2016-12

Mathematics Performance and Cognition (MPAC) Interview: Measuring First- and Second-Grade Student Achievement in Number, Operations, and Equality in Spring 2015

Robert C. Schoen, Mark LaVenia, Zachary M. Champagne, Kristy Farina and Amanda M. Tazaz

Mathematics Performance and Cognition (MPAC) Interview

Measuring First- and Second-Grade Student Achievement in Number, Operations, and Equality in Spring 2015

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

DECEMBER 2016

SECURE VERSION

Research Report No. 2016-02

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., & Tazaz, A. M. (2016). *Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015* (Research Report No. 2016–02). Tallahassee, FL: Learning Systems Institute, Florida State University. DOI: 10.17125/fsu.1493238666

Mathematics Performance and Cognition (MPAC) Interview

Measuring First- and Second-Grade Student Achievement in Number, Operations, and Equality in Spring 2015

Research Report No. 2016–02

Robert C. Schoen

Mark LaVenia

Zachary M. Champagne

Kristy Farina

Amanda M. Tazaz

December 2016

(updated August 31, 2017)

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

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. We will name the various players and their roles, starting with the report coauthors.

Robert Schoen was integrally involved with designing and implementing the interview and interviewer training, creating the coding system, determining the organization of the items for the Item Factor Analysis (IFA), 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 worked closely with Robert Schoen on the design of the interview, developed and led the training of the interviewers, conducted video coding of interviews, and assisted with writing of the report. Amanda Tazaz coordinated the interview team and data collection and provided editing and conceptual feedback for 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, and Ian Whitacre.

Special thanks are in order for the contributions of Amanda Tazaz for managing the day-to-day activities during those intense weeks of interviews. The interview team included Charity Bauduin, Wendy Bray, Anne Brown, Zachary Champagne, Kristopher Childs, Rebecca Gault, Nesrin Sahin, Makini Sutherland, Laura Tapp, Harlan Thrailkill, Alex Utecht, Pooja Vaswani, and Ian Whitacre.

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 Shelby McCrackin, 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 to 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 Tables

List of Figures

Executive Summary

The following report describes an assessment instrument called the Mathematics Performance and Cognition (MPAC) interview. As the name implies, the MPAC interview is administered in an interview setting. The 2015 MPAC interview was designed to measure first and second graders' mathematics achievement and cognitive processes in the domain of number, operations, and algebraic thinking.

The MPAC interview was designed to measure two outcomes of interest. It was designed to measure mathematics achievement in number, operations, and equality, and it was also designed to gather information about students' cognitive processes while they solved mathematics problems. The current report focuses on the content of the interview, interview protocol, scoring procedures, and psychometric properties for the achievement focus of the MPAC Interview.

The 2015 MPAC interview was administered to 856 students in spring 2015 in 22 schools located in two school districts in Florida. The school districts were implementing a curriculum based on the Mathematics Florida Standards, which are very similar to the Common Core State Standards for Mathematics.

The MPAC interview described in the current report builds upon a previous version. Although the 2014 MPAC interview had good psychometric properties, the 2015 MPAC interview represented an improvement. The current report focuses on the 2015 MPAC interview.

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.

Purpose

The immediate goal of the MPAC Interview was to measure student achievement and related thinking processes. It was primarily designed for the purpose of evaluating the impact of a teacher professional-development program on student achievement, cognition, and understanding in the domain of number, operations, and algebraic thinking in mathematics. Nonetheless, we expect the interview data and protocol to be usable in other ways to make a broader contribution in various aspects of mathematics education research.

Content

As stated previously, the MPAC Interview focuses on number, operations, and algebraic thinking at the early elementary level. The 2015 MPAC has five major sections: Number Facts, Solving Equations, Word Problems, Equations (True/False), and Multidigit Computation.

The final MPAC Interview items and coding process were the result of the iterative process of development and feedback from a variety of experts, pilot testing with students, and extensive training of interviewers.

The development process for the MPAC involved expert review that verified the alignment of the content of the interview with current research and with fundamentally important ideas in mathematics at the first and second grade level. In general, the MPAC is designed to align with the core content in the number, operations, and algebraic thinking domains in the Common Core State Standards for Mathematics (CCSS-M) at Grades 1 and 2. In a few instances, the content of the MPAC extends beyond the CCSS-M for the given grade level. These exceptions include word problems involving grouping- or ratio-type scenarios (1 item at Grade 1; 2 items at Grade 2), numbers greater than 100 in computational problems at Grade 1 (1 item at Grade 1), and true/false and open number sentence items involving more than 3 quantities at Grade 1.

The MPAC Interview was designed to measure student achievement and thinking on types of problems that tend to be more difficult for students. For example, multidigit subtraction problems involved regrouping (i.e., borrowing) at least once and sometimes involved regrouping across a zero. These types of numbers in subtraction problems are more likely to produce student errors based on limited understanding than are subtraction problems that do not involve regrouping. The problems in other sections also included more complex types and therefore more places for students to make errors. The purpose of the focus on more complex problems was to increase the ability of the MPAC Interview to identify different levels of knowledge and understanding in the area of number, operations, and equality.

Scoring

Analysis of interviewer coding agreement indicated high coding reliability and adherence to the interview protocol. Data corresponding to whether the final answers to MPAC Interview items were determined to be correct or incorrect were fitted to an item factor analysis (IFA) model with a higher-order structure. Five first-order factors—corresponding to the five sections of the interview—were regressed onto a single second-order factor. The second-order factor score was intended to serve as the overall achievement score on the interview. The RMSEA, CFI, and TLI goodness-of-fit statistics indicated that the IFA models for the two grade levels both provided a close fit to the data. The Grade 1 higher-order model fit statistics were $\chi^2(345) = 663.079$, $p < .001$; RMSEA = .05, 90% CI [.04, .05]; CFI = .97; and TLI = .96. The Grade 2 higher-order model fit statistics were $\chi^2(247) = 368.944$, $p < .001$; RMSEA = .03, 90% CI [.03, .04]; CFI = .99; and TLI = .99.

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 data from the MPAC Interview to split the outcome into more granular set of topics does appear to be a defensible approach in some situations.

In an absolute sense, the difficulty level of the grade 2 MPAC is considerably higher than that of the grade 1 MPAC. The difference was intentional and was informed by the findings of 2014 MPAC interview. The MPAC Interview was designed to allow vertical linking between grades 1 and 2, but the grade 1 and grade 2 interviews were not identical.

Although the current report does provide some information about how students' cognitive processes were recorded, the scoring procedures for cognitive processing metrics will be described in a separate report.

Reliability

The reliabilities of the final MPAC scales were determined by means of a composite reliability estimate for the higher-order factor and ordinal forms of Cronbach's alpha (α) for the subscales. The higher-order factor composite reliability was .91 for Grade 1 and .89 for Grade 2.

On the Grade 1 interview, the α estimate for one subscale was below the .7 conventional minimum (.69); ranging between .81 and .96, the other four exceeded the conventional target value of at least .80. Ranging from .71 to .97, the Grade 2 α subscale estimates all exceeded the conventional minimum value of .7, and four exceeded .80. 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 the Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008) standard scores. 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 .75 and .66 in Grade 1 for each ITBS test, respectively; and .73 and .61 in Grade 2 for each ITBS test, respectively. All correlations between the MPAC higher-order factor scores and the ITBS scores were statistically significant with *p*-values less than .001. MPAC subscale correlations with ITBS ranged from .50 to .77 in Grade 1 and from .49 to .75 in Grade 2.

Summary

The validity of the MPAC interview as an instrument for measuring first and second graders' mathematical achievement in number, operations, and equality 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 on other achievement instruments in wide use by states and districts. 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 achievement.

The description of the development process and the results of the field test of the 2015 MPAC interview described in the following report suggest that the MPAC had sufficient reliability and provide evidence in favor of the validity of the use of MPAC as an achievement measure.

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 that allows the interviewers to follow an initial script to introduce each problem and then 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.

1.1. Overview of Interview

The 2015 MPAC interview consists of 34 items in Grade 1 and 35 items in Grade 2. The items are grouped into five categories for purposes of the implementation of the interview: Number Facts, Solving Equations, Word Problems, Equations: True/False, and Multidigit Computation.[1] Table 1 provides a blueprint of the categories and number 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 2015

	Number of Items	
Section	Grade 1	Grade 2
Number Facts (NF)	10	10
Solving Equations (SE)	5	5
Word Problems (WP)	7	7
Equations: True/False (TF)	8	8
Multidigit Computation (MDC)	4	5
Total	*34*	*35*

Approximately 80% of the questions in the Grade 1 and Grade 2 interviews are identical. For the most part, when the questions are not identical, the questions in the Grade 2 interview are similar in nature but involve higher numbers in an attempt to increase the difficulty proportionally with age and to reveal information about how these older students are making sense of operations on multidigit whole numbers. With the exception of the Word Problems section, the questions that are identical are presented in the same order in the two grades. These items were generally sequenced from easier to more difficult within each subsection.

Interviewers were instructed to explain to students at the beginning of the interview that they were conducting the interview because they were interested in how students solve math problems. In the Number Facts, Solving Equations, Word Problems, and Multidigit Computation sections, unless the

[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.

student's thinking process is exceptionally clear, after an answer is provided by the student, the interviewer asks, "How did you get [insert numerical student response]?" The interviewer can make minor modifications to the exact wording and ask a follow up such as, "I think I see what you did, but can you explain to me how you were using the cubes to find out your answer?" During the Equations: True/False section, after the student provides a response, the interviewer asks the follow up question, "What makes this equation true/not true?"

The purpose of the interviewer's follow-up question is not to find out whether students can prove their answers. Rather, the purpose is to make the thinking process they actually used more salient. When the student's response is something like "I did it in my head," the interviewer asks a probing follow-up question such as, "Can you tell me what you did in your head?" If the strategy was readily apparent, and the interviewer has very high confidence in how the student solved the problem, the interviewer might instead say, "I see just how you got that answer," but the interviewer was advised to use that phrase sparingly and only when it was true.

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 do *not* ask "How do you know that is the answer?"

Sometimes a student's explanation of their strategy and what the interviewer observed them do appear to be inconsistent. Unless the interviewer has indisputable, positive evidence to the contrary, the way the student explains that he or she arrived at the answer is accepted as accurate, even when the interviewer retains some doubt whether that is exactly how the answer was generated. In attempt to minimize the instances of revisionist explanations, the tempo of the interview was kept fairly high (but the high tempo did not apply to the period between presentation of the problem to the student 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 in all cases. To avoid introducing bias, interviewers must be very careful to respond in the same way regardless of whether or not the student generated a correct answer. In the 2014 MPAC interviews (Schoen et al., 2016), the most common violation of this rule occurred when students generated incorrect answers. We found that interviewers sometimes offered to read the question again after the student generated an incorrect answer. This practice is to be avoided. The interviewer should not offer to reread the question after the student has clearly indicated an answer. Rather, the answer should be recorded, and the interview should proceed as usual; the interviewer should ask for information about how the student arrived at the answer. A more complete list of the instructions for interviewing is presented in interview protocol provided in Appendix A.

