per Measurement Instrument

	Sample size		
Measure	Grade 1	Grade 2	Total
Pretest	1,226	1,147	2,373
Iowa Test of Basic Skills	1,103	1,069	2,172
Discovery Education Assessment	391	269	660
MPAC interview	336	286	622

Table 7. Student Sample Demographics

Characteristic	Total student sample (*N* = 2,631)		Eligible sample (*n* = 2,279)		MPAC interview sample (*n = 622)*	
	Proportion	*n*	Proportion	*n*	Proportion	*n*
Gender						
Male	.48	1,254	.48	1,083	.48	301
Female	.47	1,242	.47	1,075	.51	318
Missing	.05	135	.05	121	.01	3
Grade						
1	.50	1,326	.51	1,153	.54	336
2	.50	1,305	.49	1,126	.46	286
Race/Ethnicity						
Asian	.04	115	.05	102	.05	33
Black	.17	459	.17	396	.20	122
White	.35	912	.35	791	.33	207
Other	.03	70	.03	59	.03	19
Hispanic	.35	910	.35	787	.38	238
Missing	.06	165	.06	144	.01	3
English Language Learners	.21	553	.21	471	.23	140
Eligible for Free or Reduced-Price Lunch	.58	1523	.57	1,303	.61	381
Exceptionality						
Students With Disabilities	.07	184	.07	159	.06	40
Gifted	.04	97	.04	90	.05	31
Missing	.06	165	.06	144	.01	3

Note. Proportion provided reflects percentage of each sample. Some characteristic categories are not mutually exclusive. Students with unreported demographic information are represented in the "Unknown" category. The Asian, Black, and White categories are non-Hispanic. Eligible sample refers to students in the sample with positive consent for video recording.

3.2. Sampling Procedure

The interviews were conducted with students who completed pretests at the start of the school year. To allow for later review of the students' responses, we interviewed only students with positive parental consent for video recording.

Interviews were conducted with a stratified random sample of up to four students from each participating teacher's classroom. To maintain a balanced sample within each classroom with respect to student gender, we used gender as the first stratum. Student gender data were provided by the school districts. The goal was to include two boys and two girls in the interview sample from each teacher's class. The second stratum involved splitting the class by pretest achievement level. The median achievement level for each classroom was determined, and a student of each gender was drawn from the lower half of the class (including the median) and from the upper half of the classroom.

Class rosters were divided into four subcategories: upper pretest boy, lower pretest boy, upper pretest girl, lower pretest girl. A random number was assigned to each student, and the sample was sorted by gender, pretest stratum, and random number. Then, a primary and an alternate student were selected from each stratum on the basis of the random number. The highest random number designated the primary student; the second highest the alternate. Alternate students were only called upon to be interviewed in instances where the primary student was absent or did not assent to be interviewed. Although in nearly all classes all four strata were represented, some classrooms did not have an alternate student for every stratum or even a primary for every stratum.

The interviewers were not made aware of the treatment condition of the school (or students), and they were also not aware of whether the student was from the upper or lower half of the class.

3.3. Student Interview Interrater Agreement

We also conducted an investigation as to the interrater agreement. This section describes that process and the results.

Of the 622 valid student interviews that were conducted, an initial group of 281 interviews were coded by a trained interviewer from the video recording. The data for 171 of these video-coded interviews were entered only by the video coder (not by the interviewer), so that the interviewers could focus their time and attention on following the interview protocol in the first weeks of data collection. The remaining 110 video-coded interviews were coded by both the interviewer and the video coder. Of these, we first drew a random sample of 79 interviews and used them to investigate interrater agreement. After that random sample, a set of 31 additional interviews were identified to be video coded.

Of the 79 interviews selected at random from the 451 sets of interview data entered by the interviewers for comparison between video coder and interviewer, 21 were video-coded by two different people, so that we could also compare agreement among video coders.

We calculated interrater agreement by dividing the total number of matching values by the total number of instances for each data type (e.g., correct, strategy, additive/subtractive). Interrater agreement for individual coders was examined, and an additional 31 interviews were video coded to replace data from interviewers identified to have below-average agreement on the basis of the stratified random sample used to check for interrater agreement. To improve the overall accuracy of the dataset, data obtained from video coding replaced interviewer data for all 110 cases where two sets of data existed.

Exact agreement between video coders across all codes was 92%, which is 3% higher than the overall interviewer-video coder agreement. Video coders could improve their accuracy through advantages not available to interviewers, including the ability to pause, rewind, and rewatch segments of an interview. Video coders were also able to refer to literature during coding to ensure the strategies observed were recorded correctly. As a result, the video-coded data appear slightly more reliable than the real-time, interviewer-coded data. In all 110 cases where an interview was coded, the video-coded data therefore replaced the interviewer-coded data. Tables 8 and 9 report the interrater agreement on individual items or groups of items.

Table 8. Interrater Agreement by Data Type

	Type of comparison	
Type of agreement	Video–Interviewer (*n* = 79)	Video–Video (*n* = 21)
Correct/incorrect	.96	.98
Major strategy	.83	.89
Additive or subtractive	.78	.73
Total strategy	.89	.92

The interrater agreement proportions reflected here represent agreement between video-coded data and interviewer-coded data. The achievement-score data depend only on the Correct/Incorrect evaluation, which had an interrater agreement of greater than 95%. Because data from coders with low interrater agreement were replaced by video-coded data, the proportions of interrater agreement reported in Tables 8 and 9 constitute a conservative estimate of the accuracy of the final student interview data.

Table 9. Grade 1 and Grade 2 Video Coder–to–Interviewer Interrater Agreement by Data Type, Split by Item

Item	Description	Correctness	Major strategy	Additive/ subtractive
Counting				
CNS 0	removed for test security	.95		
CNS 1_Gr1	removed for test security	.97		
CNS 1_Gr2	removed for test security	.95		
CNS 2	removed for test security	1.00		
CNS 3	removed for test security	.99	.89	
CNS 4	removed for test security	.98	.81	
CNS 5	removed for test security	.98	.84	
Word Problems				
WP 6	removed for test security	.98	.85	
WP 7	removed for test security	.99	.87	.91
WP 8	removed for test security	.98	.73	
WP 9_Gr1	removed for test security	.95	.92	
WP 9_Gr2	removed for test security	.95	.73	
WP 10	removed for test security	.98	.86	.79
WP 11	removed for test security	.96	.65	
WP 12	removed for test security	.91	.84	
WP 13_Gr2	removed for test security	.98	.73	.68
Equations and Calculations				
EC 1	removed for test security	.99	.89	
EC 2	removed for test security	.99	.94	.85
EC 3	removed for test security	1.00	.89	
EC 4_Gr1	removed for test security	.92	.89	1.00
EC 4_Gr2	removed for test security	1.00	.73	.73
EC 5	removed for test security	.95	.76	
EC 6	removed for test security	.96	.84	.71
EC 7_Gr1	removed for test security	.97	.82	.71
EC 7_Gr2	removed for test security	1.00	.76	.78
EC 8	removed for test security	1.00		
EC 9	removed for test security	.96		
EC 10	removed for test security	.97	.77	
EC 11	removed for test security	.98	.87	
EC 12	removed for test security	1.00	.82	
EC 13	removed for test security	1.00	.79	
EC 14	removed for test security	.96	.89	.77
EC 15	removed for test security	.99		
EC 16	removed for test security	.96		

Note. N = 79. Grade 1 n = 38. Grade 2 n = 41. Items with "_Gr1" in the label are unique to the Grade 1 interview; those with "_Gr2" unique to the Grade 2 interview. These percentages reflect agreement on all codes recorded, including codes for skipped items. Major strategy and Additive/subtractive data are only available for some items.

3.4. Investigation of the Factorial Validity and Scale Reliability

All analyses were performed with Mplus version 7.11 (Muthén & Muthén, 1998-2012), with the exception of the estimation of Cronbach's alpha (α), Revelle's beta (β), and McDonald's omega heirarchical (ω_h) reliability coefficients, which were performed in R 3.1.2 (R Development Core Team, 2014) with the psych package (Revelle, 2016) alpha, splithalf, omega, and polychoric functions.

Our investigation included five steps. We intended (1) to screen-out items that demonstrated outlier parameter estimates when fit to a unidimensional framework, (2) to evaluate item performance when structured in accordance with the five-factor blueprint and drop items that demonstrated low-salience with their respective factor, (3) to respecify the structure of the model from one of correlated factors to one of a single second-order factor and five first-order factors, (4) to estimate reliabilities for the interview overall and for each subscale, and (5) to estimate the concurrent validity of the MPAC interview for each grade level.

The first step was to screen the initial set of items within a 2-parameter logistic (2-pl) unidimensional item response theory (UIRT) framework. Discrimination and difficulty parameters were inspected, and items were flagged for removal if they had outlier parameter estimates or they provided little information in a region along the difficulty continuum where a number of other better discriminating items were present. Criteria of > 3 discrimination or difficulty greater than three or less than negative three were used to indicate outlier estimates, and a criterion of < 0.4 discrimination was used to indicate that it provided little information. Poorly discriminating items that appeared to fill a void along the difficulty continuum were flagged to receive special consideration for being retained.

The second step was to fit the screened data to a correlated trait item factor analysis (IFA; confirmatory factor analysis with ordered-categorical indicators) model that was in accordance with the 5-factor model structure specified by the principal investigator in consultation with project advisory board members.[3]

We used the model chi-square (χ^2), root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI) to evaluate overall model fit. Following guidelines in the structural-equation modeling literature (Browne & Cudeck, 1992; MacCallum et al., 1996), we interpreted RMSEA values of .05, .08, and .10, as thresholds of close, reasonable, and mediocre model fit, respectively, and interpreted values > .10 to indicate poor model fit. Drawing from findings and observations noted in the literature (Bentler & Bonett, 1980; Hu & Bentler, 1999), we interpreted CFI and TLI values of .95 and .90 as thresholds of close and reasonable fit, respectively, and interpreted values < .90 to indicate poor model fit. We note that little is known about the behavior of these indices when based on models fit to categorical data (Nye & Drasgow, 2011), which adds to the chorus of cautions associated with using universal cutoff values to determine model adequacy (e.g., Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004). Because fit indices were not used within any of the decision rules, a cautious application of these threshold interpretations bears on the evaluation of the final models but has no bearing on the process employed in specifying the models.

[3] Note that the final CFA reported in the Results section was a second attempt at categorization. An explanation of the initial and resulting categorization is presented in the Discussion section.

Confirmatory factor analysis models with standardized factor loadings > .7 in absolute value are optimal, as they ensure that at least 50% of the variance in responses is explained by the specified latent trait. In practice, however, this criterion is often difficult to attain while maintaining the content representativeness intended for many scales. Researchers working with applied measurement (e.g., Reise et al., 2011) have used standardized factor loadings as low as .5 in absolute value as a threshold for item salience. In accordance with this practice, we aimed to retain in the final model only items that had standardized factor loading estimates > .5 and unstandardized factor loading *p*-values < .05.

The third step was to respecify the reduced set of items with a higher-order factor structure in which the five first-order factors were regressed onto a single second-order factor. As with the correlated trait model, we evaluated the factorial validity of the higher-order model on the basis of overall goodness of fit and interpretability, size, and statistical significance of the parameter estimates. The purpose of respecifying the factor structure as a higher-order model was (a) to select a more parsimonious factor structure if warranted by goodness of fit to the data and (b) to specify a factor structure that provided the pragmatic benefit and utility of having a single underlying factor (and composite score).

The fourth step was to inspect the scale reliabilities, which we did by calculating the composite reliability for the higher-order total math factor and estimating ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h for the subscales. As a supplementary analysis, we also estimated the reliability for the total math scale, except modeled as a single factor on which the reduced set of items loaded directly. For this purpose, we estimated α, β, and ω_h reliability coefficients for a single, first-order factor. Also, we inspected the total information curve from a 2-pl UIRT model using the reduced set of items modeled as a single, first-order factor. To evaluate reliability coefficients, we applied the conventional values of .7 and .8 as the minimum and target values for scale reliability, respectively (Nunnally & Bernstein, 1994; Streiner, 2003).

Using the equation described in Geldhof et al. (2014), we calculated the composite reliability as the squared sum of unstandardized second-order factor loadings divided by the squared sum of unstandardized second-order factor loadings plus the sum of the first-order factor residual variances. Accordingly, the first-order factors are Number Facts (NF), Operations on Both Sides of the Equal Sign (OBS), Word Problems (WP), Equal Sign as a Relational Symbol (ESRS), and Computation (COMP). The formula for the composite reliability for the second-order Math factor is

$$\text{Composite reliability} = \frac{(\lambda_{\text{NF}} + \lambda_{\text{OBS}} + \lambda_{\text{WP}} + \lambda_{\text{ESRS}} + \lambda_{\text{COMP}})^2}{(\lambda_{\text{NF}} + \lambda_{\text{OBS}} + \lambda_{\text{WP}} + \lambda_{\text{ESRS}} + \lambda_{\text{COMP}})^2 + (\zeta_{\text{NF}} + \zeta_{\text{OBS}} + \zeta_{\text{WP}} + \zeta_{\text{ESRS}} + \zeta_{\text{COMP}})},$$

where *λ* is the unstandardized second-order factor loading and ζ is the residual variance for the respective first-order factor. This calculation is analogous to the classical conceptualization of reliability as the ratio of true-score-variance to the true-score-variance-plus-error-variance.

For our estimation of ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h, we executed the procedure described by Gadermann, Guhn, and Zumbo (2012). Cronbach's α is mathematically equivalent to the mean of all possible split half reliabilities and Revelle's β is the worst split half reliability. Only when essential tau equivalence (i.e., unidimensionality and equality of factor loadings) is achieved will α equal β; otherwise, α will always be greater than β. Variability in factor loadings can be attributable to microstructures (multidimensionality) in the data: what Revelle (1979) termed *lumpiness*. McDonald's ω_h models lumpiness in the data through a bifactor structure. The relation between α and

ω_h is more dynamic than that between α and β, as α can be greater than, equal to, or less than ω_h, as a result of the particular combination of scale dimensionality and factor loading variability. We investigated these scale properties by examining the relation among coefficients α, β, and ω_h through the four-type heuristic proposed by Zinbarg et al., (2005).

The reduced set of items in the final model of the MPAC interviews were fit to a 2-pl UIRT model to generate a total information curve (TIC) for each grade-level interview for the purpose of judging scale reliability across the distribution of person ability. Inspecting the TICs allowed us to make the conversion from information function to reliability along a given range of person abilities with the equation

Reliability = Information/(Information + 1).