In general, the problems in the Number Facts, Solving Equations, and Word Problems sections of the interview were ordered from easier to more difficult within each section. If a student provided incorrect answers for three successive items within any individual section, the interviewer moves on to the next section. We called this aspect of the interview protocol the *Mercy Rule*. The Mercy Rule is an attempt to avoid causing undue stress to children who are not performing well (and know it). In the Number Facts section, the interviewer skipped to the subtraction section of Number Facts if the mercy rule applied during the addition section. Because the items were generally sequenced from easier to more difficult

within each section, the Mercy Rule is based on an assumption that the student will not correctly solve the later problems after several failed attempts at earlier problems.

Interviewers must ultimately use their own clinical judgment to decide when to terminate a section or an item and move on to the next. The interviewer always has the authority to choose to end an interview because of anxiety exhibited by the student. Any interview that lasted longer than 60 minutes was politely terminated after the current problem was finished. The 2015 MPAC did not allow for the use of the mercy rule in the Equations: True/False or Multidigit Computation sections. These sections are both very short, and items within these sections are intended to assess different facets of knowledge. Therefore, the assumption that the student who fails to answer earlier problems correctly will not be able to solve the later ones is less relevant or valid. Nevertheless, the interviewer could always use his or her own clinical judgment to ensure that no undue stress was caused to the child through the interview process.

The following paragraphs provide a brief overview of the purpose and substance of each of the introduction and five subsequent sections of the interview.

1.1.1. MPAC 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). Interviewers were instructed to use a positive tone and encouraged to ask each student a nonmathematical question to break the ice and encourage students to participate in the conversation. The interviewer explained that the focus of the interview was on *how* students solve mathematics problems and not on judging correctness. The interviewer confirms the student's name and grade level and requests the student's assent to be interviewed and to be video recorded. The student's assent is recorded on the metadata sheet. If the student does not assent to participate in the interview, the interview is politely terminated without prejudice. In the 2015 sample, students with parental consent to participate in the study were included in the randomly selected sample of students to interview. Parental consent to videotaping was a separate question from consent to participate in the study. If the student (or parent) declined to allow the researchers to video record the interview, the declined video consent/assent was recorded on the metadata sheet. In the cases where the parent or student did not consent or assent to videorecording, a second interviewer observed the interview in real time and coded the interview. This enabled a reliability check for coding as well as a follow-up conversation designed to correct or complete any discrepancies in the codes.

1.1.2. MPAC Section 1: Number Facts

This section contains items that were developed to assess fluency with basic addition and subtraction facts. The items in this section were intentionally created to determine which facts students know directly from memory, which are derived from other known facts, and which are solved by means of counting strategies (Carpenter et al., 2015).

Table 2. Items in the Number Facts (NF) Section

Item	Grade 1	Grade 2
NF1		
NF2		
NF3		
NF4		
NF5		
NF6		
NF7		
NF8		
NF9		
NF10		

As Table 2 shows, the items in the Number Facts section were identical and presented in the same sequence in the two grade levels.

This section is placed first in the interview, because we find that students are fairly comfortable solving these number-fact questions, and it serves as a good warm-up for the rest of the interview. It also provides an important chance for the interviewer to establish expectations about how the student can best participate in the interview. As a result, the interviewer must demonstrate an interest in learning how the student was thinking about these problems as he or she solved them during this first section. Establishing these expectations from the outset help the child to understand how we want him or her to participate in the interview.

Students do not have access to tools in the Number Facts section (other than their minds and their fingers). The students are instructed that they can solve these problems mentally or by using their fingers. The four additional tools allowable in the 2015 MPAC (i.e., paper, markers, snap cubes, and ones, tens, and hundreds base-ten blocks) are presented to the student at the end of the Solving Equations section.

1.1.3. MPAC Section 2: Solving Equations

The items in the Solving Equations section are designed to measure a student's understanding of the equal sign and are particular focused on items that have operations on both sides of the equal sign. As Table 3 shows, the items in the Solving Equations section are identical and presented in the same sequence in the two grade levels. In the actual item on the paper seen by the student, each blank is replaced by an open box. The students are instructed to determine the number that goes in the box to make the equation or number sentence correct.

Table 3. Items in the Solving Equations (SE) Section

Item	Grade 1	Grade 2
SE1		
SE2		
SE3		
SE4		
SE5		

The first item is modeled with the Number Facts factor, but it seems to fit best in this section of the interview, because it serves to ease students into the items that follow, some of which are not typical of problems or structures students have seen previously. The other four items in this section provide an opportunity for students to demonstrate whether they think about the equal sign as a relational symbol or as some sort of operator symbol.

For this section, the interviewer reads each equation to the student exactly as it is written. The subtraction symbol is read as "minus," and the addition symbol is read as "plus." The equal sign is read as "equals," and the blank line is read as, "what number."

Students do not yet have access to tools in the Solving Equations section (other than their minds and their fingers). The additional tools are presented to the student at the end of this section and before the start of the next section.

Interviewers code the answer provided by the student and how the student arrived at the answer. The interviewers keep track of the difference between (a) students who determined the solution to the expression to the left of the equal sign and then solved to find the value of the missing number on the right and (b) those who saw the relationship between the numbers on two sides of the equal sign. For example, when solving the item , students who solved using the description in (b) may say, "since the is than the , the missing number must be than the ."

1.1.4. MPAC Section 3: Word Problems

This section contains a set of word problems representing a range of difficulty and three subtypes: (1) standard addition and subtraction, (2) standard multiplication and division (grouping and measurement type problems), and (3) multistep problems. The problems are sequenced from easier to more difficult. Table 4 provides a list of the word problems by showing, for the sake of brief comparison, the type of each problem and the numbers presented in it.

Table 4. Types of Items (and Given Numbers) in the Word Problem (WP) Section

Item	Grade 1	Grade 2
WP1		
WP2		
WP3		
WP4		
WP5		
WP6		
WP7		

Note. See Appendix D for a glossary of the abbreviations used for the problem types.

Only two of the seven problems in the Word Problems section are identical in the Grade 1 and Grade 2 interviews. When the questions are not identical, the questions in the Grade 2 interview are typically similar in nature but involve higher greater numbers in attempt to increase the difficulty proportionally with age and to reveal information about how these older students are making sense of operations on multidigit whole numbers. The problem involving ratios is offered only to Grade 2 students, in attempt to improve the ability of the MPAC to discriminate among those second graders with the highest amounts of knowledge.

All word problems on the MPAC interviews are read aloud to the students and are repeated as often as the student asks. The interviewer must read the word problem in its entirety. Even if the student requests only a portion of the problem to be repeated, the interviewer must repeat the entire problem.

Before beginning this section, students are presented with paper, markers, snap cubes, and ones, tens, and hundreds base-ten blocks. At this time, the interviewer briefly explains and names each tool as it is presented to the student. The students are instructed that they are not required to use any of these tools but may do so if they choose. Students are also reminded that they may still use their fingers or solve the problems mentally if they choose to do so.

1.1.5. Section 4: Equations: True/False

The items in this section are intended to measure a student's understanding of the meaning of the equal sign in mathematics. As Table 5 shows, the items in this section are identical and presented in the same sequence in the two grade levels. The equations in this section are revealed to the student one at a time. For each equation, after the student correctly reads the equation from left to right, the interviewer asks, "Is that equation true or not true?" After the student answers, the interviewer asks, "What makes this equation [true or not true]?"

Table 5. Items in the Equations: True/False (TF) Section

Item	Grade 1	Grade 2
TF1		
TF2		
TF3		
TF4		
TF5		
TF6		
TF7		
TF8		

The first two items in the section are intended to allow students to become familiar with the item format and to serve as a warm-up. For each of the items in this section, the interviewer presents a piece of paper with the equation on it to the student and asks the student to read the equation aloud. The interviewer ensures that the student correctly reads each equation exactly as it is written, from left to right. If the student is unable to read it correctly, as written, after two tries, the interviewer reads it to the student and then asks the student to read it again. The purpose of this procedure is to direct the student's attention to the way the equation is actually written.

In pilot interviews for the 2014 MPAC (Schoen et al., 2016), we found that many students read atypical equations incorrectly. For example, many students read the equation as "" The careful attention to the way students read the equation aloud seems to provide many students the opportunity to self-correct, and we think it is an important step in the process for the sake of instilling confidence that we are measuring student understanding of the equal sign rather than students' reading ability or attention to detail.

1.1.6. MPAC Section 5: Multidigit Computation

The final section of the interview is designed to measure a student's ability to compute sums and differences with higher numbers as well as to provide opportunities to demonstrate whether they think flexibly about subtraction as take away or distance.

As Table 6 shows, the majority of the items in this section are identical. Grade 2 has one more item than the Grade 1 section has, and the final item requires Grade 2 students to cross the century mark rather than just the decade.

Table 6. Items in the Multi-digit Computation Section

Item	Grade 1	Grade 2
MDC1	■	■
MDC2	■	■
MDC3	■	■
MDC4	■	■
MDC5		■

For this section, the interviewer reads each equation to the student exactly as it is written. Just as in the other sections, the subtraction symbol is read as "minus," and the addition symbol is read as "plus." The equal sign is read as "equals," and the interviewer reads the unknown value as "what number." The student copy of the items shows a box in place of the unknown number.

In addition, students have access to paper, markers, snap cubes, and ones, tens, and hundreds base-ten blocks for all but the last problem. Students are instructed that they are not required to use any of these tools, but the tools remain available. For the last problem on each interview, the student is asked to forego the tools and determine the solution using mental strategies. Because the final problem in both grades was intended to assess how the student uses place value and/or relational thinking, we limit the use of manipulatives to encourage these types of thinking.

The selection of numbers in the subtraction items in the computation section deliberately included number combinations that were challenging to students at this grade level, for several reasons. First, the more routine problems generally do not discriminate among students with higher or average levels of knowledge. Second, the challenging items provide more insight into thinking processes.

2. Procedures

2.1. Instrument Development

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

1. Identification of the student learning goals for primary grades in the area of number, operations, and equality according to the CCSS-M (NGACBP & CCSSO, 2010) and find the intersection between those goals and the focus of the Cognitively Guided Instruction professional development program being evaluated
2. Review of literature related to student thinking and assessment in the areas identified in step 1
3. Review of the psychometric properties of the 2014 MPAC interview
4. Development of a test blueprint for each grade level that may allow for vertical scaling across the two forms
5. Development of a first written draft of the interview items and protocol
6. Review of draft interview and protocol by internal members of the evaluation team and several members of the project advisory board
7. Revision of protocol based on feedback
8. Pilot testing of interview protocol and training of interviewers
9. Revision of protocol and development of electronic data entry system

Because the interview was used in spring 2015 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 was reviewed (e.g., Carpenter et al., 1989; Carpenter et al., 1999; Carpenter et al., 2003; Falkner et al., 1999; Jacobs et al., 2007). 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 great care to avoid using problems that were encountered by teachers during the CGI professional development workshops or other ancillary materials. In addition, the workshop leaders and coordinators did not have access to the items on the interview.

The conceptual categories in the MPAC were determined on the basis of a review of scholarly literature related to student thinking in the domains of number, operations, and equality as well as a review of the results from the 2014 MPAC Interview. From these sources, the major categories of Number Facts, Solving Equations, Word Problems, Equations: True/False, and Multidigit Computation were determined to be likely to provide important information about the effect of the CGI Professional Development 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, Susan Empson, 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 grade 1 and grade 2 curriculum. It was designed to be valid for use as a

mathematics achievement measure for students in grade 1 and grade 2 classrooms. Although a couple of items on the grade 1 and grade 2 forms include content below or above grade-level expectations, the topics are consistent with the CCSS-M (NGACBP & CCSSO, 2010), which formed the basis for the accountability system in place in the schools where the field study was conducted.