Accordingly, information of 2.33 converts to reliability of approximately .7 and information of 4 converts to a reliability of .8, for example. This equation derives from the classical test theory equation of reliability = true variance / (true variance + error variance). Applied to an IRT framework, where error variance = 1 / information, the equation works out to reliability = 1 / 1 + (1 / information), which coverts algebraically to information / (information + 1) (http://www.lesahoffman.com; cf. Embretson & Reise, 2000).

The reliability estimates directly relevant to the scales as described and presented as the final models in this research report are the composite reliability for the higher-order total math factor and the α, β, and ω_h reliability coefficients for the subscales. That is, the α, β, and ω_h reliability coefficients and the 2-pl UIRT information-based reliability estimates for the total math scale apply to structures and modeling approaches different from that of the higher-order structure described in this research report. These supplementary analyses of reliability for the total math scale were conducted as part of our endeavor toward obtaining a broad understanding of how the items from the final model worked together and are presented principally with the purpose of thoroughness and transparency in reporting.

The fifth step was to investigate the concurrent validity of the interviews by correlating their factor scores with scores from the DEA Common Core Edition (Discovery Education Assessment, 2010) and Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008). We used correlations > .7 to indicate scale correspondence. The procedure involved saving the factor scores from the final higher-order factor model for the Grade 1 and Grade 2 interviews. Then, as manifest variables, the factor scores were merged into a file containing the DEA and ITBS scores. For the DEA, five variables were used: the numbers of items answered correctly on (a) the total skill set and the (b) Operations & Algebra, (c) Number/Operations Base Ten, (d) Measurement & Data, and (e) Geometry skill subtests. For the ITBS, we used the Math Problems and Math Computation tests for Level 7 and Level 8 at Grade 1 and Grade 2, respectively.

Because only one of the two participating school districts administered the DEA, analyses using DEA data were applied to a smaller sample, and because student interviews were administered to only about one-fourth of the participating students in sample classrooms, the number of interviewed students with DEA was less than 100 students within each grade level. The Grade 1 sample sizes were Interview n = 336; DEA n = 391; ITBS n = 1149; Interview with DEA correlation n = 95; Interview with ITBS correlation n = 309; and DEA with ITBS correlation n = 321. The Grade 2 sample sizes were Interview n = 286; DEA n = 269; ITBS n = 1104; Interview with DEA correlation n = 78; Interview with ITBS correlation n = 271; and DEA with ITBS correlation n = 244.

4. Results

4.1. Five-factor Test Blueprint

Table 10 provides an overview of the number of items in Grade 1 and Grade 2 that remained after undergoing the full procedure of screening, evaluation, and respecification. Initially, the Grade 1 interview included 29 items and the Grade 2 interview 30 items. The first item on each interview, CNS 0, was designed to be administered to English language learning students only. Item CNS 0 demonstrated limited utility and was dropped before any data modeling. The first two true/false questions about equations were purposefully easy and designed as a way to ease the transition to that section of the interview and provide the interviewees with an opportunity to be exposed to those less common types of questions in the event that they were novel to them. Those two items were dropped from the overall blueprint, but they serve an important purpose for the operationalization of the interview and should be retained when the interview is conducted. The other items were dropped as a result of UIRT and IFA screening and scale-refinement procedures.

Table 10. Number of Items that Remained on the Spring 2014 MPAC Interview Blueprint After Screening

Factor	Grade 1	Grade 2	Common items
Number Facts (NF)	3	3	3
Operations on Both Sides of the Equal Sign (OBS)	3	3	3
Word Problems (WP)	6	7	6
Equal Sign as a Relational Symbol (ESRS)	3	3	3
Computation (COMP)	7	7	5
Total	22	23	20

4.2. Item Screening

Tables 11 and 12 present the full set of items on the Grade 1 and Grade 2 student interviews, respectively. The tables report the proportion answered correctly as well as the 2-pl UIRT discrimination and difficulty parameter estimates for each item on each grade level interview. For ease of reference, we have presented in boldface the entries for items that remained in the final model after undergoing the full procedure of screening, evaluation, and respecification. Also for ease of reference, we have inserted a column that names which factor each item belonged to, according to the item blueprint. Tables 11 and 12 present the items in the order administered and shows them organized according to whether the item structure was that of a counting prompts, word problem, or equations and computation problem.

Although we conceptualized the full set of items to indicate the single construct of student math ability in early elementary number, operations, and equality, we hypothesized a 5-factor substructure. The five factors were Number Facts (NF), Operations on Both Sides of the Equal Sign (OBS), Word Problems (WP), Equal Sign as a Relational Symbol (ESRS), and Computation (COMP).

4.2.1. Grade 1 Interview Item Screening

Table 11 reveals that no items on the Grade 1 interview had outlier discrimination estimates (> 3), but two items (EC 8 and EC 9) were near the outlier minimum and maximum acceptable value (>|3|) for item difficulty. One item (EC 2) fell below the discrimination minimum acceptable value (< 0.4). The high proportions correct observed for EC 8 (.99) and EC 9 (.97) are consistent with marginal outlier estimates of their difficulty parameters. As expected from the original design, EC 8 and EC 9 were dropped from the Grade 1 measurement model.

Table 11. Grade 1 MPAC Interview Item Descriptions, Descriptives, and Item Response Theory (2-pl UIRT) Parameters

			Proportion	2-pl UIRT parameters	
Item	Factor	Item description	correct	Discrimination	Difficulty
Counting					
CNS 1	—	removed for test security	.90	0.63	-2.44
CNS 2	**COMP**	**removed for test security**	**.55**	**1.47**	**-0.14**
CNS 3	**COMP**	**removed for test security**	**.34**	**1.55**	**0.49**
CNS 4	**COMP**	**removed for test security**	**.42**	**1.59**	**0.24**
CNS 5	**COMP**	**removed for test security**	**.34**	**1.91**	**0.45**
Word problems					
WP 6	—	removed for test security	.84	0.70	-1.75
WP 7	**WP**	**removed for test security**	**.34**	**1.31**	**0.51**
WP 8	**WP**	**removed for test security**	**.27**	**1.19**	**0.81**
WP 9	**WP**	**removed for test security**	**.46**	**0.85**	**0.16**
WP 10	**WP**	**removed for test security**	**.39**	**0.87**	**0.42**
WP 11	**WP**	**removed for test security**	**.30**	**1.40**	**0.66**
WP 12	**WP**	**removed for test security**	**.30**	**1.17**	**0.70**
Equations and computation					
EC 1	**NF**	**removed for test security**	**.95**	**0.94**	**-2.47**
EC 2	—	removed for test security	.69	0.39	-1.34
EC 3	**NF**	**removed for test security**	**.92**	**0.85**	**-2.19**
EC 4	**COMP**	**removed for test security**	**.46**	**0.83**	**0.17**
EC 5	—	removed for test security	.49	0.55	0.06
EC 6	**COMP**	**removed for test security**	**.64**	**0.75**	**-0.59**
EC 7	**COMP**	**removed for test security**	**.15**	**0.66**	**1.88**
EC 8	—	removed for test security	.99	0.82	-3.61
EC 9	—	removed for test security	.97	0.56	-4.01
EC 10	**OBS**	**removed for test security**	**.27**	**0.54**	**1.30**
EC 11	**ESRS**	**removed for test security**	**.59**	**0.62**	**-0.40**
EC 12	**ESRS**	**removed for test security**	**.52**	**0.80**	**-0.08**
EC 13	**ESRS**	**removed for test security**	**.48**	**0.64**	**0.11**
EC 14	**NF**	**removed for test security**	**.72**	**1.32**	**-0.73**
EC 15	**OBS**	**removed for test security**	**.15**	**1.00**	**1.45**
EC 16	**OBS**	**removed for test security**	**.28**	**0.99**	**0.83**

Note. N = 336 valid Grade 1 student interviews conducted. Discrimination estimates use a 1.7 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994). Items that remained after factor analysis are presented in boldface type. NF = Number Facts; OBS = Operations on Both Sides of the Equal Sign; WP = Word Problems; ESRS = Equal Sign as a Relational Symbol; COMP = Computation

We plotted the discrimination and difficulty parameters to inform our decision to retain or drop items with special attention to EC 2 because of its low discrimination (0.39). Figure 1, the Grade 1 difficulty-vs.-discrimination scatterplot, does reveal EC 2 to be an item that provides relatively little information, but the item was located in a region on the difficulty continuum where few others were located. Accordingly, EC 2 was given special consideration and was retained and used in the initial correlated trait model for further evaluation.

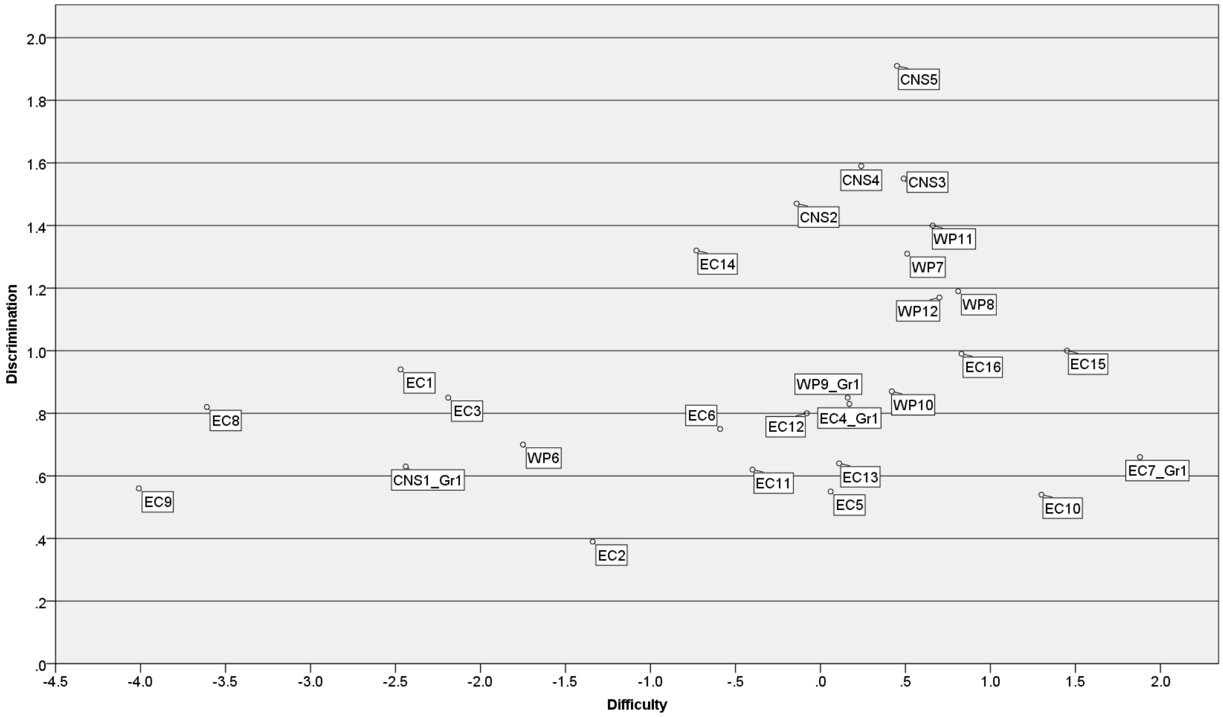


Figure 1. Grade 1 MPAC interview 2-parameter logistic unidimensional item response theory (2-pl UIRT) difficulty-versus-discrimination scatterplot. Items with "_Gr1" in the label are unique to the Grade 1 interview.

4.2.2. Grade 2 Interview Item Screening

Table 12 reveals that one item, EC 8, on the Grade 2 MPAC interview had an outlier discrimination estimate (> 3) and another, EC 9, had an outlier item-difficulty estimate (>|3|). The discrimination estimate for EC 8 was 72, and the difficulty estimate for EC 9 was -8.08. As was also the case with the Grade 1 MPAC interview, high proportions correct were observed for both of these items: > .99 at Grade 2. Also as expected in the original design, we dropped EC 8 and EC 9 from the measurement model for the Grade 2 MPAC interview.

Table 12. Grade 2 MPAC Interview Item Descriptions, Descriptives, and 2-pl UIRT Parameters

			Proportion	2-pl UIRT parameters	
Item	Factor	Item description	correct	Discrimination	Difficulty
Counting					
CNS 1	**COMP**	**removed for test security**	**.79**	**0.85**	**-1.23**
CNS 2	**COMP**	**removed for test security**	**.77**	**1.36**	**-0.88**
CNS 3	**COMP**	**removed for test security**	**.62**	**1.68**	**-0.33**
CNS 4	**COMP**	**removed for test security**	**.75**	**1.42**	**-0.81**
CNS 5	**COMP**	**removed for test security**	**.65**	**1.39**	**-0.44**
Word problems					
WP 6	—	removed for test security	.93	0.69	-2.63
WP 7	**WP**	**removed for test security**	**.63**	**1.01**	**-0.43**
WP 8	**WP**	**removed for test security**	**.55**	**0.81**	**-0.21**
WP 9	**WP**	**removed for test security**	**.72**	**1.05**	**-0.78**
WP 10	**WP**	**removed for test security**	**.56**	**0.90**	**-0.20**
WP 11	**WP**	**removed for test security**	**.59**	**1.52**	**-0.26**
WP 12	**WP**	**removed for test security**	**.69**	**0.79**	**-0.79**
WP 13	**WP**	**removed for test security**	**.55**	**0.87**	**-0.16**
Equations and computation					
EC 1	**NF**	**removed for test security**	**.98**	**1.42**	**-2.69**
EC 2	—	removed for test security	.90	0.51	-2.82
EC 3	**NF**	**removed for test security**	**.97**	**0.84**	**-2.94**
EC 4	—	removed for test security	.57	0.43	-0.39
EC 5	—	removed for test security	.56	0.59	-0.29
EC 6	**COMP**	**removed for test security**	**.82**	**1.03**	**-1.26**
EC 7	**COMP**	**removed for test security**	**.26**	**0.91**	**0.98**
EC 8	—	removed for test security	>.99	72.00	-2.83
EC 9	—	removed for test security	>.99	0.43	-8.08
EC 10	**OBS**	**removed for test security**	**.33**	**0.58**	**0.90**
EC 11	**ESRS**	**removed for test security**	**.65**	**0.88**	**-0.56**
EC 12	**ESRS**	**removed for test security**	**.61**	**0.87**	**-0.40**
EC 13	**ESRS**	**removed for test security**	**.59**	**0.60**	**-0.40**
EC 14	**NF**	**removed for test security**	**.91**	**0.68**	**-2.37**
EC 15	**OBS**	**removed for test security**	**.17**	**1.16**	**1.26**
EC 16	**OBS**	**removed for test security**	**.28**	**0.99**	**0.84**

Note. N = 286 valid Grade 2 student interviews conducted. Discrimination estimates use a 1.7 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994). Items that remained after factor analysis are presented in boldface type. NF = Number Facts; OBS = Operations on Both Sides of the Equal Sign; WP = Word Problems; ESRS = Equal Sign as a Relational Symbol; COMP = Computation

Two items were just above the 0.4 discrimination minimum acceptable value: EC 2 (0.51) and EC 4 (0.43). Figure 2 presents the Grade 2 difficulty-vs.-discrimination scatterplot with EC 8 and EC 9 included and Figure 3 presents the same scatterplot with EC 8 and EC 9 removed. Figure 3 reveals EC 2 and EC 4 to be located in regions on the difficulty continuum where other items are also located. Accordingly, EC

2 and EC 4 warranted no special consideration for retention, but they were nevertheless used in the initial correlated trait models for further evaluation.