We learned a tremendous amount from our work on the 2014 MPAC. Several aspects of the 2014 MPAC Interview protocol were revised or deleted for the 2015 MPAC Interview. Some of the most salient changes we recommended to make to the 2014 MPAC Interview (Schoen et al., 2016) are:

1. Rethink some of the equations used in True/False questions to decrease the likelihood that students can provide correct answers based on incorrect reasoning.
2. Remove the Counting section and replace it with a set of questions asking students to give some basic number facts as a strategy to ease the students into the interview.
3. Replace some of the items that were not used in the final measurement model for the Grade 1 interview with some lower-difficulty problems to improve the instrument's ability to discriminate reliably among students at lower levels of knowledge.
4. Replace some of the items that were not used in the final measurement model for the Grade 2 interview with some higher-difficulty problems to improve the instrument's ability to discriminate reliably among students at high levels of knowledge.
5. In general, drop items that were eliminated during the screening and IFA modeling for the 2014 MPAC, but retain the ones in the True/False section that were designed to serve as a warm-up and practice for that type of problem.

Tables 7 and 8 provide an item-by-item accounting of the relation between items on the 2015 MPAC interview and their history with respect to the 2014 MPAC interview. In the following paragraphs, we discuss each of these recommendations and some of the specific changes that were made in response to them.

In the 2014 MPAC interview, we encountered many students who answered true/false questions about equations correctly but provided incorrect reasoning. This phenomenon occurred most commonly on items involving equation structures with less common formats (e.g., $c = b + a$, $a = a$, and $a + b = c + d$) and when those equations were false. For example, in the 2014 MPAC Interview, a student might judge the item [redacted] false either because [redacted] does not equal [redacted] or because "you can't have a minus sign after the equal sign," a situation that introduced complications in scoring. We decided to only use true statements in equations with these types of structures in the 2015 MPAC. This decision guarded against students providing a correct response based on incorrect reasoning. We also made adjustments to the warm-up True/False items that made them clearer to the students, and we introduced a more detailed coding scheme to capture students' responses to the True/False items in the 2015 MPAC interview. See Appendices B and C for more details of that scheme.

The second recommended change addressed some Counting items that were removed from the 2014 MPAC Interview. These items seemed to confuse some students (who were able to count), so we decided they did not serve their purpose as a warm-up set. We also judged these items to be unsatisfactory in providing information about student thinking. We therefore replaced the Counting items with a Number Facts section. The Number Facts items were intended to provide a more familiar warm-up for students and also to allow us to gain insight into student knowledge and relational thinking strategies used by the students.

Table 7. Item by Item Analysis of the 2014 MPAC Interview – First Grade

Item	Factor	Item Description	Kept, Adapted, or Dropped for 2015 MPAC Interview
CNS 1	*		Dropped
CNS 2	COMP		Dropped
CNS 3	COMP		Dropped
CNS 4	COMP		Dropped
CNS 5	COMP		Dropped
WP 6	*		Dropped
WP 7	WP		Kept
WP 8	WP		Kept
WP 9	WP		Dropped
WP 10	WP		Adapted to
WP 11	WP		Kept
WP 12	WP		Adapted to
EC 1	NF		Dropped
EC 2	*		Dropped
EC 3	NF		Dropped
EC 4	COMP		Dropped
EC 5	*		Adapted to
EC 6	COMP		Adapted to
EC 7	COMP		Adapted to
EC 8	*		Adapted to
EC 9	*		Adapted to
EC 10	OBS		Adapted to
EC 11	ESRS		Kept
EC 12	ESRS		Kept
EC 13	ESRS		Adapted to
EC 14	NF		Kept
EC 15	OBS		Kept
EC 16	OBS		Dropped

Note. * Indicates items that were not retained in the final factor analysis model for the 2014 MPAC Interview.

To widen the range of student knowledge the MPAC could address, we followed our own recommendations 3 and 4 (as listed previously within this section). The addition of the Number Fact items lowered the Grade 1 difficulty. To increase the difficulty for Grade 2, we included more difficult word problem items, including a multistep word problem and a word problem involving ratios. We also made slight adjustments to the numbers in the Multidigit Computation section as a strategy to increase the difficulty of the Grade 2 test. For example, we changed the item from the 2014 MPAC to . This change was designed to retain the intent of the item while making the difference slightly more difficult to determine.

Table 8. Item by Item Analysis of the 2014 MPAC Interview – Second Grade

Item	Factor	Item Description	Kept, Adapted, or Dropped for 2015 MPAC Interview
CNS 1	COMP		Dropped
CNS 2	COMP		Dropped
CNS 3	COMP		Dropped
CNS 4	COMP		Dropped
CNS 5	COMP		Dropped
WP 6	*		Dropped
WP 7	WP		Kept
WP 8	WP		Adapted to
WP 9	WP		Dropped
WP 10	WP		Adapted to
WP 11	WP		Adapted to
WP 12	WP		Dropped
EC 1	NF		Dropped
EC 2	*		Adapted to
EC 3	NF		Adapted to
EC 4	*		Dropped
EC 5	*		Adapted to
EC 6	COMP		Adapted to
EC 7	COMP		Adapted to
EC 8	*		Adapted to
EC 9	*		Adapted to
EC 10	OBS		Adapted to
EC 11	ESRS		Kept
EC 12	ESRS		Kept
EC 13	ESRS		Adapted to
EC 14	NF		Kept
EC 15	OBS		Kept
EC 16	OBS		Dropped

Note. * Indicates items that were not retained in the final factor analysis model for the 2014 MPAC Interview.

Last, the large number of items that are common to the Grade 1 and Grade 2 MPAC forms allows for the vertical scaling of the two forms, opening the possibility for analyses that pool across grade level. A promising technique for this pooling is the Bayesian measurement invariance modeling described by Muthén and Asparouhov (2013). 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.

2.2. Interviewer Training

Gathering data for a semistructured interview in a way that allows fair comparison among interviewees requires considerable skill and coordination on the part of the interviewers.

Many of the personnel involved in interviewing were faculty or graduate students in mathematics education or elementary education. Others included project staff, university faculty members, and former elementary and middle-school teachers. All interviewers had some experience in 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. The total number of interviewers on the 2015 MPAC interview team was 14. Thirteen individuals completed the following training procedures and conducted interviews in spring 2015. One interviewer who had been a fully trained interviewer for the 2014 MPAC Interview started working on the interview team toward the end of the data-collection window. Project staff conducted four hours of training to refresh and update her understanding of the interview protocol.

2.2.1. Phase one of interviewer training

The training procedures for the interviewers consisted of three phases. The first phase involved two six-hour days of classroom-style orientation and introduction to the interview and related research on student thinking. This phase also included a discussion and guidelines for how the interviewers were expected to behave in schools. See Appendix A for the specific Interview Guiding Principles the interviewers were expected to follow.

The first day of training included a discussion of general principles of interviewing children, including guidelines for behaviors. Several ideas from the chapter "Guidelines for Clinical Interviews" from *Entering the Child's Mind: The Clinical Interview in Psychological Research and Practice* (Ginsburg, 1997) were used to frame the discussion.

Each interviewer received a copy of *Children's Mathematics: Cognitively Guided Instruction* (Carpenter et al., 1999) and was assigned to read chapters on how students solve addition, subtraction, multiplication, and division problems involving single- and multidigit numbers. The interviewers were also provided with copies of the Grade 1 and Grade 2 interview protocols for review. We discussed each item, the coding scheme for student strategies, and other guidelines for implementing the protocol. We used a Grade 2 pilot interview that was previously conducted by the project staff to practice coding together as a team. After coding individually, we reviewed each of the codes as a team and worked on finding consensus about what the student did.

The second day of Phase One included learning how to use Microsoft SharePoint, our data-entry system, and how to operate the video camera to capture good-quality video footage. The team members observed a second video of an interview of a grade 1 student and coded the student's responses individually. The team members then practiced entering the data into SharePoint and reviewed the coding decisions with the goal of maintaining high fidelity to the coding protocol and high consistency among interviewers.

2.2.2. Phase two 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 coding scheme used during the interview.

Data collected during the two days of pilot interviews were not used for data analysis. In the first wave of pilot interviews, one of the more experienced interviewers conducted the interview while the less experienced interviewers observed. Subsequent waves of 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 Two 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.

Perhaps most importantly, Phase Two allowed for two interviewers to discuss the proper coding options for the student responses that were provided. Feedback from the interviewers confirmed that this experience was very helpful in understanding how to code student responses and undoubtedly helped foster consistency among our interviewers after formal interviewing began.

2.2.3. Phase three of interviewer training

The third phase of training occurred throughout the period of actual data collection. During the first week of 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 at the conclusion of the interview, after the student was returned to the classroom, 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.

Throughout the interview process, the video recordings of a stratified sample of videos were coded by the project principal investigator and another member of the project staff. The data that these two individuals coded for those interviews as well as written feedback concerning the observed adherence to the interview protocol were sent to each of the interviewers during this period.

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 the conclusion of the student interviews, the video recordings of the interviews were coded from May 2015 through August 2015. 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 *invalid item*, which indicated that the interview strayed from the protocol in a way that invalidated the item. Percentage agreement between video coders and interviewers was calculated, and those results and the rate of incidence of items flagged as invalid are available in the Results section of this report.

2.3. Coding Scheme

The interview was designed to be coded in real time by the interviewer. Data categories include the given answer as well as descriptive codes for the observed strategies. The full interview was pilot tested with 35 students who did not attend schools included in the analytic sample for the efficacy study. These pilot tests resulted in several rounds of 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 of the data-recording and coding system were also further refined during this pilot testing with input from the interviewers.

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.

Observed strategies included, for example, named strategies such as join all, separate from, incrementing, compensation, and standard algorithm. A more detailed description of each strategy and its substrategies appears in the following section. Although the body of literature surrounding many of these strategies defines them as resulting in correct solutions, we encountered many students who attempted to use them in the pilot-testing phase and generated incorrect answers. As a result, strategies are coded on the basis of the strategy used by the student regardless of whether the answer was correct.

For the Number Fact section and first item in the Solving Equations section on the interview, we collected data on:

- The answer the student provided
- The major strategy used by the student (Counting, Derived Fact, Recalled Fact)
- Selected substrategies (where applicable)
- The use of fingers when determining the sum or difference
- Whether an additive or subtractive strategy was used (where applicable)

For the remaining items in the Solving Equations section, we collected data on:

- The answer the student provided
- The explanation provided for how the student arrived at the given answer.

For the Equations: True/False section, we collected data on:

- The student's response for each equation
- How the student decided on that answer (common responses were included for each item and are presented in Appendices B and C)

For the items in the Word Problems and Multidigit Computation sections, we collected data on:

[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, so 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.

- The answer the student provided
- The major strategy used by the student (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)

For a complete list of strategy and substrategy descriptions, see Appendix E.