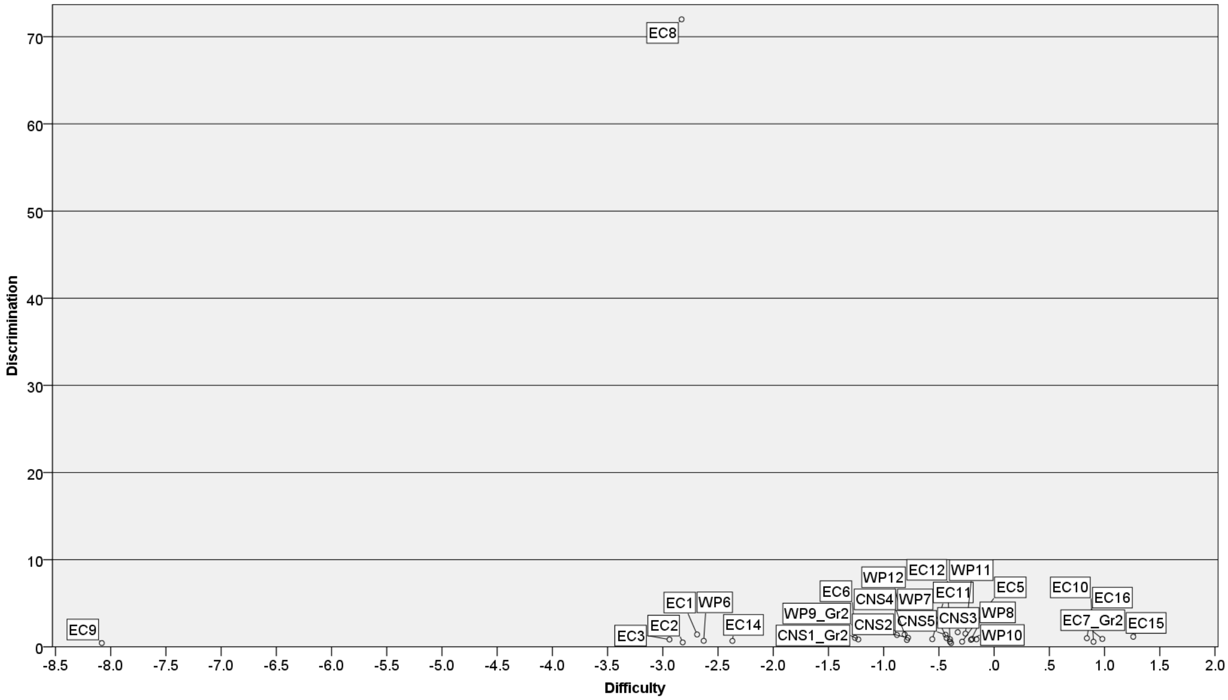


Figure 2. Grade 2 MPAC interview 2-pl UIRT difficulty-vs.-discrimination scatterplot (all items). Items with "_Gr2" in the label are unique to the Grade 2 interview.

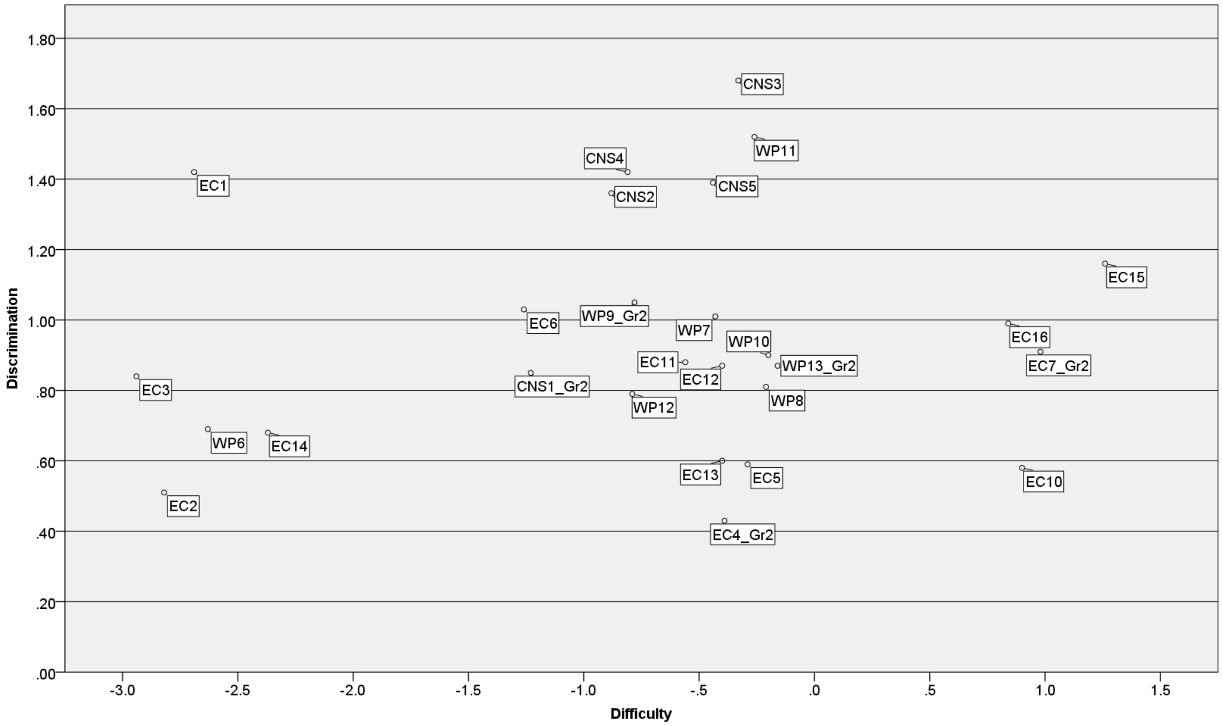


Figure 3. Grade 2 MPAC interview 2-pl UIRT difficulty-versus-discrimination scatterplot minus outliers. Items with "_Gr2" in the label are unique to the Grade 2 interview.

4.3. Correlated Trait Model Evaluation

4.3.1. Grade 1 Correlated Trait Model Evaluation

The initial Grade 1 model contained all items except EC 8 and EC 9, which were dropped during the item screening step. All items in the initial model had statistically significant unstandardized factor loading (*p* < .05). Five items had standardized factor loadings that were below or near the factor loading minimum acceptable value of .5: EC 2 (.42), WP 6 (.58), CNS 1 (.46), EC 5 (.52), EC 7 (.55). Upon inspection of their standardized loadings and their representation of the range of item difficulty, as well as consideration of their relative contribution toward the content validity of the scale, we decided that all but one of these items (EC 7) could be dropped for the revised model.

We then fit the data for the final set of Grade 1 items to a correlated trait structure and evaluated the factorial validity of the model on the basis of overall goodness of fit and interpretability, size, and statistical significance of the parameter estimates. The revised Grade 1 correlated trait model RMSEA, CFI, and TLI indicated close fit: $\chi^2(199) = 263.091$, *p* = .002; RMSEA = .031, 90% CI [.020,.041]; CFI = .987; and TLI = .985. All unstandardized factor loadings for the revised Grade 1 model were statistically significant. Table 13 presents the standardized factor loadings for the initial and revised correlated trait model. All standardized factor loadings for the revised Grade 1 model were above the minimum acceptable value of .5, and most were well above the target of .7.

Table 13. Grade 1 Standardized Factor Loadings for Initial and Revised Correlated Trait Model

		Initial model		Revised model	
Factor	Indicator description	Estimate	(*SE*)	Estimate	(*SE*)
Number Facts by					
EC 1	Removed for security	.728	(.089)	.684	(.085)
EC 2	Removed for security	.419	(.078)	—	—
EC 3	Removed for security	.711	(.077)	.698	(.074)
EC 14	Removed for security	.991	(.057)	.954	(.063)
Operations on Both Sides of the Equal Sign by					
EC 10	Removed for security	.612	(.073)	.611	(.073)
EC 15	Removed for security	.896	(.068)	.899	(.067)
EC 16	Removed for security	.893	(.050)	.890	(.049)
Word Problems by					
WP 6	Removed for security	.578	(.075)	—	—
WP 7	Removed for security	.832	(.038)	.837	(.037)
WP 8	Removed for security	.801	(.041)	.804	(.041)
WP 9	Removed for security	.693	(.051)	.684	(.052)
WP 10	Removed for security	.696	(.050)	.695	(.051)
WP 11	Removed for security	.853	(.036)	.852	(.037)
WP 12	Removed for security	.807	(.043)	.814	(.043)
Equal Sign as a Relational Symbol by					
EC 11	Removed for security	.716	(.067)	.716	(.067)
EC 12	Removed for security	.883	(.045)	.876	(.045)
EC 13	Removed for security	.782	(.046)	.789	(.048)
Computation by					
CNS 1	Removed for security	.455	(.110)	—	—
CNS 2	Removed for security	.851	(.033)	.849	(.033)
CNS 3	Removed for security	.861	(.030)	.866	(.030)
CNS 4	Removed for security	.873	(.028)	.873	(.028)
CNS 5	Removed for security	.909	(.025)	.915	(.024)
EC 4	Removed for security	.681	(.051)	.675	(.053)
EC 5	Removed for security	.521	(.062)	—	—
EC 6	Removed for security	.633	(.057)	.612	(.060)
EC 7	Removed for security	.548	(.078)	.545	(.081)

Note. *N* = 336. Items EC 8 and EC 9 were dropped before fitting the initial model.

Table 14 presents the correlations among the factors for the Grade 1 model. All interfactor correlations were statistically significant and moderate to large in size. No interfactor correlations were so large as to suggest colinearity, but two were notably high: Word Problems with Number Facts (*r* = .843) and Word Problems with Computation (*r* = .858). Figure 4 illustrates the correlated factor structure and standardized factor loadings for the revised Grade 1 model.

Table 14. Grade 1 Factor Correlations for the Revised Correlated Trait Model

Factors	NF	OBS	WP	ESRS	COMP
Number Facts	—				
Operations on Both Sides of the Equal Sign	.558	—			
Word Problems	.843	.674	—		
Equal Sign as a Relational Symbol	.564	.708	.636	—	
Computation	.761	.724	.858	.658	—

Note. *N* = 336.

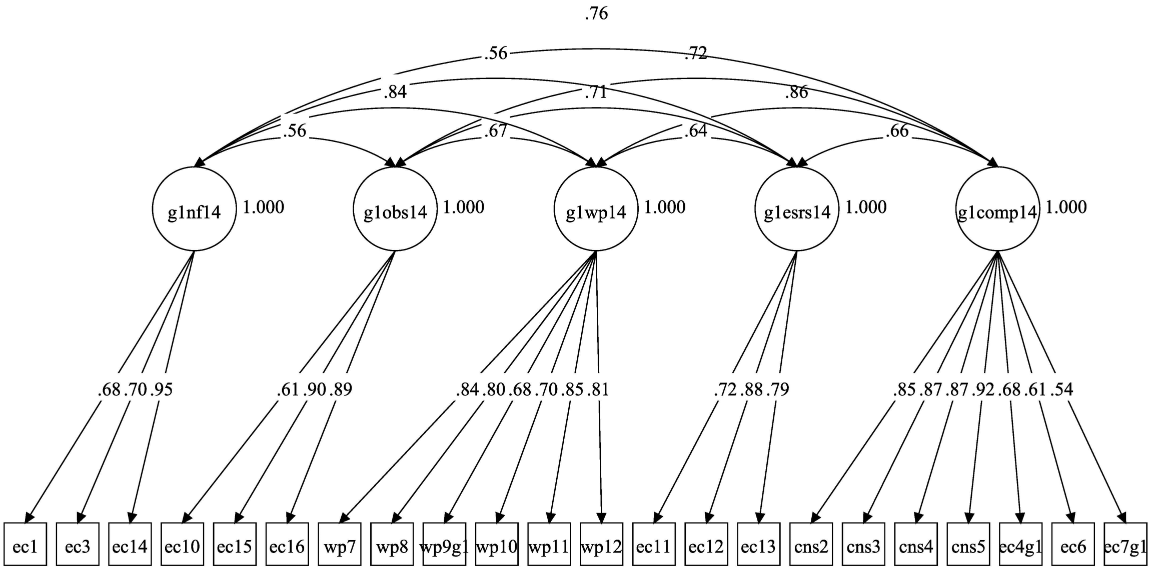


Figure 4. Grade 1 revised model: correlated-trait model diagram with standardized parameter estimates.

4.3.2. Grade 2 Correlated-Trait Model Evaluation

The initial Grade 2 model contained all items except EC 8 and EC 9, which were dropped during the item screening step. All items in the initial model had statistically significant unstandardized factor loading ($p < .05$). Four items had standardized factor loadings that were below or near the factor loading minimum acceptable value of 0.5: EC 2 (.54), WP 6 (.60), EC 4 (.42), EC 5 (.56). Upon inspection of their standardized loadings and their representation of the range of item difficulty, as well as consideration of their relative contribution toward the content validity of the scale, we determined that all could be dropped for the revised model.

We then fit the data for the final set of Grade 2 items to a correlated trait structure and evaluated the factorial validity of the model on the basis of overall goodness of fit and interpretability, size, and statistical significance of the parameter estimates. The revised Grade 2 correlated-trait model RMSEA, CFI, and TLI indicated close fit: $\chi^2(220) = 263.951$, $p = .023$; RMSEA = .026, 90% CI [.011, .038]; CFI = .988; and TLI = .986. All unstandardized factor loadings for the revised Grade 2 model were statistically significant. Table 15 presents the standardized factor loadings for the initial and revised correlated trait model. All standardized factor loadings for the revised Grade 2 model were above the minimum acceptable value of .5, and most were well above the target of .7.