2.4. Digression from Protocol

Each interviewer was expected to adhere to the script and interviewer guidelines (Appendix A) at all times. The video coders were instructed to flag items on which they felt interviewers digressed from the script dramatically enough to affect the student's response, either positively or negatively, we coded those digressions from protocol as *invalid item*. These digressions were infrequent, but they did occur. When they did, the data for that item was recoded as missing. Below are two examples of the more common digressions from the protocol:

1. In the Equations: True/False section, if students read the equation in a manner that was not exactly as it was written and the interviewer failed to prompt the student to reread the equation, we considered it a digression from the protocol. For example, if the student read the equation [redacted] as [redacted], and the interviewer did not prompt the student to reread it as it is written, we coded the item as invalid.
2. In instances when an interviewer read a number or operation symbol incorrectly, we coded the item as invalid.

Out of the 210 video-coded interviews, each including approximately 35 items, six items were coded as invalid for digressions from protocol, an incidence rate of less than one item per 1,000.

3. Data Analysis

3.1. Description of the Sample

The students who completed the 2015 MPAC Interview were selected through a stratified random sampling procedure from a larger sample composed of 3,681 students (1,933 Grade 1 and 1,748 Grade 2) for whom signed parental consent was obtained. The larger student sample came from 22 schools in two diverse public school districts (7 schools in one district; 15 in the other) in Florida. Grade 1 and grade 2 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 four measurement instruments as part of their participation in the study: a whole-group-administered, written pretest at the beginning of the 2014–2015 school year (EMSA; Schoen et al., 2016); the Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008) Math Problems and Math Computation tests, also administered in a whole-class setting at the end of the 2014–2015 school year; and a student interview (2015 MPAC), which was administered in an individual, one-on-one setting at the end of the 2014–2015 school year.

Table 9 reports the sample sizes for each of the measurement instruments. Table 10 reports the demographics for the sample of participating students.

Table 9. 2015 Student Sample Size per Measurement Instrument

	Sample size		
Measure	Grade 1	Grade 2	Total
Pretest	1,597	1,486	3,083
ITBS Math Problems test	1,599	1,491	3,090
ITBS Math Computation test	1,571	1,482	3,053
MPAC Student Interview	440	416	856

Note. ITBS = Iowa Test of Basic Skills

Table 10. Student Sample Demographics

Characteristic	Total student sample (*n* = 3,681)		Student interview sample (*n* = 856)	
	Proportion	*n*	Proportion	*n*
Gender				
Male	.46	1,694	.50	424
Female	.49	1,790	.50	432
Missing	.05	197	.00	0
Grade				
1	.53	1,933	.51	440
2	.47	1,748	.49	416
Race/Ethnicity				
Asian	.04	146	.06	51
Black	.15	562	.17	144
White	.31	1,126	.33	281
Other	.03	93	.03	21
Hispanic	.30	1,112	.32	276
Missing	.17	642	.10	83
English Language Learners	.16	579	.17	142
Eligible for free or reduced-price lunch	.52	1,902	.56	478
Exceptionality				
Students with disabilities	.06	231	.06	52
Gifted	.03	97	.03	28
Missing	.17	642	.10	83

Note. Proportion provided reflects percentage of total sample. Some characteristic categories are not mutually exclusive. Students with unreported demographic information are represented in the "Missing" category. The Asian, Black, and White categories are non-Hispanic.

3.2. Sampling Procedure

Interviews were conducted with a stratified random sample of up to four students from each participating teacher's classroom. In an attempt to maintain a balanced sample within each classroom with respect to student gender, the first stratum was gender. Student gender data were provided by the school districts. The goal was to have 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, where available. 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 on the day of the interview or did not assent to be interviewed. Although all four strata were represented for almost every classroom in the sample, some classrooms did not have an alternate student for every stratum or even a primary for every stratum. In the event that not enough students fell into each stratum to allow both a primary and an alternate selection, the remaining students in the classroom were sampled at random. For example, if a classroom did not include a girl who scored below the whole-class median pretest achievement level, a primary and an alternate girl were chosen at random from the participating students who did not have a pretest score on record.

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 Percentage Agreement

A total of 856 student interviews were conducted for the purpose of data collection in spring 2015. Interviewers coded data for all interviews they conducted and submitted their data to a Microsoft SharePoint site.

A stratified random sample was selected to be video coded for investigation of interrater agreement. So that the sample would be representative, the first stratum selected was interviewer. The second stratum was whether the interview was conducted during the first week of Phase 3 of interview training (when interviews were conducted in pairs) or after it. Recorded interviews were divided into categories on the basis of primary interviewer, stage-one interviews (conducted in pairs), and stage-two interviews (conducted individually). A random number was assigned to each interview, and the data were sorted on interviewer, stage, and random number. At least 20% of the videos conducted by each interviewer in each stage were selected for review. The data for 210 student interviews were coded by trained interviewers using video recording, all of which were used in the comparisons between video coder and interviewer. In 23 of these cases, the interviews were video-coded by two different people so that agreement among video coders could also be assessed.

Interrater agreement was calculated as the total number of matching values divided by the total number of instances for each data type (e.g., correct, strategy, additive/subtractive). Exact agreement between video coders across all codes was 89% for Grade 1 and 90% for Grade 2, 2% and 5% higher, respectively, than the overall interviewer-video coder agreement. Video coders had advantages over interviewers that improved their accuracy, 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 cases where an interview was video-coded, the video-coded data therefore replaced the interviewer-coded data.

Tables 11 and 12 report the interrater agreement on groups of items. Tables 13 and 14 report the interrater agreement on individual items. The interrater agreement proportions reported here represent agreement between video-coded data and interviewer-coded data. The achievement-score data depend only on the Response Correct evaluation, which had an interrater agreement of greater than 99%. Because data from coders with low interrater agreement were replaced by video-coded data,

the proportions of interrater agreement reported in Tables 11 through 14 are a conservative estimate of the accuracy of the final student interview data.

Table 11. Grade 1 Interrater Agreement by Data Type

	Type of comparison	
Type of agreement	Video–interviewer (*n* = 104)	Video–video (*n* = 14)
Response code	.97	.99
Response correct	>.99	>.99
Major strategy	.86	.85
Substrategy	.79	.79
Additive or subtractive	.84	.90
All categories/items	.87	.89

Table 12. Grade 2 Interrater Agreement by Data Type

	Type of comparison	
Type of agreement	Video–interviewer (*n* = 106)	Video–video (*n* = 9)
Response code	.98	.97
Response correct	.99	.99
Major strategy	.86	.87
Substrategy	.77	.80
Additive or subtractive	.86	.92
All categories/items	.85	.90

Table 13. Grade 1 Video Coder-to-Interviewer Interrater Agreement by Data Type, Split by Item

Item	Description	Response code	Response correct	Major strategy	Substrategy	Additive or subtractive
Number Facts (NF)						
NF1		1.00	1.00	.82	.64	
NF2		1.00	1.00	.82	.67	
NF3		1.00	1.00	.85	.77	
NF4		1.00	1.00	.85	.66	
NF5		.98	.98	.89	.78	
NF6		.98	.98	.86	.84	.87
NF7		1.00	1.00	.85	.76	.86
NF8		.98	.99	.88	.77	.78
NF9		.96	.99	.80	.63	.69
NF10		.99	1.00	.88	.82	.88
Solving Equations (SE)						
SE1		.99	.99	.84	.62	.89
SE2		.97	1.00	.94		
SE3		1.00	1.00	.90		
SE4		.92	.99	.92		
SE5		.86	.99	.95		
Word Problems (WP)						
WP1		.99	.99	.93	.86	.92
WP2		.99	1.00	.85	.72	
WP3		.99	1.00	.86	.66	.82
WP4		.94	1.00	.76	.71	
WP5		.92	1.00	.94	.79	.88
WP6		.89	.99	.89	.81	
WP7		.88	1.00	.92		
Equations: True/False (TF)						
TF1		1.00	1.00	.93		
TF2		1.00	1.00	.72		
TF3		.98	.98	.86		
TF4		.98	.99	.78		
TF5		.98	.99	.81		
TF6		.98	1.00	.95		
TF7		.97	.99	.81		
TF8		.97	.99	.84		
Multidigit Computation (MDC)						
MDC1		.97	.98	.84	.73	.86
MDC2		.96	1.00	.85	.73	.87
MDC3		.94	1.00	.92	.92	
MDC4		.97	1.00	.88	.77	

Note. $N = 104$. These percentages reflect agreement on all codes recorded, including codes for skipped items. Substrategy and additive/subtractive data are only available for some items.

Table 14. Grade 2 Video Code-to-Interviewer Interrater Agreement by Data Type, Split by Item

Item	Description	Response code	Response correct	Major strategy	Substrategy	Additive or subtractive
Number Facts (NF)						
NF1		.99	.99	.77	.61	
NF2		.98	.98	.93	.83	
NF3		.99	.99	.91	.85	
NF4		1.00	1.00	.91	.80	
NF5		1.00	1.00	.89	.85	
NF6		1.00	1.00	.85	.79	.78
NF7		.99	.99	.93	.87	.89
NF8		.98	.98	.94	.82	.86
N 9		.98	.98	.95	.71	.79
NF10		.96	.98	.93	.85	.91
Solving Equations (SE)						
SE1		.98	.99	.89	.67	.90
SE2		.97	.99	.99	.99	
SE3		.95	1.00	.93	.93	
SE4		.96	.99	.92	.92	
SE5		.92	1.00	.90	.90	
Word Problems (WP)						
WP1		.98	.98	.89	.77	.87
WP2		1.00	1.00	.86	.70	.81
WP3		.97	.99	.81	.61	
WP4		.96	.99	.69	.59	
WP5		.94	1.00	.81	.65	.85
WP6		.94	1.00	.91		
WP7		.92	.98	.89		
Equations: True or False (TF)						
TF1		1.00	1.00	.91		
TF2		1.00	1.00	.85		
TF3		.99	.99	.87		
TF4		.99	.99	.79		
TF5		1.00	1.00	.76		
TF6		1.00	1.00	1.00		
TF7		1.00	1.00	.87		
TF8		1.00	1.00	.88		
Multi-Digit Computation (MDC)						
MDC1		.95	.97	.89	.70	.84
MDC2		.95	.96	.90	.71	.83
MDC3		.96	.99	.84	.84	
MDC4		.95	1.00	.89	.60	.84
MDC5		.98	.99	.89	.61	

Note. N = 106. These percentages reflect agreement on all codes recorded, including codes for skipped 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 α, Revelle's β, and McDonald's 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 that is 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.

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.

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 only to retain items in the final model 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 by 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), and the equation for the composite reliability for the second-order Math factor is

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

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 et al. (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 was 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 here. 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 standard scores from the 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 ITBS scores. 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.

Sample sizes for correlations varied across measures. The Grade 1 sample sizes were MPAC *n* = 440, ITBS Math Problems *n* = 1,599, ITBS Math Computation *n* = 1,571, MPAC with ITBS Math Problems correlation *n* = 412, MPAC with ITBS Math Computation correlation *n* = 407, and ITBS Math Problems with ITBS Math Computation correlation *n* = 1,570. The Grade 2 sample sizes were MPAC *n* = 416, ITBS Math Problems *n* = 1,491, ITBS Math Computation *n* = 1,482, MPAC with ITBS Math Problems correlation *n* = 395, MPAC with ITBS Math Computation correlation *n* = 393, and ITBS Math Problems with ITBS Math Computation correlation *n* = 1482.

4. Results

4.1. Five-factor Test Blueprint

Table 15 provides an overview of the number of items in Grade 1 and Grade 2 that remained after the full screening, evaluation, and respecification.

Table 15. Number of Items that Remained on the Spring 2015 MPAC Interview Blueprint After Screening and Respecification

Factor	Grade 1	Grade 2	Common items
Number Facts (NF)	11	6	6
Operations on Both Sides of the Equal Sign (OBS)	5	6	5
Word Problems (WP)	6	6	1
Equal Sign as a Relational Symbol (ESRS)	3	3	3
Computation (COMP)	3	3	1
Total	*28*	*24*	*16*

4.2. Item Screening

Tables 16 and 17 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 inserted a column that names which factor each item belongs to, according to the item blueprint. The items with factor association remained in the final model after screening, evaluation, and respecification, ones without factor associations were dropped from the scale before models were run.

4.2.1. Grade 1 interview item screening

Table 16 reveals that, on the Grade 1 interview, one item (SE5) was slightly above the maximum acceptable value (> 3) for item discrimination. No items fell below the discrimination minimum acceptable value (< 0.4). Eight items (NF1, NF6, TF1, TF2, and TF6) were near or above the maximum acceptable value (>|3|) for item difficulty. The low proportion correct observed for SE5 (.06) may be the reason for the near outlier discrimination estimate, and the high proportions correct observed for NF1 (.98), NF6 (.93), TF1 (.94), TF2 (.97), and TF6 (.92) are consistent with the near outlier estimates of their difficulty parameters.

Table 16. Grade 1 MPAC Interview Item Descriptions, Descriptives, and 2-pl UIRT Parameters

Item	Factor	Item description	Proportion correct	2-pl UIRT parameters	
				Discrimination	Difficulty
Number Facts (NF)					
NF1	NF		.98	1.420	−2.643
NF2	NF		.81	0.865	−1.361
NF3	NF		.83	0.954	−1.382
NF4	NF		.76	0.675	−1.244
NF5	NF		.75	0.997	−0.956
NF6	NF		.93	0.827	−2.438
NF7	NF		.63	0.887	−0.479
NF8	NF		.56	0.664	−0.282
NF9	NF		.48	0.788	0.092
NF10	NF		.56	0.712	−0.259
Solving Equations (SE)					
SE1	NF		.69	0.850	−0.768
SE2	OBS		.15	1.466	1.233
SE3	OBS		.09	1.808	1.519
SE4	OBS		.07	1.532	1.788
SE5	—		.06	3.070	1.615
Word Problems (WP)					
WP1	—		.74	0.557	−1.320
WP2	WP		.32	0.982	0.662
WP3	WP		.41	1.238	0.287
WP4	WP		.33	2.097	0.481
WP5	WP		.35	1.159	0.514
WP6	WP		.26	1.601	0.731
WP7	WP		.15	1.154	1.357
Equations: True/False (TF)					
TF1	—		.94	0.643	−2.913
TF2	—		.97	0.864	−3.050
TF3	ESRS		.68	0.627	−0.883
TF4	ESRS		.53	0.834	−0.110
TF5	ESRS		.52	0.686	−0.066
TF6	—		.92	0.801	−2.327
TF7	OBS		.31	1.079	0.676
TF8	OBS		.33	1.155	0.589
Multidigit Computation (MDC)					
MDC1	COMP		.51	1.088	−0.033
MDC2	COMP		.43	0.549	0.340
MDC3	—		.26	0.524	1.259
MDC4	COMP		.55	1.076	−0.153

Note. N = 440 valid Grade 1 student interviews conducted. 2-pl UIRT, 2-parameter logistic unidimensional item response theory model. Discrimination estimates use a 1.702 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994).

We plotted the discrimination and difficulty parameters to inform our decision about retaining or dropping items. Figure 1 presents the Grade 1 difficulty-versus-discrimination scatterplot. The cluster of items with near outlier difficulty estimates (NF1, NF6, TF1, TF2, and TF6) were determined to pass the item screening, but they were flagged for further scrutiny in subsequent model evaluation. Likewise, the marginally acceptable high-discrimination item (SE5) passed the screening but was flagged for further scrutiny. Moreover, all items for the Grade 1 interview were included in subsequent model evaluation.

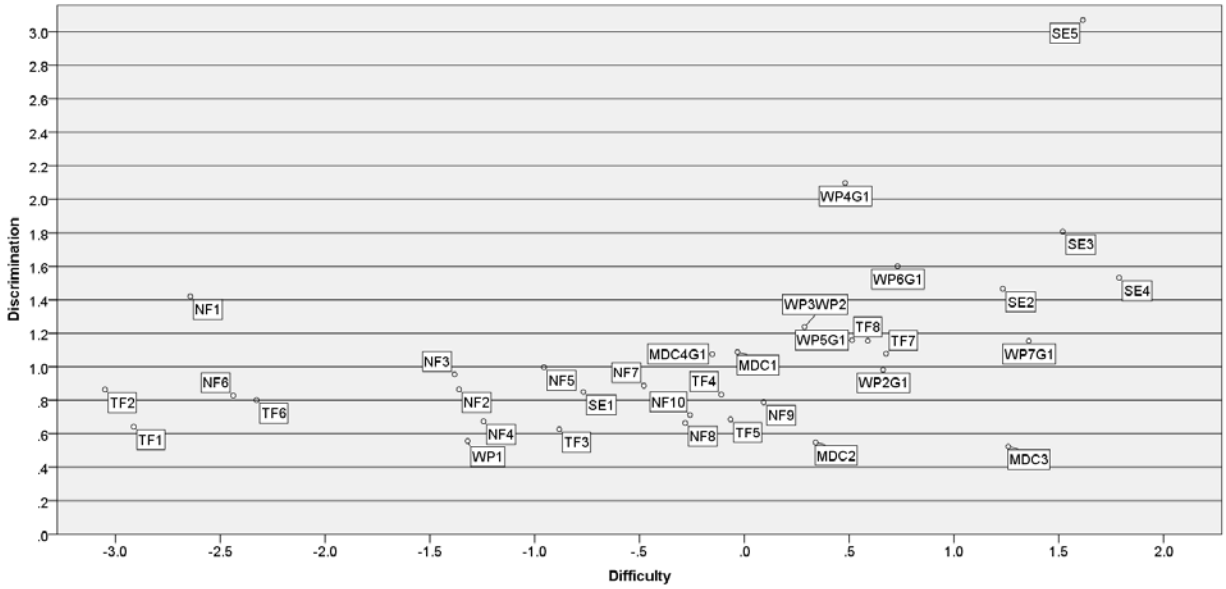


Figure 1. Grade 1 MPAC interview 2-parameter logistic unidimensional item response theory (2-pl UIRT) difficulty-vs-discrimination scatterplot. Items with labels ending in "G1" are unique to the Grade 1 interview.

4.2.2. Grade 2 interview item screening

Table 17 reveals that, on the Grade 2 MPAC interview, seven items (NF1, NF2, NF3, NF6, TF1, TF2, and TF6) had outlier item difficulty estimates (>|3|). High proportions correct were observed for all of these items. Extreme difficulty parameters estimated for items NF1 (−10.6), NF3 (−4.8), NF6 (−4.7), TF1 (−8.5), and TF6 (27.2) resulted in our decision to screen them out from subsequent modeling for the Grade 2 MPAC interview. The other items that exceeded the >|3| criterion for acceptable difficulty estimates (NF2 and TF2) were determined to be marginal enough to pass the item screening, but they were flagged for further scrutiny in subsequent model evaluation. Two other items (NF4 and WP1) were near the maximum acceptable value for item difficulty and were determined to pass the item screening but were flagged for further scrutiny.

Five items (NF1, NF3, TF1, TF6, and MDC2) fell below the discrimination minimum acceptable value (< 0.4), and four of them also had outlier difficulty parameters estimates. Item MDC2 had a discrimination parameter estimate below the minimum acceptable value, but it was retained for subsequent modeling until we could determine whether it should be retained through special consideration to fill a void along the difficulty continuum.

Table 17. Grade 2 MPAC Interview Item Descriptions, Descriptives, and 2-pl UIRT Parameters

Item	Factor	Item description	Proportion correct	2-pl UIRT parameters	
				Discrimination	Difficulty
Number Facts (NF)					
NF1	—		1.00[a]	0.342	−10.642
NF2	—		.94	0.564	−3.267
NF3	—		.94	0.363	−4.785
NF4	—		.93	0.579	−3.026
NF5	NF		.93	0.749	−2.491
NF6	—		.99	0.702	−4.735
NF7	NF		.85	0.614	−1.963
NF8	NF		.75	0.567	−1.339
NF9	NF		.77	0.803	−1.147
NF10	NF		.83	0.791	−1.559
Solving Equations (SE)					
SE1	NF		.87	0.684	−1.982
SE2	OBS		.19	1.869	1.018
SE3	OBS		.14	2.540	1.196
SE4	OBS		.08	2.143	1.564
SE5	OBS		.11	2.266	1.340
Word Problems (WP)					
WP1	—		.84	0.421	−2.501
WP2	WP		.69	0.669	−0.845
WP3	WP		.38	0.928	0.441
WP4	WP		.42	1.016	0.274
WP5	WP		.41	0.792	0.361
WP6	WP		.31	0.852	0.783
WP7	WP		.11	0.775	2.045
Equations: True/False 9TF)					
TF1	—		.99	0.315	−8.484
TF2	—		.99	0.795	−3.929
TF3	ESRS		.80	0.730	−1.409
TF4	ESRS		.65	0.832	−0.575
TF5	ESRS		.63	0.801	−0.507
TF6	—		.97	0.076	−27.218
TF7	—		.37	1.454	0.392
TF8	—		.42	1.514	0.260
Multidigit Computation (MDC)					
MDC1	COMP		.71	0.764	−0.869
MDC2	—		.64	0.381	−0.928
MDC3	—		.48	0.573	0.119
MDC4	COMP		.24	0.782	1.159
MDC5	COMP		.56	0.522	−0.282

Note. N = 416 valid Grade 2 student interviews conducted. 2-pl UIRT, 2-parameter logistic unidimensional item response theory model. Discrimination estimates use a 1.702 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994).
[a]Rounded to three digits, this figure is .998.

Figure 2 presents the Grade 2 difficulty-versus-discrimination scatterplot with NF1, NF3, NF6, TF1, and TF6 included. Figure 3 presents the same plot for Grade 2 with NF1, NF3, NF6, TF1, and TF6 removed. Figure 3 reveals WP1 to be located in a region on the difficulty continuum where few other items were located and accordingly was given special consideration and retained in the initial correlated trait model for further evaluation. Figure 3 reveals MDC2 to be located in a region on the difficulty continuum where other items were located, so it did not warrant special consideration, though was still used in the initial correlated-trait model for further evaluation.