Table 15. Grade 2 Standardized Factor Loadings for Initial and Revised Correlated Trait Model

		Initial Model		Revised Model	
Factor	Indicator Description	Estimate	(*SE*)	Estimate	(*SE*)
Number Facts by					
EC 1	Removed for security	.872	(.126)	.760	(.148)
EC 2	Removed for security	.542	(.117)	—	—
EC 3	Removed for security	.708	(.153)	.642	(.162)
EC 14	Removed for security	.666	(.120)	.575	(.146)
Operations on Both Sides of the Equal Sign by					
EC 10	Removed for security	.632	(.085)	.638	(.084)
EC 15	Removed for security	.910	(.070)	.909	(.070)
EC 16	Removed for security	.909	(.055)	.908	(.056)
Word Problems by					
WP 6	Removed for security	.595	(.113)	—	—
WP 7	Removed for security	.751	(.048)	.742	(.050)
WP 8	Removed for security	.675	(.052)	.682	(.052)
WP 9	Removed for security	.774	(.055)	.774	(.056)
WP 10	Removed for security	.725	(.048)	.722	(.050)
WP 11	Removed for security	.865	(.036)	.864	(.038)
WP 12	Removed for security	.678	(.058)	.677	(.059)
WP 13	Removed for security	.710	(.050)	.688	(.052)
Equal Sign as a Relational Symbol by					
EC 11	Removed for security	.847	(.069)	.847	(.068)
EC 12	Removed for security	.896	(.042)	.893	(.042)
EC 13	Removed for security	.735	(.056)	.738	(.055)
Computation by					
CNS 1	Removed for security	.661	(.063)	.659	(.063)
CNS 2	Removed for security	.833	(.044)	.842	(.043)
CNS 3	Removed for security	.875	(.030)	.883	(.031)
CNS 4	Removed for security	.850	(.040)	.861	(.039)
CNS 5	Removed for security	.840	(.036)	.848	(.036)
EC 4	Removed for security	.423	(.069)	—	—
EC 5	Removed for security	.558	(.061)	—	—
EC 6	Removed for security	.735	(.060)	.740	(.060)
EC 7	Removed for security	.689	(.055)	.666	(.057)

Note. *N* = 286. Items EC 8 and EC 9 were dropped before fitting of the initial model.

Table 16 presents the correlations among the factors for the Grade 2 model. All interfactor correlations were statistically significant and moderate to large in size. No interfactor correlations were so large as to suggest colinearity, but three were notably high: Word Problems with Number Facts (*r* = 0.904), Word

Problems with Computation (*r* = .891), and Number Facts with Computation (*r* = 0.868). Figure 5 illustrates the correlated factor structure and standardized factor loadings for the revised Grade 2 model.

Table 16. Grade 2 Factor Correlations for the Revised Correlated Trait Model

Factors	NF	OBS	WP	ESRS	COMP
Number Facts	—				
Operations on Both Sides of the Equal Sign	.432	—			
Word Problems	.904	.568	—		
Equal Sign as a Relational Symbol	.713	.750	.659	—	
Computation	.868	.667	.891	.657	—

Note. *N* = 286.

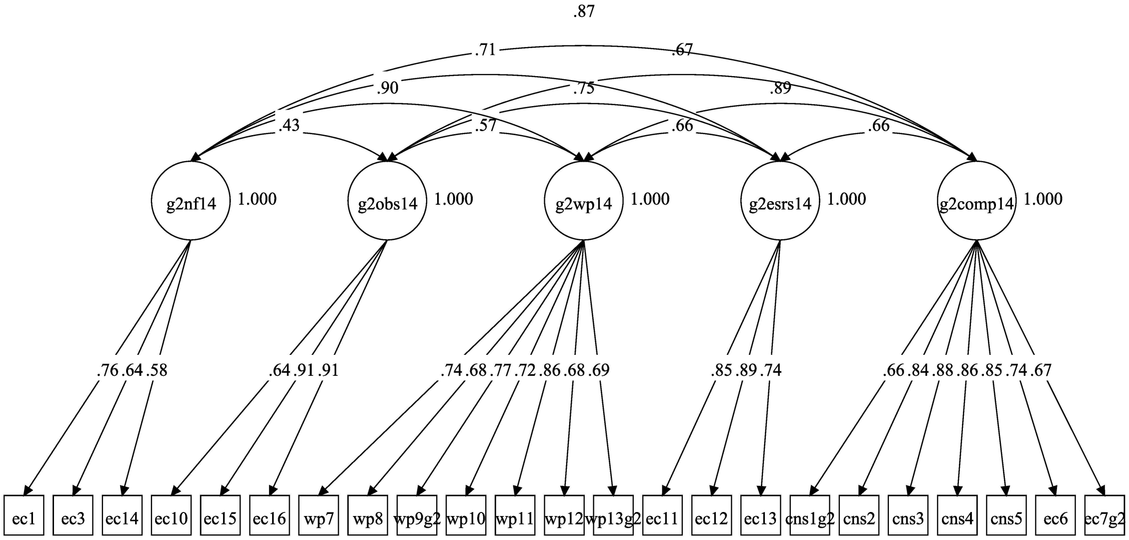


Figure 5. Grade 2 revised model: correlated-trait model diagram with standardized parameter estimates.

4.4. Higher-Order Model Evaluation

4.4.1. Grade 1 Higher-Order Model Evaluation

The Grade 1 higher-order model RMSEA, CFI, and TLI indicated close fit: $\chi^2(204) = 281.690$, $p < .001$; RMSEA = .034, 90% CI [.023, .043]; CFI = .984; and TLI = .982. The differences between factor loading estimates for the correlated-trait and higher-order factor model were negligible, varying less than .01 in absolute value for each item. Given the negligible decrement in model fit and maintenance of close fit for all indices and negligible variation in factor loading estimates between models, we determined the more parsimonious higher-order structure to be appropriate for modeling the Grade 1 interview data. Table 17 presents the standardized factor loadings for the Grade 1 (and Grade 2) higher-order measurement model. Figure 6 illustrates the higher-order factor structure and standardized factor loadings for the final Grade 1 model.

Table 17. Standardized Factor Loadings for Grade 1 and Grade 2 Higher-Order Measurement Models

		Grade 1 interview		Grade 2 interview	
Factor	Indicator description	Estimate	(*SE*)	Estimate	(*SE*)
		Standardized first-order factor loadings			
Number Facts by					
EC 1	Removed for security	.683	(.084)	.745	(.148)
EC 3	Removed for security	.696	(.075)	.646	(.165)
EC 14	Removed for security	.956	(.063)	.583	(.147)
Operations on Both Sides of the Equal Sign by					
EC 10	Removed for security	.607	(.073)	.637	(.086)
EC 15	Removed for security	.898	(.067)	.905	(.070)
EC 16	Removed for security	.894	(.050)	.912	(.056)
Word Problems by					
WP 7	Removed for security	.837	(.037)	.743	(.050)
WP 8	Removed for security	.804	(.041)	.681	(.052)
WP 9_Gr1	Removed for security	.685	(.052)		
WP 9_Gr2	Removed for security			.773	(.056)
WP 10	Removed for security	.695	(.051)	.722	(.050)
WP 11	Removed for security	.852	(.037)	.867	(.038)
WP 12	Removed for security	.814	(.043)	.675	(.059)
WP 13_Gr2				.686	(.052)
Equal Sign as a Relational Symbol by					
EC 11	Removed for security	.716	(.067)	.846	(.069)
EC 12	Removed for security	.878	(.045)	.894	(.043)
EC 13	Removed for security	.788	(.049)	.737	(.057)
Computation by					
CNS 1_Gr2	Removed for security			.659	(.063)
CNS 2	Removed for security	.848	(.033)	.842	(.043)
CNS 3	Removed for security	.866	(.030)	.883	(.031)
CNS 4	Removed for security	.873	(.028)	.862	(.039)
CNS 5	Removed for security	.915	(.024)	.848	(.036)
EC 4_Gr1	Removed for security	.675	(.053)		
EC 6	Removed for security	.612	(.060)	.740	(.060)
EC 7_Gr1	Removed for security	.545	(.081)		
EC 7_Gr2	Removed for security			.665	(.057)
		Standardized second-order factor loadings			
Math by					
NF	NF latent variable	.843	(.058)	.924	(.164)
OBS	OBS latent variable	.776	(.043)	.708	(.048)
WP	WP latent variable	.924	(.033)	.918	(.031)
ESRS	ESRS latent variable	.724	(.048)	.749	(.047)
COMP	COMP latent variable	.923	(.029)	.948	(.027)

Note. Grade 1 *n* = 336. Grade 2 *n* = 286. NF = Number Facts; OBS = Operations on Both Sides of the Equal Sign; WP = Word Problems; ESRS = Equal Sign as a Relational Symbol; COMP = Computation

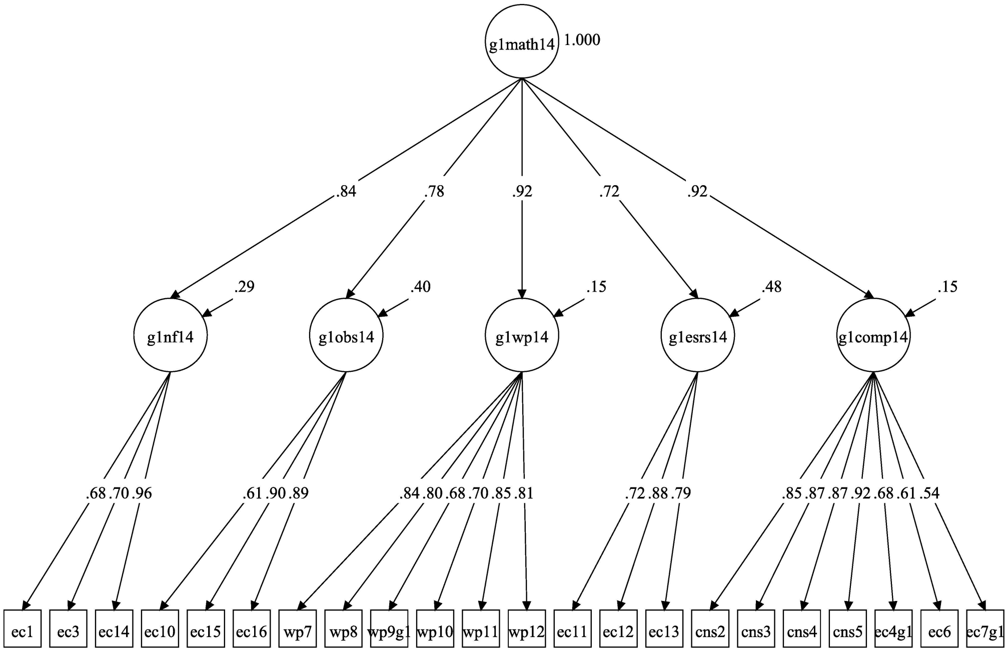


Figure 6. Grade 1 final model: higher-order factor diagram with standardized parameter estimates.

4.4.2. Grade 2 Higher-Order Model Evaluation

The Grade 2 higher-order model RMSEA, CFI, and TLI indicated close fit: $\chi^2(225) = 301.747$, $p < .001$; RMSEA = .035, 90% CI [.023, .044]; CFI = .979; and TLI = .976. The differences between factor loading estimates for the correlated-trait and higher-order factor model were negligible, typically varying less than 0.01 in absolute value. Given the negligible decrement in model fit and maintenance of close fit for all indices and negligible variation in factor loading estimates between models, we determined the more parsimonious higher-order structure to be appropriate for modeling the Grade 2 interview data. Table 17 presents the standardized factor loadings for the Grade 2 (and Grade 1) higher-order measurement model. Figure 7 illustrates the higher-order factor structure and standardized factor loadings for the final Grade 2 model.

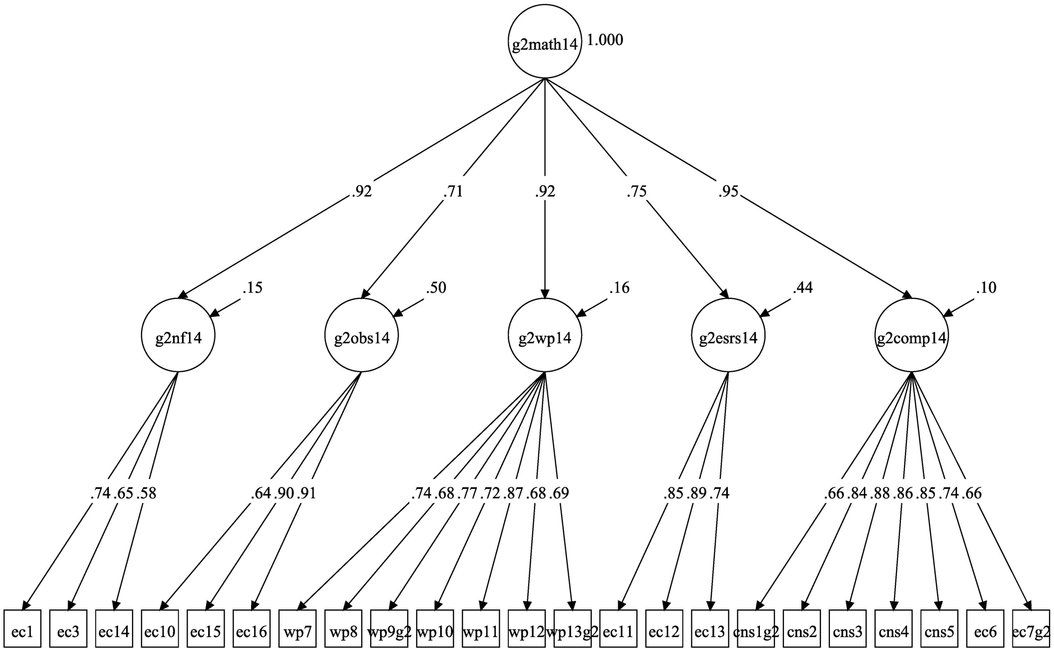


Figure 7. Grade 2 final model: Higher-order factor diagram with standardized parameter estimates.

4.5. Scale Reliability Evaluation

4.5.1. Grade 1 Scale Reliabilities

The scale reliabilities for the Grade 1 MPAC interview suggested acceptable reliability for all scales. Using the following equation for Grade 1 higher-order Math factor composite reliability,

$$\frac{(0.806+0.693+0.753+0.570+0.503)^2}{(0.806+0.693+0.753+0.570+0.503)^2+(0.044+0.295+0.097+0.318+0.265)}=0.916,$$

where the numerator is the squared sum of the unstandardized second-order factor loadings and the denominator is the squared sum of the unstandardized second-order factor loadings plus the sum of the first-order factor residual variances, we calculated a composite reliability for the Grade 1 higher-order Math factor of .92, which exceeds the target reliability of .8.

Table 18 presents the α, β, and ω_h ordinal reliability coefficients for the reduced set of items by subscale and for the total scale. The α estimates for all subscales exceeded the target of .8, except for the ESRS scale, which had an α reliability of .78. Comparison between the αs and βs revealed a range of discrepancies, some small (such as for the WP scale, where α = .90 and β = .88), some moderate (such as for the COMP scale, where α = .90 and β = .83), and others large (such as for the OBS scale, where α = .84 and β = .66). The magnitudes of discrepancies indicate heterogeneity among the factor loadings, challenging the assumption of essential tau equivalence. Comparison between the α and ω_h coefficients revealed discrepancies to be small to moderate for the subscales (range .01 to .07) and large for the total scale (.18). Where α exceeds ω_h (i.e., for WP, ESRS, COMP, and Math), the α to ω_h discrepancies indicate the presence of multidimensionality within the scales. Where ω_h exceeds α (i.e., for NF and

OBS), variability was present in the general factor loadings, but group factor loadings were relatively small, indicating that lumpiness in the scale was not attributable to multidimensionality. In every case, ω_h exceeded the conventional minimum value of .7. As demonstrated by Gustafsson and Aberg-Bengtsson (2010), high values of ω_h indicate that composite scores can be interpreted as reflecting a single common source of variance despite evidence of some within-scale multidimensionality.

Table 18. Grade 1 MPAC Interview Scale Reliability Estimates

		Reliability		
Scale	*N* items	α	β	ω_h
Number Facts	3	.82	.73	.83
Operations on Both Sides of the Equal Sign	3	.84	.66	.89
Word Problems	6	.90	.88	.84
Equal Sign as a Relational Symbol	3	.78	.54	.75
Computation	7	.90	.83	.83
Math	22	.95	.88	.77

Note. Sample *N* = 336. α, β, and ω_h are ordinal forms of Cronbach's alpha, Revelle's beta, and McDonald's omega hierarchical, respectively.

Inspection of the 2-pl UIRT TIC in Figure 8 reveals the information curve for the Grade 1 MPAC interview to exceed 2.33 (reliability of .7) for the ability range of approximately -2.0 through 2.2. Given the sample descriptives (M = -0.002, SD = 0.942, Min = -2.345, and Max = 2.320), this suggests acceptable reliability of the scale for over 97% of the sample and nearly the full range of observed abilities. The information curve exceeds 4 (reliability of .8) for the ability range of approximately -1.2 through 1.9, indicating target reliability of the scale was achieved for approximately 89% of the sample.[4]

[4] Areas under the normal distribution were calculated with the online normal distribution calculator found at http://onlinestatbook.com/2/calculators/normal_dist.html

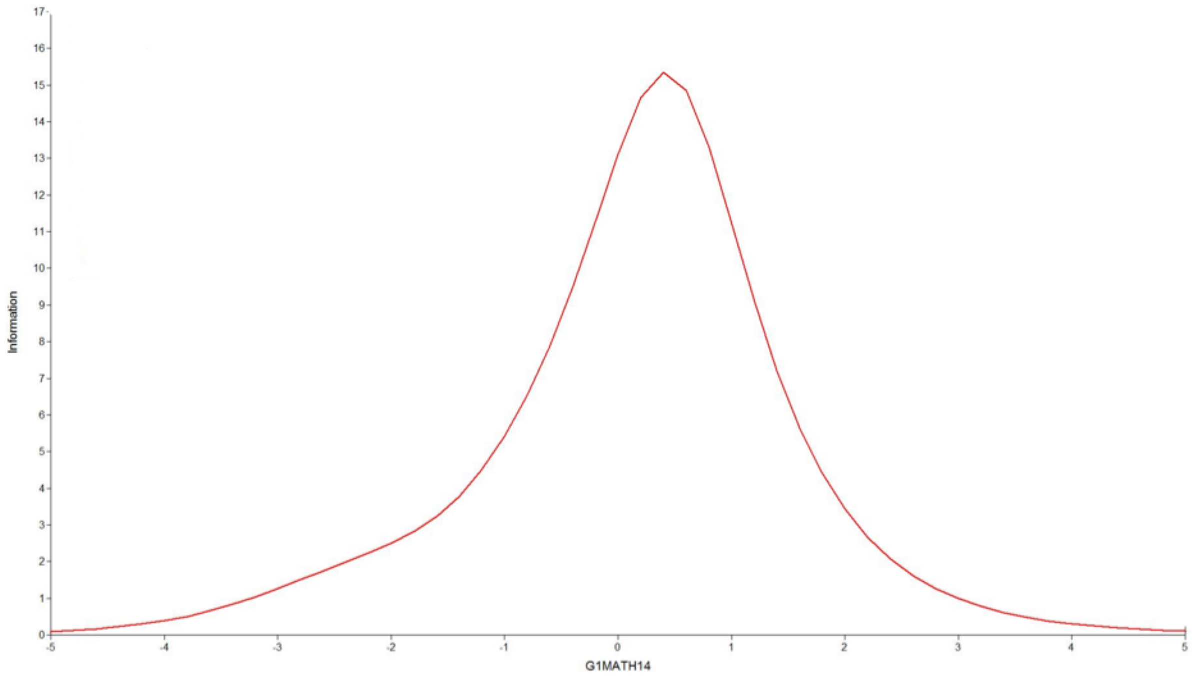


Figure 8. Grade 1 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

Figure 9 presents the overall distribution of number of items answered correctly in Grade 1 for the reduced set of items. Similar figures for each subscale are provided in Appendix E. Interested readers will find information about the most common incorrect responses to the various items in Appendix F.

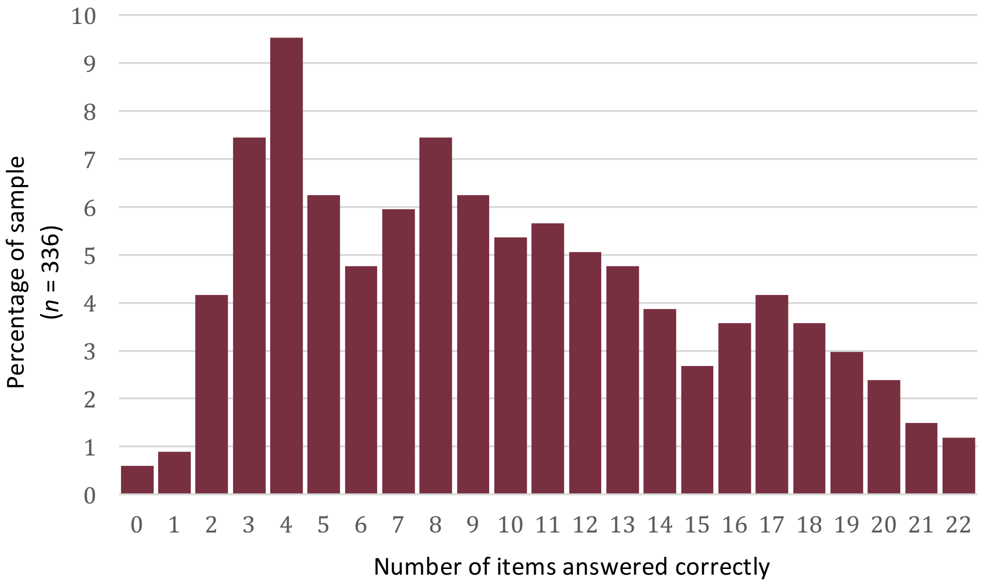


Figure 9. Distribution of the number of items individual students in the Grade 1 sample answered correctly on the reduced set of items.

4.5.2 Grade 2 Scale Reliabilities

The scale reliabilities for the Grade 2 MPAC interview suggested acceptable reliability for all scales. Using the following equation for Grade 2 higher-order Math factor composite reliability,

$$\frac{(0.539+0.646+0.630+0.552+0.631)^2}{(0.539+0.646+0.630+0.552+0.631)^2+(0.045+0.239+0.074+0.415+0.050)}=0.916,$$

where the numerator is the squared sum of the unstandardized second-order factor loadings and the denominator is the squared sum of the unstandardized second-order factor loadings plus the sum of the first-order factor residual variances, we calculated a composite reliability for the Grade 2 higher-order Math factor of .92, which exceeds the target reliability of .8.

Table 19 relays the α, β, and ω_h ordinal reliability coefficients for the reduced set of items by subscale and for the total scale. All α estimates for all subscales exceeded the target of .80, except for the NF scale, which had an α reliability of .65. As with the Grade 1 interview, comparison between the αs and βs revealed a range of discrepancies (range .00 to .24), challenging the assumption of essential tau equivalence where the discrepancy was sizable. Comparison between the α and ω_h coefficients also revealed a range of discrepancies (range .01 to .25). Where α exceeded ω_h (i.e., OBS, WP, COMP, and Math), the α to ω_h discrepancies indicated the presence of multidimensionality within the scales. Where ω_h exceeded α (i.e., NF and ESRS), variability was present in the general factor loadings, but group factor loadings were relatively small, indicating that lumpiness in the scale was not attributable to multidimensionality. In every case, ω_h met or exceeded the conventional minimum value of .70, suggesting that composite scores can be interpreted as reflecting a single common source of variance in spite of evidence for some within-scale multidimensionality (Gustafsson & Aberg-Bengtsson, 2010).

Table 19. Grade 2 Scale Reliability Estimates

	Number	Reliability		
Scale	of items	α	β	ω_h
Number Facts	3	.65	.65	.70
Operations on Both Sides of the Equal Sign	3	.84	.61	.83
Word Problems	7	.89	.82	.77
Equal Sign as a Relational Symbol	3	.80	.56	.86
Computation	7	.91	.85	.78
Math	23	.95	.88	.70

Note. Sample *N* = 286. α, β, and ω_h are ordinal forms of Cronbach's alpha, Revelle's beta, and McDonald's omega hierarchical, respectively.

Inspection of the 2-pl UIRT TIC in Figure 10 reveals the information curve for the Grade 2 MPAC interview to exceed 2.33 (reliability of .70) for the ability range of approximately -3.1 through 1.9. Given the sample descriptives (M = -0.003, SD = 0.941, Min = -2.929, and Max = 1.965), this suggests acceptable reliability of the scale for over 97% of the sample and nearly the full range of observed abilities. The information curve exceeds 4 (reliability of .80) for the ability range of approximately -2.2 through 1.2, indicating that target reliability of the scale was achieved for approximately 89% of the sample.

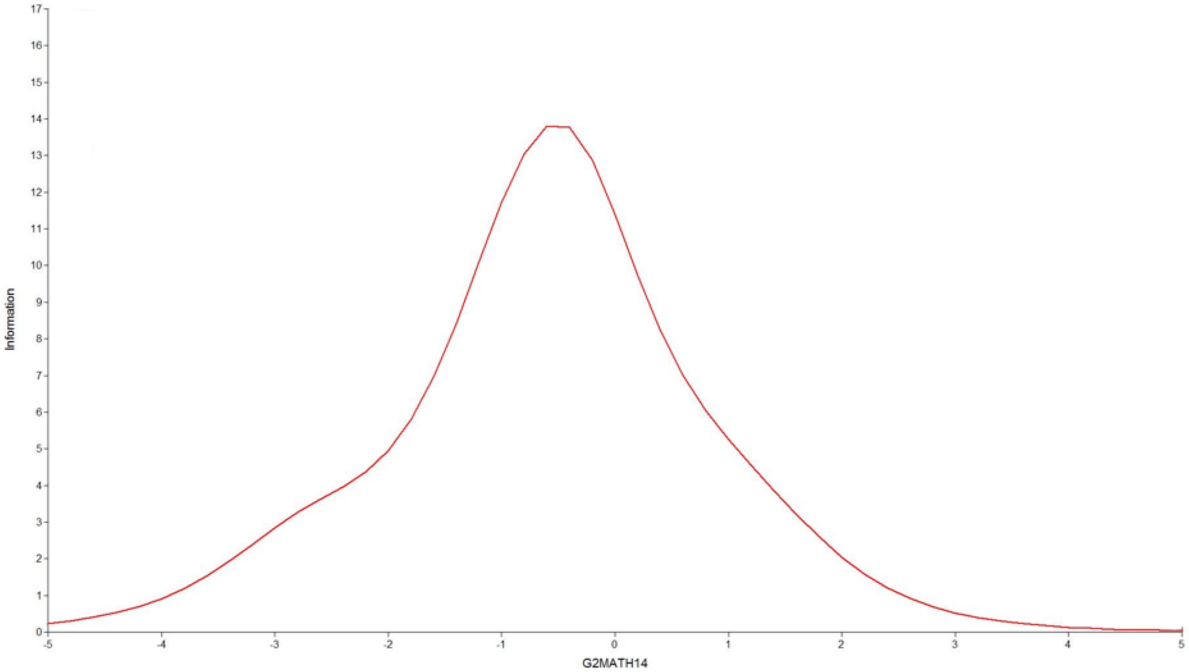


Figure 10. Grade 2 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

Figure 11 presents the overall distribution of number of items answered correctly in Grade 2 for the reduced set of items. Similar figures for each subscale are provided in Appendix E. Interested readers will find information about the most common incorrect responses to the various items in Appendix F.

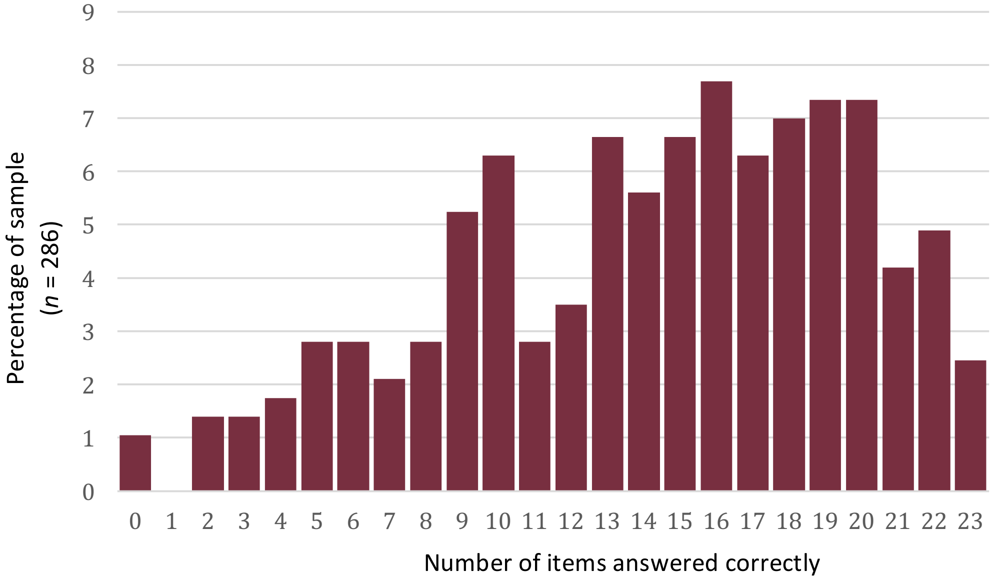


Figure 11. Distribution of the number of items individual students in the Grade 2 sample answered correctly on the reduced set of items.