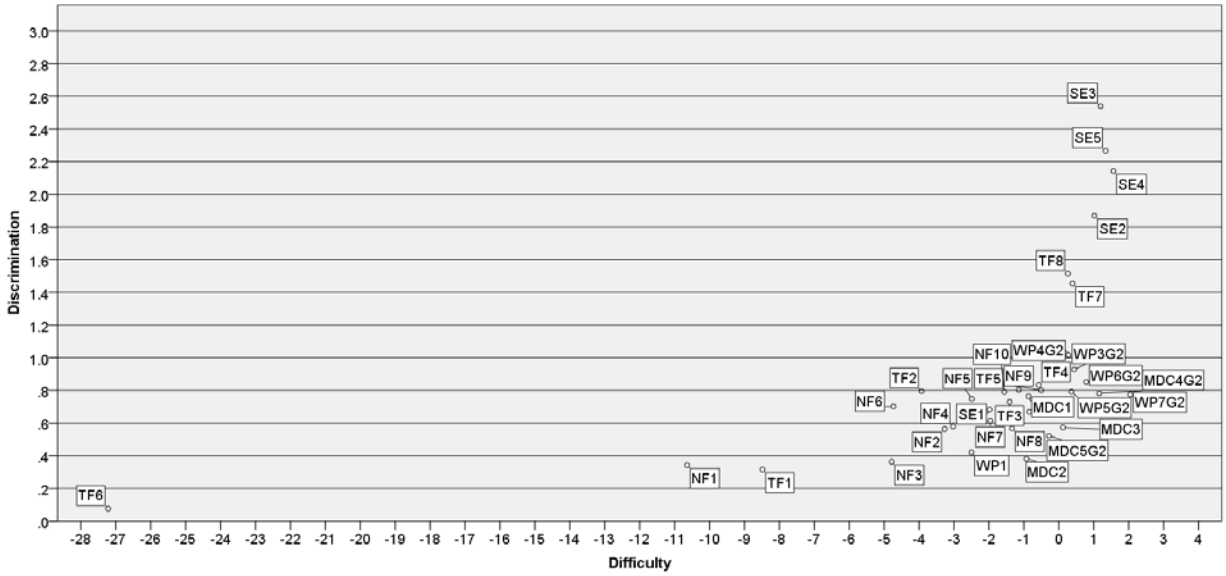


Figure 2. Grade 2 MPAC interview 2-pl UIRT difficulty-vs-discrimination scatterplot (all items). Items with labels ending in "G2" are unique to the Grade 2 interview.

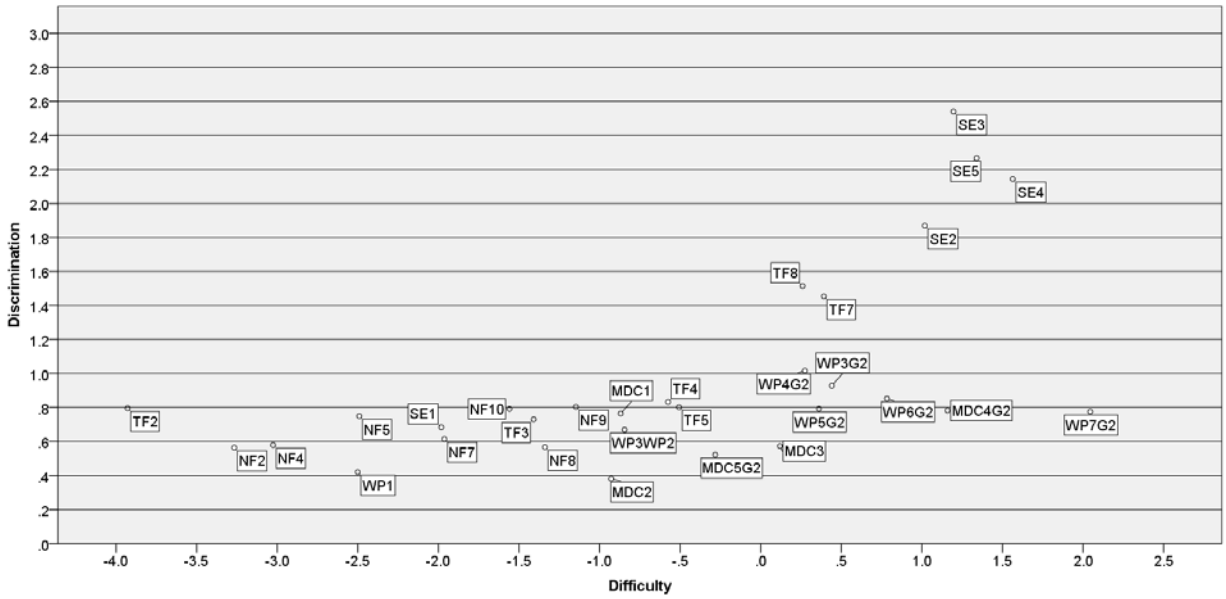


Figure 3. Grade 2 MPAC interview 2-pl UIRT difficulty-vs-discrimination scatterplot (minus outliers). Items with labels ending in "G2" are unique to the Grade 2 interview.

4.3. Correlated Trait Model Evaluation

4.3.1. Grade 1 correlated trait model evaluation

The initial Grade 1 correlated trait model included all items that were administered on the Grade 1 MPAC student interview. All items in the initial model had statistically significant unstandardized factor loadings ($p < .001$). Three items (WP1, MDC2, and MDC3) had standardized factor loadings below or near the minimum acceptable value of 0.5. On inspection of their standardized loadings (.56, .51, and .48, respectively) 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 WP1 and MDC3 could be dropped for the revised model but that MDC2 should be retained. Items TF1, TF2, and TF6 had item parameters that indicated adequate item performance, but their marginal outlier difficulty parameter estimates weighed against them, and we dropped them for the revised model. We also determined that item SE5 should be dropped for the revised model, because of its marginal outlier discrimination parameter estimate and the presence of other items in the same region of difficulty with more reasonable discrimination parameter estimates.

We then fitted the data for the reduced set of Grade 1 items to a revised 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: $\chi^2(340) = 580.478$, $p < .001$; RMSEA = .040, 90% CI [.034, .046]; CFI = .975; and TLI = .972. All unstandardized factor loadings for the revised Grade 1 model were statistically significant. Table 18 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 .50, and most were well above the target of .70.

Table 18. 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 (NF) by					
NF 1		.730	(.119)	.723	(.120)
NF 2		.759	(.051)	.761	(.050)
NF 3		.799	(.049)	.803	(.049)
NF 4		.619	(.057)	.613	(.058)
NF 5		.779	(.041)	.778	(.044)
NF 6		.672	(.083)	.662	(.083)
NF 7		.808	(.037)	.813	(.037)
NF 8		.696	(.041)	.693	(.041)
NF 9		.701	(.043)	.707	(.043)
NF 10		.697	(.043)	.698	(.042)
SE1		.707	(.049)	.702	(.049)
Operations on both sides of the equal sign (OBS) by					
SE2		.940	(.023)	.938	(.028)
SE3		.949	(.035)	.958	(.040)
SE4		.903	(.045)	.909	(.049)
SE5		.989	(.040)	—	—
TF7		.943	(.022)	.952	(.021)
TF8		.954	(.018)	.959	(.018)
Word problems (WP) by					
WP1		.528	(.058)	—	—
WP2		.731	(.042)	.732	(.042)
WP3		.824	(.034)	.833	(.034)
WP4		.943	(.022)	.950	(.022)
WP5		.799	(.035)	.794	(.036)
WP6		.880	(.028)	.880	(.028)
WP7		.798	(.040)	.793	(.039)
Equal sign as a relational symbol (ESRS) by					
TF1		.630	(.077)	—	—
TF2		.716	(.095)	—	—
TF3		.731	(.048)	.739	(.051)
TF4		.874	(.036)	.901	(.033)
TF5		.795	(.039)	.822	(.037)
TF6		.697	(.086)	—	—
Computation (COMP) by					
MDC1		.735	(.040)	.749	(.042)
MDC2		.511	(.053)	.512	(.055)
MDC3		.482	(.058)	—	—
MDC4		.739	(.042)	.740	(.044)

Note. $N = 440$.

Table 19 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 collinearity, but two correlations were notably high: Computation with number facts (r = .87) and Computation with word problems (r = .93). Figure 4 illustrates the correlated factor structure and standardized factor loadings for the revised Grade 1 model.

Table 19. Grade 1 Factor Correlations for the Revised Correlated Trait Model

Factors	NF	OBS	WP	ESRS	COMP
Number Facts (NF)	—				
Operations on Both Sides of the Equal Sign (OBS)	.503	—			
Word Problems (WP)	.707	.678	—		
Equal Sign as a Relational Symbol (ESRS)	.525	.735	.574	—	
Computation (COMP)	.872	.675	.929	.632	—

Note. N = 440.

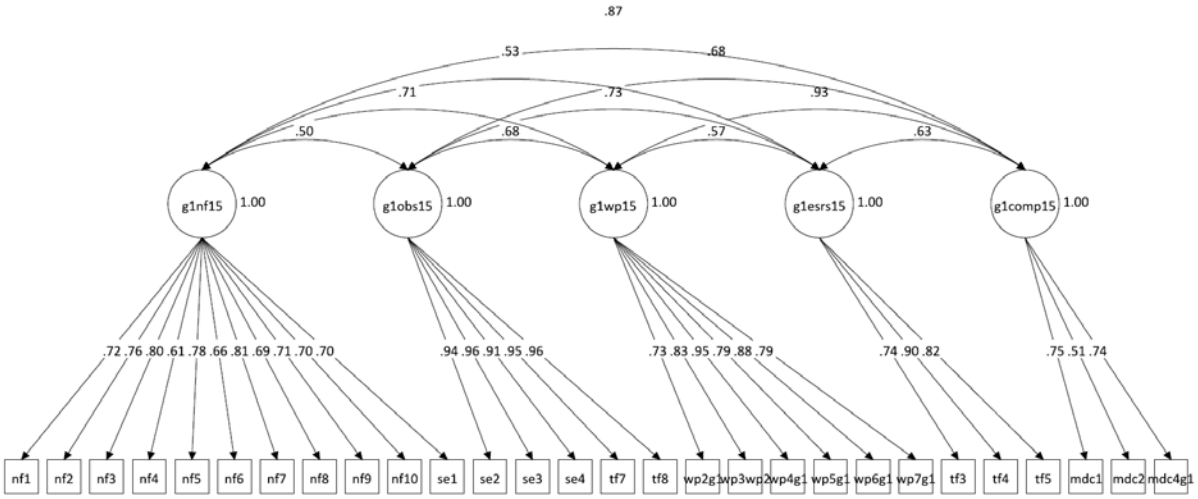


Figure 4. *Grade 1 revised model—correlated trait model diagram with standardized parameter estimates.*

4.3.2. Grade 2 correlated trait model evaluation

The initial Grade 2 model contained all items except NF1, NF3, NF6, TF1, and TF6, which were dropped during the item screening step. All items in the initial model had statistically significant unstandardized factor loading (p < .001). Six items (NF2, NF4, WP1, TF2, MDC2, and MDC5) had standardized factor loadings that were below or near the factor-loading minimum acceptable value of .50. On inspection of their standardized loadings (.55, .54, .41, .44, .47, and .57, respectively) 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 dropped all of these items except MDC5 for the revised model. Items MDC3 had item parameters that indicated adequate item performance, but we determined it should be dropped for the revised model because of concern about whether it performed as intended on the basis of on anomalies in how some students responded to it.