4.6. Concurrent Validity Evaluation

4.6.1. Grade 1 MPAC Concurrent Validity

The correlations between the Grade 1 MPAC student interview and the DEA and ITBS were consistently moderate to large in size, providing evidence of concurrent validity of the student interview. See Table 20 for correlations between manifest factor scores for the interview scales, numbers items answered correct for the DEA scales, and standard scores for the ITBS tests. Using correlations > .70 to indicate scale correspondence revealed a pattern of correspondence between the MPAC interview Total, NF, WP, and COMP subscales and the DEA total and ITBS Math Problems (ITBS-MP) test. Note that the correlation between the DEA total and the ITBS-MP test was smaller (r = .66) than that observed for the correlation between the student interview total and the ITBS-MP test (r = .74).

Note also that, although moderate correlations were found between the MPAC interview and the ITBS Math Computation (ITBS-MC) test (range .53 to .65), none of the correlations surpassed the .7 correspondence criterion. Nevertheless, the same is true for the correlations between the DEA and the ITBS-MC, where a pattern of smaller correlations was observed (range .42 to .59). The correlation between the ITBS tests was r = .60, larger than any of the DEA-with-ITBS-MC correlations and smaller than four of the six interview-with-ITBS-MC correlations. Moreover, the MPAC interview appeared to correspond better with the DEA and ITBS tests than the DEA and ITBS did with each other. All correlations were statistically significant at p < .001.

4.6.2. Grade 2 MPAC Concurrent Validity

The findings for the MPAC interview for Grade 2 were nearly identical to those for Grade 1. The correlations between the Grade 2 MPAC and the DEA and ITBS were consistently moderate to large, providing evidence of concurrent validity of the student interview. See Table 21 for correlations between manifest factor scores for the MPAC scales, number of items answered correctly for the DEA scales, and standard scores for the ITBS tests. Using correlations > .70 to indicate scale correspondence revealed a pattern of correspondence between the MPAC interview Total, NF, WP, and COMP subscales and the DEA total and ITBS Math Problems (ITBS-MP) test. Note that the correlation between the DEA total and the ITBS-MP test was smaller (r = .74) than that observed between the student interview total and the ITBS-MP test (r = .79).

Note also that although moderate correlations were found between the MPAC and the ITBS Math Computation (ITBS-MC) test (range .55 to .63), none of the correlations surpassed the .70 correspondence criterion. Nevertheless, the same is true for the correlations between the DEA and the ITBS-MC, where a pattern of smaller correlations was observed (range .34 to .60). The correlation between the ITBS tests was r = .60: as large as the largest DEA with ITBS-MC correlation and smaller than four of the six interview-with-ITBS-MC correlations. Moreover, like that for Grade 1, the Grade 2 MPAC student interview appeared to correspond better with the DEA and ITBS tests than the DEA and ITBS did with each other. All correlations were statistically significant at p < .01, except for the Grade 2 MPAC ESRS-with-DEA geometry correlation (p = .131).

Table 20. Correlations Among Grade 1 MPAC interview, Discovery Education Assessment (DEA), and Iowa Test of Basic Skills (ITBS)

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Grade 1 MPAC Interview												
(1) MPAC Total	—											
(2) Number Facts	.96	—										
(3) Operations on Both Sides of the Equal Sign	.89	.84	—									
(4) Word Problems	.98	.93	.84	—								
(5) Equal Sign as a Relational Symbol	.84	.79	.77	.79	—							
(6) Computation	.98	.91	.85	.93	.79	—						
Discovery Education Assessment												
(7) Total	**.72**	**.71**	.67	**.71**	.60	.69	—					
(8) Operations & Algebra	.68	.66	.66	.63	.54	.67	.82	—				
(9) Number/Operations Base Ten	.51	.48	.46	.50	.44	.49	.84	.56	—			
(10) Measurement & Data	.44	.43	.44	.45	.36	.39	.75	.48	.56	—		
(11) Geometry	.58	.59	.47	.60	.50	.53	.73	.45	.50	.40	—	
Iowa Test of Basic Skills												
(12) Math Problems, Level 7	**.74**	.69	.64	**.75**	.61	**.72**	.66	.54	.54	.50	.50	—
(13) Math Computation, Level 7	.65	.64	.53	.63	.56	.64	.59	.49	.49	.42	.44	.60

Note. Grade 1 MPAC interview *n* = 336. Discovery Education Assessment DEA *n* = 391. ITBS *n* = 1,103. MPAC with DEA correlation *n* = 95. MPAC with ITBS correlation *n* = 309. DEA with ITBS correlation *n* = 321. All correlations are statistically significant at *p* < .001. Correlations within borders signify correlations that indicate potential concurrent validity between measures. Values in boldface are concurrent validity correlations > .70, indicating ≥ .50 shared variance between measures. DEA version was the Common Core 2010 Skill Set. ITBS was Form C Level 7. The MPAC interview, DEA, and ITBS were all administered during spring 2014.

Table 21. Correlations Among Grade 2 MPAC interview, Discovery Education Assessment (DEA), and Iowa Test of Basic Skills (ITBS)

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Grade 2 MPAC Interview												
(1) MPAC Total	—											
(2) Number Facts	1.00	—										
(3) Operations on Both Sides of the Equal Sign	.86	.85	—									
(4) Word Problems	.98	.97	.80	—								
(5) Equal Sign as a Relational Symbol	.87	.86	.78	.82	—							
(6) Computation	.99	.98	.83	.95	.84	—						
Discovery Education Assessment												
(7) Total	.69	.69	.59	**.71**	.46	.67	—					
(8) Operations & Algebra	.50	.51	.42	.55	.39	.46	.78	—				
(9) Number/Operations Base Ten	.52	.52	.44	.51	.41	.50	.78	.48	—			
(10) Measurement & Data	.58	.58	.49	.59	.34	.58	.80	.49	.47	—		
(11) Geometry	.38	.37	.35	.38	.17	.38	.53	.27	.26	.25	—	
Iowa Test of Basic Skills												
(12) Math Problems, Level 8	**.79**	**.77**	.67	**.77**	.67	**.78**	**.74**	.52	.61	.59	.40	—
(13) Math Computation, Level 8	.63	.62	.55	.63	.55	.61	.60	.43	.52	.44	.34	.60

Note. Grade 2 MPAC Interview n = 286. DEA n = 269. ITBS n = 1, 069. MPAC-with-DEA correlation n = 78. MPAC-with-ITBS correlation n = 271. DEA-with-ITBS correlation n = 244. The correlation between the MPAC ESRS scale and the DEA Geometry scale was not statistically significant (p = .131); all other correlations were statistically significant at p < .01. Correlations within borders signify correlations that indicate potential concurrent validity between measures. Values given in boldface are concurrent validity correlations > .70, indicating ≥ .50 shared variance between measures. DEA version was the Common Core 2010 Skill Set. ITBS was Form C Level 8. The MPAC interview, DEA, and ITBS were all administered spring 2014.

5. Summary and Discussion

Overall, we have high confidence in the quality of data generated through the MPAC interviews. The interviewers exhibited high fidelity to the protocol, the coding had high reliability, the alpha-reliability and model fit indicated high internal consistency of the data, the reliability was sufficiently high across a broad range of ability levels, and the MPAC scores were highly correlated with those measured concurrently by means of other instruments with high policy relevance and established reliability and validity. The development process, including the 34 pilot interviews, surely contributed to the high reliability and validity. Any subsequent use of this interview will require extensive interviewer training to maintain scoring validity. Ultimately, we are confident that the MPAC interview is valid for use with first- and second-grade students to measure their achievement in the mathematical domain of number, operations, and equality.

Apart from offering a well-designed, field-tested, focused interview, we think that our conceptual and empirical specification of the five factors offers a new insight into important psychological constructs in mathematics and the measurement of these constructs. The five factors were the second attempt at a CFA model. A thorough discussion of the process and rationale for how we decided to organize the items into the five factors is beyond the scope of this report, but subsequent manuscripts will delve deeper into that particular line of inquiry.

5.1. Reliability and Validity

Our analyses indicate that (a) the measurement models met target criteria for factorial validity, (b) the subscales and total scores had acceptable reliability of measurement, and (c) the interviews were significantly correlated with policy-relevant, standardized measures of student mathematics.

In our investigation of the validity and reliability of the Grade 1 and Grade 2 MPAC interviews, our respecification procure resulted in measurement models with close model fit and factor loadings and factor correlations all within acceptable range. Although some items were dropped, the final set of items achieved the content coverage initially intended for each factor and the assessments as a whole. Scale reliability estimates suggested acceptable reliability at both the subscale level and total score. How these items and scales would perform with students other than first or second graders in the U.S. is not known, so all results herein should be interpreted with respect to the population in our sample.

The validity of the MPAC interview as an instrument with which to measure student achievement is supported by expert review of the content of the assessment, good reliability and model fit statistics, and observed correlations between student achievement on the interview and achievement on other instruments in wide use by states and districts. Statistically significant and moderately sized correlations were found between the Grade 1 and Grade 2 MPAC interview and two standardized measures of student mathematics: the Discovery Education Assessment and Iowa Test of Basic Skills.

We chose to use the higher-order model to define an overall achievement score on the interview, but the correlated-traits model also had close fit. Using a correlated traits model with this interview to split the outcome into a more granular set of topics does appear to be a defensible approach.

5.2. Development Process

The reliability and validity of the achievement data resulting from the interviews was most certainly strengthened by the development process. As a starting point, the development team had a strong knowledge of the content area (i.e., number, operations, and equality at the early elementary level), research on student thinking in this area, and experience working with children. Nonetheless, the many rounds of feedback from experts in the field, including Thomas Carpenter, Victoria Jacobs, and the interviewers themselves served to improve the interview. Without a doubt, the students who participated in the pilot tests of the interview provided very important feedback that improved the interview. The interviewer training procedures are critically important for the validity and reliability of the interview data.

5.3. Reflections and Next Steps

We have learned a tremendous amount through the work we have done in developing the interview protocol, training the interview team, implementing the interviews with students, and analyzing the resulting data. Several aspects of the interview might be revised in future endeavors. Some examples are discussed here.

Among the True/False items, we encountered several instances where students provided correct answers based on faulty reasoning. The most common culprits were those equations that were false and included operator symbols on the right-hand side of the equal sign. In these cases, we were able to use student reasoning data to improve the response by adding reasoning as a criterion: to be counted as correct the student had to provide valid reasoning as well as a correct response. Although we do not delve here into student reasoning, future reports focused on students' reasoning processes will provide further analysis of this issue. Future versions of the MPAC interview will be revised to use items that minimize this problem. True/False tasks are explicitly referenced in the Common Core State Standards for Mathematics (NGA & CCSSO, 2010), so we expect other tests to include these types of items. The developers and users of those tests must watch for this problem or risk low reliability and misinterpretation of their assessment results.

The items that were removed in post interview analyses and did not contribute to the final achievement model fall into two categories. Some (e.g., EC 8 and EC 9) were designed to help students make the transition between sections of the interview and were not intended to be used for measuring achievement. These transition-oriented items (or items similar to them) should be retained in future versions of the interview. Other items were expected to be included in the achievement measurement, but they were removed as a result of poor item statistics such as low factor loadings. For the sake of efficiency, those items will not be used in future versions of the interview.

The counting items did not seem to be a very effective introductory section to the MPAC interview. The students seemed surprised or confused by the questions, perhaps because they are not part of their experience in mathematics class. Basic number-fact questions might be more familiar and serve as a better warm-up, and they lend themselves better to asking students to explain their thinking. (Those first few interactions involve showing the student what the interviewer expects them to do, and counting items do not offer as many opportunities to ask students to explain their thinking—a major goal of this interview.) We do maintain our original conviction that helping the student get comfortable at the start of the interview is important. We think the best way to do that is to start with something they will recognize as normal and not too difficult. Starting with a set of basic number facts might be a better option.

Overall, the reliability of the MPAC interview is higher than that of most instruments used in educational research, but its reliability might still be improved. Examination of Figures 8–11 suggests that the difficulty level of the interviews for the two grade levels could be adjusted to achieve high reliability across a larger proportion of the sample. The Grade 2 interview had acceptable reliability (i.e., >.70) for over 97% of the sample and nearly the full range of observed abilities, but the total information curve exceeded the target threshold for reliability of .80 between the ability range of -2.2 to 1.2, which constitutes approximately 89% of the sample. The Grade 1 interview also had acceptable reliability for over 97% of the sample and nearly the full range of observed abilities, but the information curve exceeded the target threshold for reliability of .80 for the ability range of approximately -1.2 through 1.9, which constitutes approximately 89% of the sample once again. Although this result may be satisfactory, potential revisions to the instrument will attempt to set the bar higher, so to speak, and exceed a target threshold of reliability of .80 for the entire ability range of ± 2 on both tests. Doing so may involve adding some items at higher ranges of difficulty to the Grade 2 interview and at lower ranges of difficulty to the Grade 1 interview. These items might replace those that were dropped during model respecification in the current study.

With such high proportion of items common to both the Grade 1 and Grade 2 interview, the prospect of creating a vertical scale across the grades is compelling. Nevertheless, our immediate evaluation goals prompted us to conduct analyses within each grade level, so at this time, we have not investigated the viability of this prospect

The following list summarizes areas that we might consider trying to improve if we are so fortunate as to have a chance to repeat this study.

1. Rethink some of the equations used in True-False questions to decrease the probability that students will arrive at correct answers based on incorrect reasoning.
2. Consider dropping items that were eliminated during screening and IFA modeling, but keep the ones in the True-False section designed to serve as a warm-up and practice.
3. Consider removing the Counting section and replacing it with a set of questions asking students to give some basic number facts as a strategy to ease the students into the interview.
4. Examine first-grade difficulty. The first-grade interview might have been slightly too difficult. Future development may involve replacing the items that were not used in the final measurement model with some less difficult problems to improve the instrument's ability to discriminate reliably among students at lower levels of knowledge.
5. Examine second-grade difficulty. The second grade interview might have been slightly too easy. Future development may involve replacing the items that were not used in the final measurement model with some more difficult problems to improve the instrument's ability to discriminate reliably among students at high levels of knowledge.
6. Examine the feasibility of vertical-scaling of the two interviews.

5.4. In Closing

The interview uncovered a wealth of information about children's thinking processes in mathematics, and we are publishing the entire protocol, history of its development, descriptions of the scoring procedures, and some lessons learned in the hope that others might find it useful for their own work.

The present report focuses on analyzing the data as an achievement measure, so the focus of analysis here falls on that relatively simplistic aspect of whether students generated correct answers and how to model patterns in correct and incorrect answers to yield insight into how much knowledge and ability in mathematics a student presents to the interviewer. The study described herein also investigates how that knowledge is organized in the child's mind through factor-analytic techniques aimed at clarifying the constructs. Future reports will delve into other aspects of what we found in the interviews, including relational thinking, frequency of various student answers and strategies for each problem, and additive or subtractive approaches to solving problems involving subtraction.

References[5]

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293–295.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26* (4), 385-531.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction*. Portsmouth, NH: Heinemann.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Portsmouth, NH: Heinemann.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494.

Discovery Education Assessment (2010). *Discovery Education's Common Core Mathematics grade 1 and grade 2 interim benchmark assessment*. Silver Spring, MD: Discovery Education.

Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., & Bray, G. B. (2008). *Iowa Test of Basic Skills®, Form C, Level 7*. Rolling Meadows, IL: Riverside Publishing.

Embretson, S.E. & Reise, S. P. (2000). *Item response theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children’s understanding of equality: A foundation for algebra. *Teaching Children Mathematics 6,* 232–236.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3). Available online: http://pareonline.net/getvn.asp?v=17&n=3

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.

Ginsburg, H. (1997). *Entering the child’s mind: The clinical interview in psychological research and practice*. Cambridge, UK: Cambridge University Press.

[5] Additional readings that may be helpful for understanding the content presented in the present report are listed in Appendix G.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association.

Hu L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55, doi: 10.1080/10705519909540118

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38*(3), 258–288.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341.

Muthén, L. K. and Muthén, B. O. (1998-2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, D.C.: Author.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nye, C. D. & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570.

R Development Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of personality assessment, 93*(3), 213-224.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57–74.

Revelle, W. (2016). *psych: Procedures for personality and psychological research* (Version 1.6.6). Evanston, Illinois: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Smith, J. P., III (1995). Competent reasoning with rational numbers. *Cognition and Instruction, 13(1),* 3–50.

Streiner, D. L. (2003) Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.

Appendix A—Instructions for Interviewers

Appendix A is not included in this version for the purpose of maintaining test security.

Contact Robert Schoen (rschoen@lsi.fsu.edu) with requests for access to a version of the report will full information.

Appendix B – Grade 1 Interview Script

Appendix B is not included in this version for the purpose of maintaining test security.

Contact Robert Schoen (rschoen@lsi.fsu.edu) with requests for access to a version of the report will full information.

Appendix C – Grade 2 Interview Script

Appendix C is not included in this version for the purpose of maintaining test security.

Contact Robert Schoen (rschoen@lsi.fsu.edu) with requests for access to a version of the report will full information.

Appendix D—Word Problem Types and Their Respective Abbreviations

(Carpenter et al., 1999)

Problem Type	Abbreviation
Join (result unknown)	JRU
Join (change unknown)	JCU
Join (start unknown)	JSU
Separate (result unknown)	SRU
Separate (change unknown)	SCU
Separate (start unknown)	SSU
Part-part-whole (whole unknown)	PWU
Part-part-whole (part unknown)	PPU
Compare (difference unknown)	CDU
Compare (compare quantity unknown)	CQU
Compare (referent unknown)	CRU
Multiplication grouping	MG
Measurement division	MD

Appendix E—Distributions of Number of Items Answered Correctly Within Each Factor

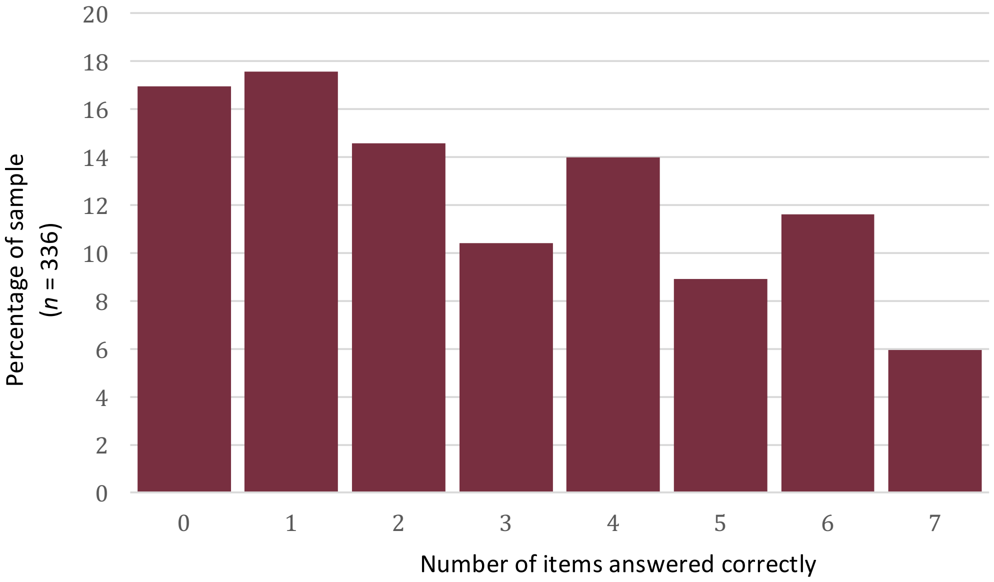


Figure 12. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Computation factor.

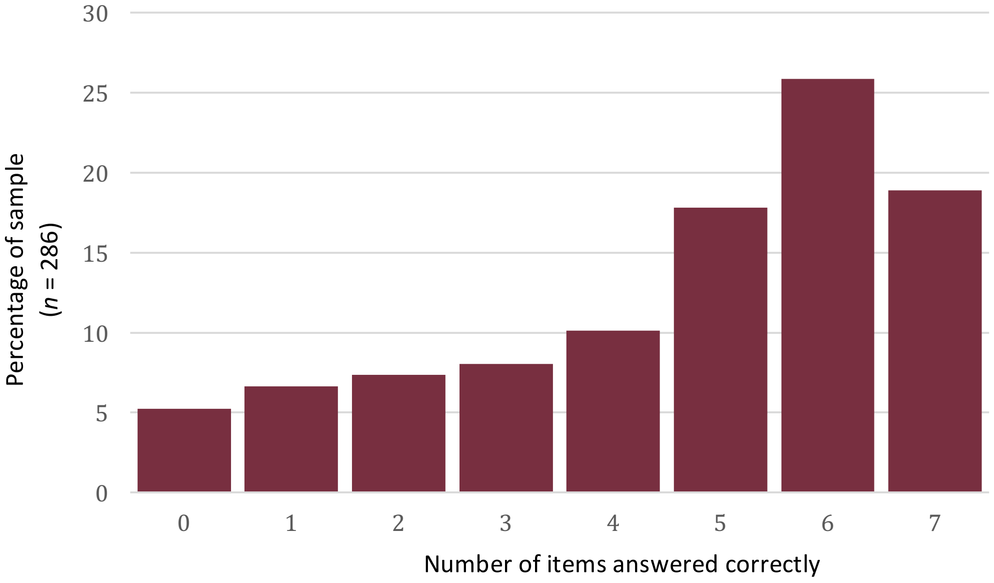


Figure 13. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Computation factor.

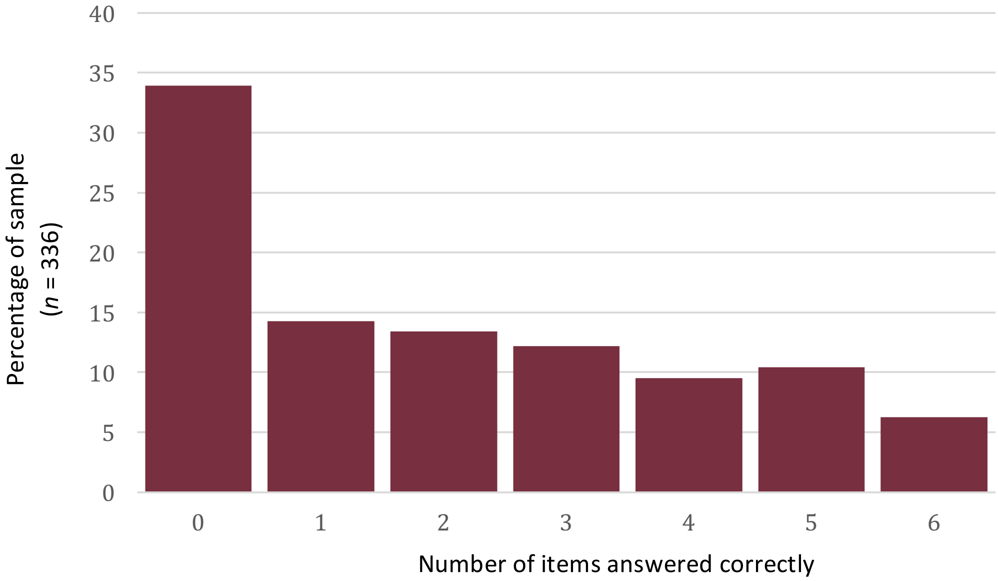


Figure 14. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Word Problems factor.

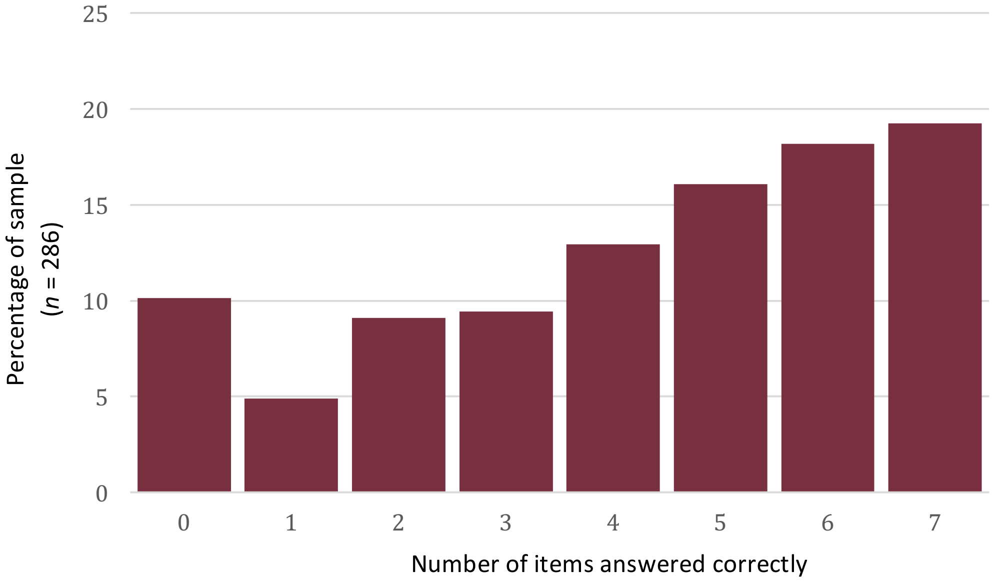


Figure 15. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Word Problems factor.

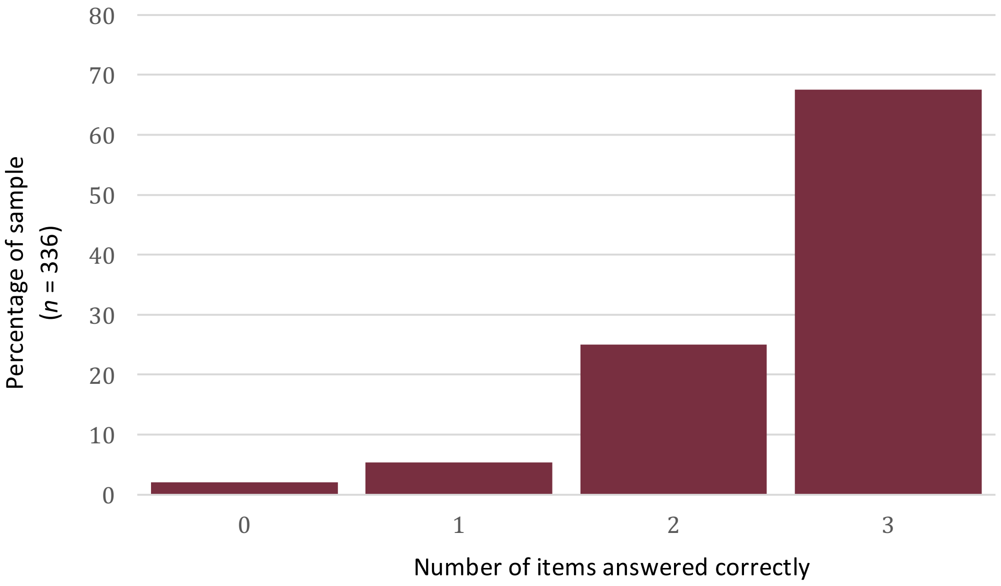


Figure 16. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Number Facts factor.

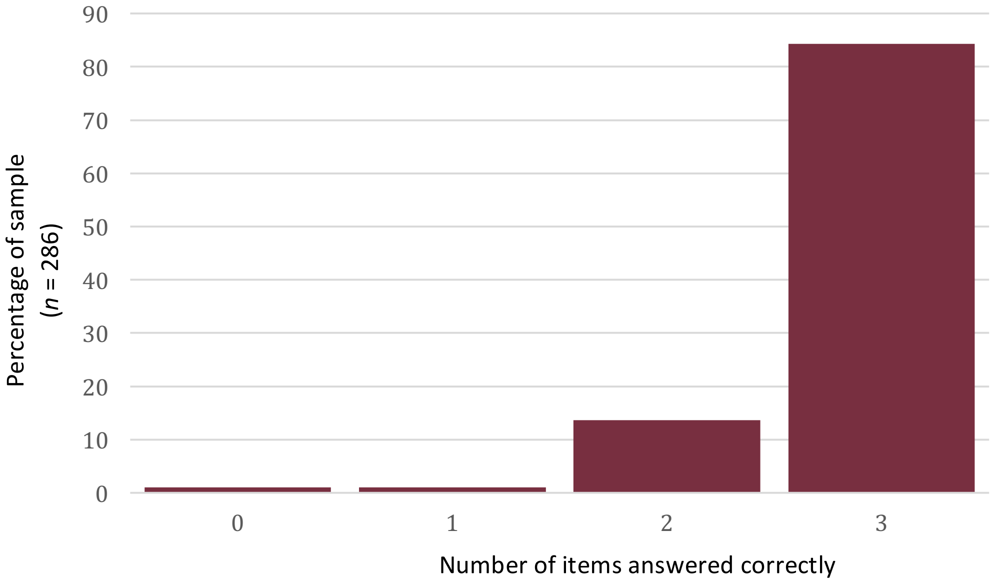


Figure 17. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Number Facts factor.