We then fitted the data for the reduced set of Grade 2 items to a revised 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: $\chi^2(242) = 340.221$, $p < .001$; RMSEA = .031, 90% CI [.023, .039]; CFI = .988; and TLI = .987. All unstandardized factor loadings for the revised Grade 2 model were statistically significant. Table 20 presents the standardized factor loadings for the initial and revised correlated-trait models. All standardized factor loadings for the revised Grade 2 model were above the minimum acceptable value of .50, and most were well above the target of .70.

Table 20. 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 (NF) by					
	NF2	.547	(.113)	—	—
	NF4	.537	(.103)	—	—
	NF5	.657	(.096)	.649	(.098)
	NF7	.635	(.069)	.632	(.070)
	NF8	.655	(.056)	.633	(.058)
	NF9	.736	(.058)	.741	(.059)
	NF10	.772	(.058)	.768	(.060)
	SE1	.637	(.073)	.629	(.074)
Operations on Both Sides of the Equal Sign (OBS) by					
	SE2	.952	(.023)	.953	(.023)
	SE3	.941	(.026)	.939	(.026)
	SE4	.930	(.028)	.932	(.027)
	SE5	.976	(.026)	.977	(.026)
	TF7	.949	(.017)	.951	(.017)
	TF8	.964	(.015)	.960	(.015)
Word Problems (WP) by					
	WP1	.410	(.076)	—	—
	WP2	.647	(.054)	.646	(.057)
	WP3	.780	(.041)	.787	(.041)
	WP4	.806	(.040)	.808	(.041)
	WP5	.716	(.044)	.719	(.045)
	WP6	.763	(.042)	.760	(.043)
	WP7	.680	(.064)	.697	(.064)
Equal Sign as a Relational Symbol (ESRS) by					
	TF2	.443	(.120)	—	—
	TF3	.729	(.058)	.705	(.058)
	TF4	.916	(.032)	.915	(.033)
	TF5	.894	(.033)	.902	(.034)
Computation (COMP) by					
	MDC1	.707	(.054)	.720	(.060)
	MDC2	.471	(.062)	—	—
	MDC3	.621	(.055)	—	—
	MDC4	.728	(.055)	.738	(.059)
	MDC5	.572	(.060)	.568	(.063)

Note. $N = 416$.

Table 21 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 collinearity, but one correlation was notably high: Word problems with Computation ($r = .85$). Figure 5 illustrates the correlated factor structure and standardized factor loadings for the revised Grade 2 model.

Table 21. Grade 2 Factor Correlations for the Revised Correlated Trait Model

Factors	NF	OBS	WP	ESRS	COMP
Number Facts (NF)	—				
Operations on Both Sides of the Equal Sign (OBS)	.531	—			
Word Problems (WP)	.699	.636	—		
Equal Sign as a Relational Symbol (ESRS)	.528	.655	.561	—	
Computation (COMP)	.720	.608	.848	.531	—

Note. N = 416.

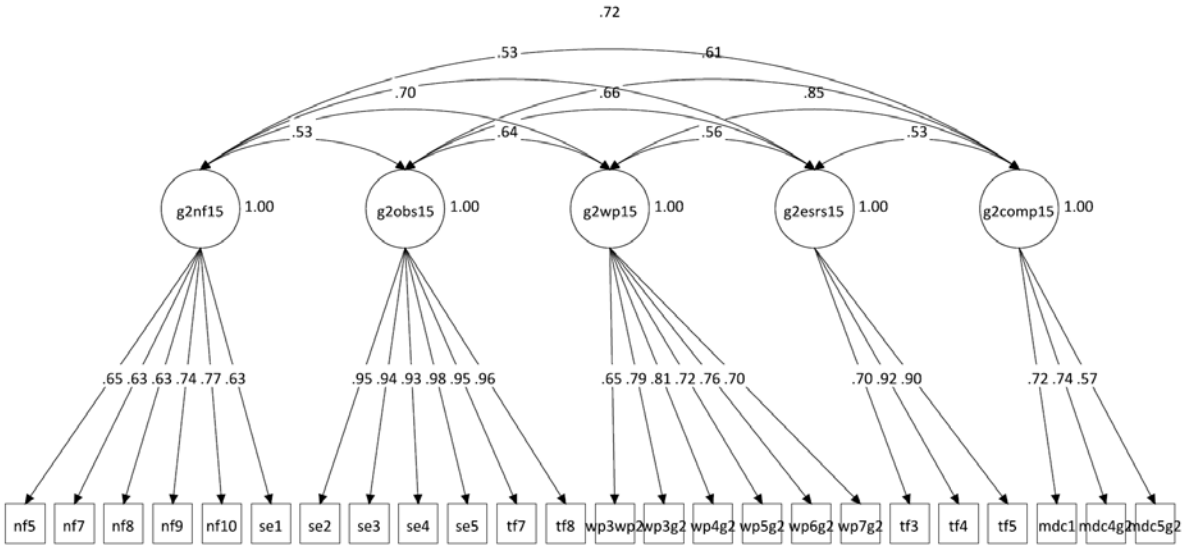


Figure 5. *Grade 2 revised model—correlated trait model diagram with standardized parameter estimates.*

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: $\chi^2(345) = 663.079$, $p < .001$; RMSEA = .046, 90% CI [.041, .051]; CFI = .966; and TLI = .963. 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 for each item. We therefore determined the more parsimonious higher-order structure to be appropriate for modeling the Grade 1 interview data. Table 22 presents the standardized first-order factor loadings for the Grade 1 (and Grade 2) higher-order measurement model. The corresponding second-order factor loadings are presented in Table 23. Figure 6 illustrates the higher-order factor structure and standardized factor loadings for the final Grade 1 model.

The initial fitting of the Grade 1 higher-order model resulted in a linear dependency between the higher-order math factor and the lower-order computation factor, indicated by a standardized loading greater than one (1.02) and negative residual variance (−.04) for the computation factor. To resolve the not positive definite latent variable covariance matrix, we constrained the residual variance for the computation factor to be greater than zero in the final Grade 1 higher-order model. Fit statistics

reported in the paragraph above pertain to the final model that included the constrained computation-factor residual variance.

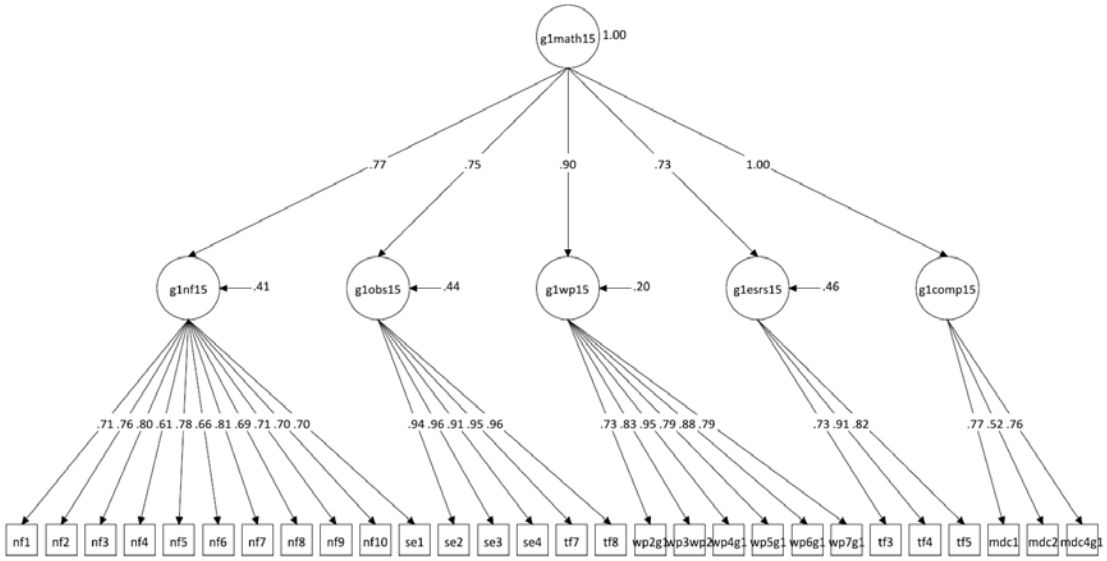


Figure 6. Grade 1 final model—higher-order factor diagram with standardized parameter estimates.

Table 22. 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*)
Number Facts (NF) by					
NF1		.715	(.121)		
NF2		.761	(.050)		
NF3		.802	(.049)		
NF4		.613	(.058)		
NF5		.779	(.045)	.644	(.098)
NF6		.660	(.083)		
NF7		.815	(.037)	.634	(.069)
NF8		.692	(.041)	.633	(.059)
NF9		.708	(.043)	.743	(.059)
NF10		.698	(.043)	.767	(.061)
SE1		.703	(.049)	.629	(.074)
Operations on Both Sides of the Equal Sign (OBS) by					
SE2		.938	(.028)	.952	(.023)
SE3		.956	(.040)	.939	(.026)
SE4		.909	(.049)	.932	(.027)
SE5				.977	(.026)
TF7		.951	(.022)	.951	(.017)
TF8		.960	(.018)	.961	(.015)
Word Problems (WP) by					
WP2_Gr1		.732	(.042)		
WP3/WP2		.833	(.034)	.647	(.058)
WP3_Gr2				.786	(.041)
WP4_Gr1		.950	(.022)		
WP4_Gr2				.808	(.041)
WP5_Gr1		.794	(.036)		
WP5_Gr2				.719	(.045)
WP6_Gr1		.881	(.028)		
WP6_Gr2				.759	(.044)
WP7_Gr1		.792	(.040)		
WP7_Gr2				.697	(.065)
Equal Sign as a Relational Symbol (ESRS) by					
TF3		.730	(.054)	.708	(.058)
TF4		.908	(.035)	.915	(.033)
TF5		.820	(.039)	.901	(.034)
Computation (COMP) by					
MDC1		.765	(.038)	.721	(.061)
MDC2		.515	(.054)		
MDC4_Gr1		.755	(.040)		
MDC4_Gr2				.740	(.060)
MDC5_Gr2				.565	(.064)

Note. Grade 1 n = 440. Grade 2 n = 416.

Table 23. Standardized Second-Order Factor Loadings and First-Order Factor Residual Variances for Grade 1 and Grade 2 Higher-Order Measurement Models

		Grade 1 Interview		Grade 2 Interview	
	Indicator description	Estimate	(*SE*)	Estimate	(*SE*)
Math by					
NF	NF latent variable	.767	(.030)	.760	(.045)
OBS	OBS latent variable	.751	(.037)	.756	(.038)
WP	WP latent variable	.896	(.027)	.864	(.036)
ESRS	ESRS latent variable	.732	(.038)	.736	(.043)
COMP	COMP latent variable	1.000	(.000)	.875	(.054)
Residual variance					
NF		.411	(.046)	.422	(.068)
OBS		.436	(.056)	.428	(.058)
WP		.198	(.048)	.253	(.063)
ESRS		.464	(.056)	.458	(.064)
COMP		.000	(.000)	.232	(.094)

Note. Grade 1 *n* = 440. Grade 2 *n* = 416. NF = Number Facts; OBS = Operations on Both Sides of the Equal Sign; WP = Word Problems; ESRS = Equal Sign as a Relational Symbol; COMP = Computation

4.4.2. Grade 2 higher-order model evaluation

The Grade 2 higher-order model RMSEA, CFI, and TLI indicated close fit: $\chi^2(247) = 368.944$, $p < .001$; RMSEA = .034, 90% CI [.027, .042]; CFI = .985; and TLI = .985. 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. We therefore determined the more parsimonious higher-order structure to be appropriate for modeling the Grade 2 interview data. Table 22 presents the standardized first-order factor loadings for the Grade 2 (and Grade 1) higher-order measurement model. The corresponding second-order factor loadings are presented in Table 23. Figure 7 illustrates the higher-order factor structure and standardized factor loadings for the final Grade 2 model.