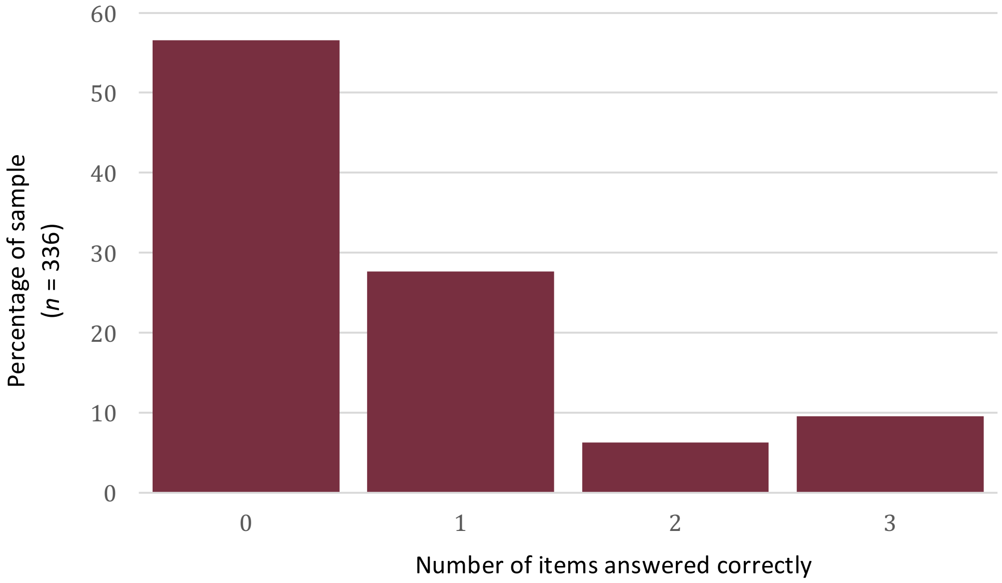


Figure 18. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Operations on Both Sides of the Equal Sign factor.

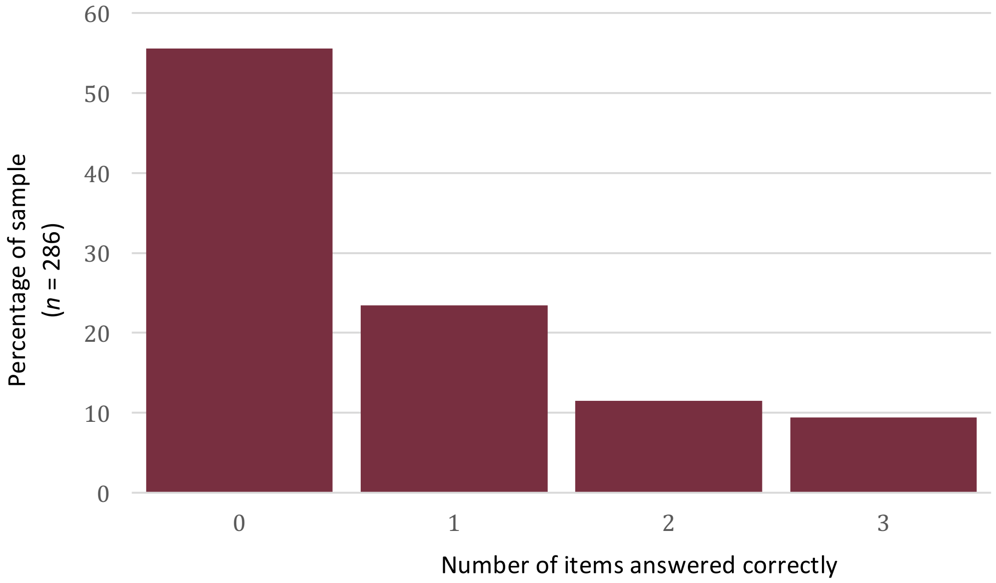


Figure 19. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Operations on Both Sides of the Equal Sign factor.

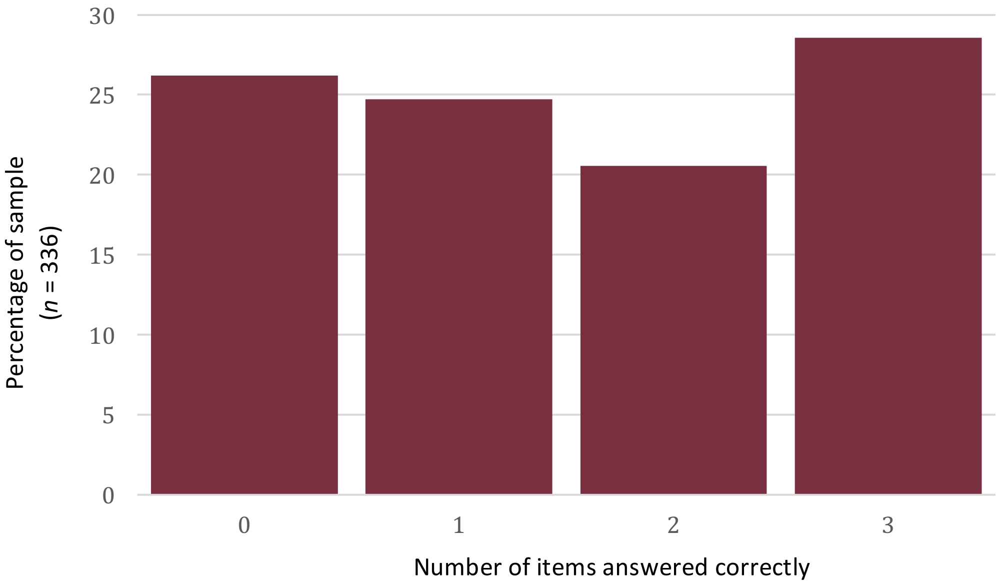
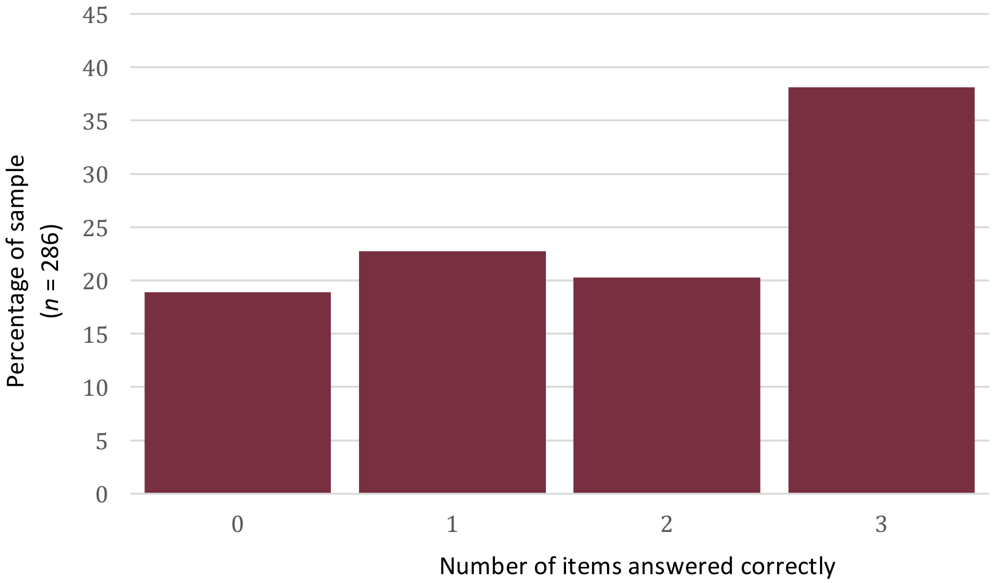


Figure 20. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Equal Sign as a Relational Symbol factor.



Appendix F—Most Common Student Responses by Item

Appendix F is not included in this version for the purpose of maintaining test security.

Contact Robert Schoen (rschoen@lsi.fsu.edu) with requests for access to a version of the report will full information.

Appendix G—A Selection of Additional Readings Relevant to this Report

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction of children's understanding of the "equals" sign. *The Elementary School Journal, 84*(2), 198–212.

Behr, M. (1976). *How children view equality sentences.* PMDC Technical Report No. 3.

Berglund-Gray, G., & Young, R. V. (1940). The effect of process sequence on the interpretation of two-step problems in arithmetic. *Journal of Educational Research, 34*(1), 21–29.

Bergeron, J. C., & Herscovics, N. (1990). Psychological aspects of learning early arithmetic. *Mathematics and Cognition*, 31–52.

Blanton, M., Stephens, A., Knuth, E., Gardiner, A. M., Isler, I., & Kim, J. S. (2015). The development of children's algebraic thinking: The impact of a comprehensive early algebra intervention in third grade. *Journal for Research in Mathematics Education*, *46*(1), 39–87.

Carpenter, T. P. (1985). Learning to add and subtract: An exercise in problem solving. In E. A. Silver (Ed.), *Teaching and learning problem solving: Multiple research perspectives* (pp. 17–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, T. P., & Moser, J. M., (1979). *An investigation of the learning of addition and subtraction* (Theoretical paper No. 79). Madison, WI: Wisconsin Research and Development Center for Individualized Schooling.

Carpenter, T. P., Hiebert, J., & Moser, J. M. (1981). Problem structure and first-grade children's initial solution processes for simple addition and subtraction problems. *Journal for Research in Mathematics Education, 12*(1), 27–39.

Carpenter, T. P., Moser, J. M., & Romberg, A. (1982). *Addition and subtraction: A cognitive perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, T. P., & Moser, J. M., (1983). The acquisition of addition and subtraction concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 7–44). New York: Academic Press.

Carpenter, T. P., & Levi, L. (2000). *Developing conceptions of algebraic reasoning in the primary grades.* Research Report.

Carpenter, T. P., Levi, L., Franke, M. L., & Zeringue, J. K. (2005). Algebra in elementary school: Developing relational thinking. *Zentralblatt für Didaktik der Mathematik*, *37*(1), 53–59.

Caldwell, J. H., & Goldin, G. A. (1979). Variables affecting word problem difficulty in elementary school mathematics. *Journal for Research in Mathematics Education, 10*(5), 323–336.

Christou, C., & Philippou, G. (1998). The developmental nature of ability to solve one-step word problems. *Journal for Research in Mathematics Education, 29*(4), 436–442.

De Corte, E., & Verschaffel, L. (1987). The effect of semantic structure on first graders' strategies for solving addition and subtraction word problems. *Journal for Research in Mathematics Education, 18*(5), 363–381.

Fuson, K. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243–275). Reston, VA: National Council of Teachers of Mathematics.

Gibb, E. G. (1956). Children's thinking in the process of subtraction. *Journal of Experimental Education*, *25*(1), 71–80.

Herscovics, N., & Kieran, C. (1980). Constructing meaning for the concept of equation. *Mathematics Teacher*, *73*(8), 572–580.

IBM Corp. (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Jerman, M. E., & Mirman, S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics, 5*(3), 317–362.

Jones, I., & Pratt, D. (2012). A substituting meaning for the equals sign in arithmetic notating tasks. *Journal for Research in Mathematics and Science Education, 43*(1), 2–33.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12*(3), 317–326.

Koehler, J. (2002). Algebraic reasoning in the elementary grades: Developing an understanding of the equal sign as a relational symbol. (Unpublished master's thesis). *University of Wisconsin-Madison, Madison, WI*.

Koehler, J. L. (2004). *Learning to think relationally: Thinking relationally to learn*. (Unpublished doctoral dissertation). University of Wisconsin-Madison, Madison, WI.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics and Science Education, 37*(4), 297–312.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*(4), 361–371.

Light, G. S. (1980). = [equal sign]. *Mathematics in School, 9*(4), 27.

Mann, R. L. (2004). The truth behind the equals sign. *Teaching Children Mathematics, 11*(2), 65–69.

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43*(3), 316–350.

McLean, R. C. (1964). Third-graders and the equal sign: Report of an experience. *Arithmetic Teacher, 11*(1), 27.

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-school students' understanding of the equal sign: The books they read can't help. *Cognition and Instruction, 24*(3), 367–385.

Molina, M., & Ambrose, R. C. (2006). Fostering relational thinking while negotiating the meaning of the equals sign. *Teaching Children Mathematics, 13*(2), 111–117.

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics, 13*(4), 373–394.

Powell, S. (2012). Equations and the equal sign in elementary mathematics textbooks. *Elementary School Journal, 112*(4), 627–648.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment,* 92(6), 544–559.

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*(1), 49–101.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85.

Saenze-Ludlow, A., & Walgamuth, C. (1998). Third graders' interpretations of equality and the equal symbol. *Educational Studies in Mathematics, 35*, 153–187.

Secada, W. G. (1991). Degree of bilingualism and arithmetic problem solving in Hispanic first graders. *Elementary School Journal, 92*(2), 213–231.

Secada, W. G., & Brendefur, J. L. (2000, Fall). CGI student achievement in region VI evaluation findings. *The Newsletter of the Comprehensive Center-Region VI, 5*(2).

Tamburino, J. L. (1980). An analysis of the modeling processes used by kindergarten children in solving simple addition and subtraction story problems. (Unpublished master's thesis). University of Pittsburgh, Pittsburgh, PA.

Van Dooren, W., De Bock, D., & Verschaffel, L. (2010). From addition to multiplication ... and back: The development of students' additive and multiplicative reasoning skills. *Cognition and Instruction, 28*(3), 360–381.

Verschaffel, L., De Corte, E., & Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education, 30*(3), 265–285.

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole numbers concepts and operations. In F. K. Lester, Jr. (Ed.), Second handbook of research on mathematics teaching and learning. Reston, VA: National Council of Teachers of Mathematics.

Wheeler, G. D. (2010). Assessment of college students' understanding of the equals relation: Development and validation of an instrument. (Unpublished doctoral dissertation). Utah State University, Logan, UT.