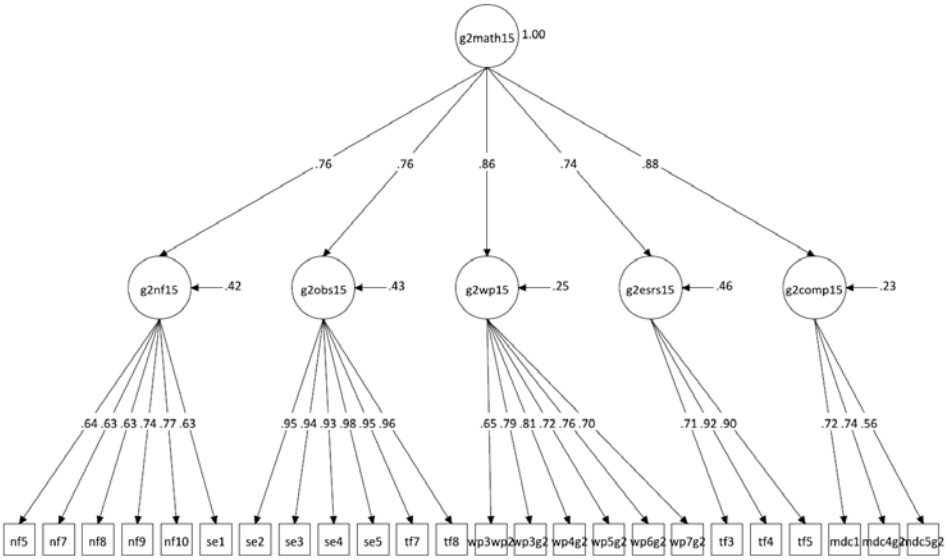


Figure 7. Grade 2 final model—higher-order factor diagram with standardized parameter estimates.

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. Grade 1 higher-order math factor composite reliability was calculated as

$$\frac{(0.539 + 0.721 + 0.710 + 0.601 + 0.755)^2}{(0.539 + 0.721 + 0.710 + 0.601 + 0.755)^2 + (0.203 + 0.402 + 0.124 + 0.312 + 0.000)} = 0.91\,,$$

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 .91, which exceeds the target reliability of .80.

Table 24 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 COMP scale, which had an estimated α reliability of .69. Comparison between the αs and βs revealed a range of discrepancies, some small (such as for the WP scale, where α = .93 and β = .90), some moderate (such as for the OBS scale, where α = .96 and β = .90), and others large (such as for the ESRS scale, where α = .81 and β = .59). 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 (range .02 to .09) for most subscales and large for the NF subscale (.21) and total scale (.15). Where α exceeds ω_h (i.e., Math, WP, ESRS, and COMP), the α to ω_h discrepancies indicate the presence of multidimensionality within the scales. Where ω_h exceeds α (i.e., COMP), variability was present in the general factor loadings but group factor loadings were relatively small, indicating that lumpiness in the scale is 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 in spite of evidence of some within-scale multidimensionality.

Table 24. Grade 1 MPAC Interview Scale Reliability Estimates

Scale	*N* items	Reliability		
		α	β	ω_h
Number Facts (NF)	11	.91	.79	.70
Operations on Both Sides of the Equal Sign (OBS)	5	.96	.90	.87
Word Problems (WP)	6	.93	.90	.87
Equal Sign as a Relational Symbol (ESRS)	3	.81	.59	.79
Computation (COMP)	3	.69	.50	.71
Total (Math)	28	.96	.85	.81

Note. Sample *N* = 440. α, β, and ω_h are ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h, 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 −3.1 through 2.7. Given the sample descriptives (M = −0.005, SD = 0.953, Min = −2.830, and Max = 2.410), reliability of the scale is probably acceptable for approximately 100% of the sample and full range of observed abilities. The information curve exceeds 4 (reliability of .8) for the ability range of approximately −2.4 through 2.3, indicating that target reliability of the scale was achieved for approximately 98% of the sample and nearly the full range of observed abilities.[3]

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

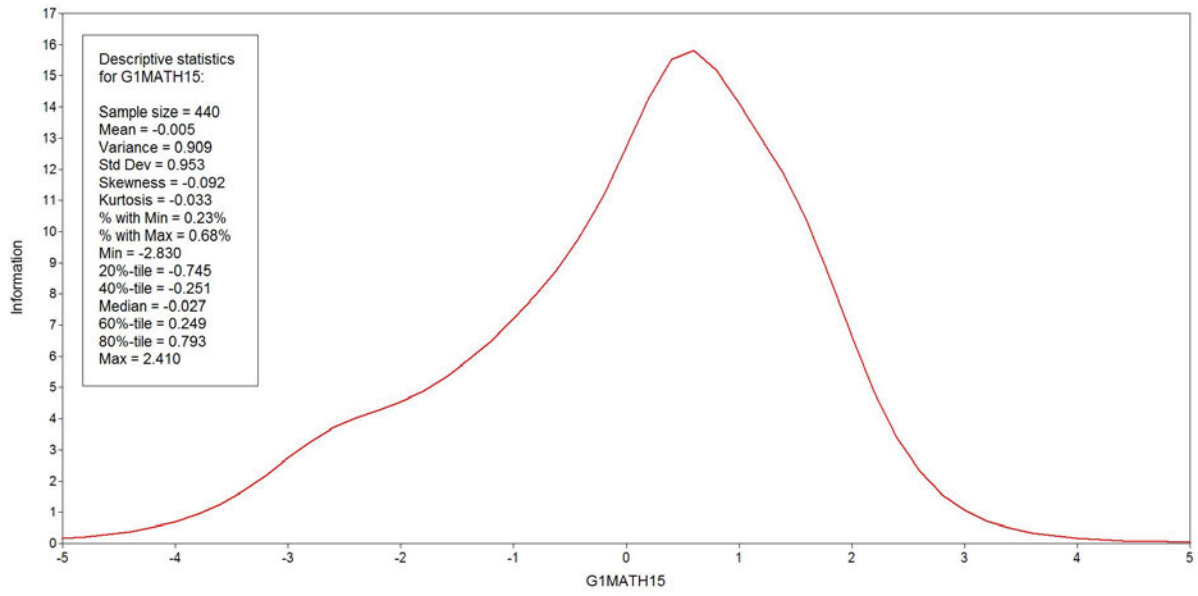


Figure 8. Grade 1 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

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

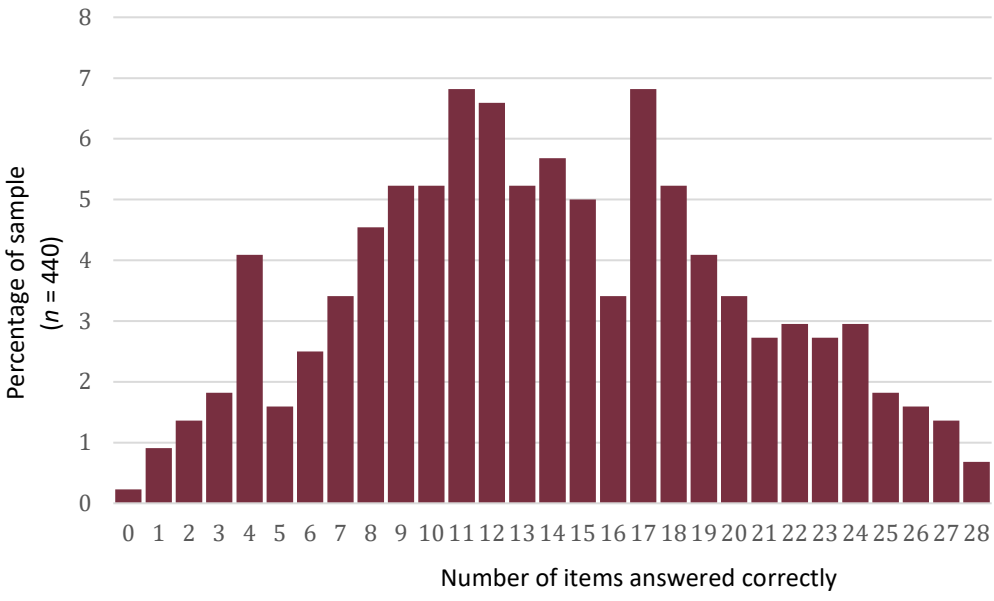


Figure 9. Distribution of the number of items individual students in the Grade 1 sample answered correctly on the reduced set of items.

4.5.2. Grade 2 scale reliabilities

The scale reliabilities for the Grade 2 MPAC interview suggested acceptable reliability for all scales. We calculated a composite reliability for the Grade 2 higher-order math factor of .89. The Grade 2 higher-order math factor composite reliability was calculated as

$$\frac{(0.478 + 0.727 + 0.603 + 0.663 + 0.495)^2}{(0.478 + 0.727 + 0.603 + 0.663 + 0.495)^2 + (0.167 + 0.395 + 0.123 + 0.372 + 0.075)} = 0.89\,,$$

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.

Table 25 relays the α, β, and ω_h ordinal reliability coefficients for reduced set of items by subscale and for the total scale. All α estimates for all subscales met or exceeded the target of .8, except for the COMP scale, which had an estimated α reliability of .71. As with the Grade 1 interview, comparison between the αs and βs revealed a range of discrepancies (range .02 to .28), 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 .03 to .26). Where α exceeds ω_h (i.e., Math, NF, OBS, and WP), the α to ω_h discrepancies indicate the presence of multidimensionality within the scales. Where ω_h exceeds α (i.e., ESRS and COMP), variability was present in the general factor loadings but group factor loadings were relatively small, indicating that lumpiness in the scale is not attributable to multidimensionality. In every case, except for the NF scale, ω_h met or exceeded the conventional minimum value of .7, suggesting composite scores can be interpreted as reflecting a single common source of variance in spite of evidence of some within-scale multidimensionality (Gustafsson & Aberg-Bengtsson, 2010).

Table 25. Grade 2 MPAC Interview Scale Reliability Estimates

		Reliability		
Scale	*N* items	α	β	ω_h
Number Facts (NF)	6	.81	.70	.66
Operations on Both Sides of the Equal Sign (OBS)	6	.97	.94	.93
Word Problems (WP)	6	.88	.84	.72
Equal Sign as a Relational Symbol (ESRS)	3	.81	.55	.81
Computation (COMP)	3	.71	.64	.72
Total (Math)	24	.94	.84	.71

Note. Sample *N* = 416. α, β, and ω_h are ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h, respectively.

Inspection of the 2-pl UIRT TIC in Figure 10, reveals the information curve for the Grade 2 interview to exceed 2.33 (reliability of .7) for the ability range of approximately −2.7 through 2.4. Given the sample descriptives (M = 0.00, SD = 0.942, Min = −2.746, and Max = 2.368), reliability of the scale was probably acceptable for over 98% 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 −1.7 through 2.1, indicating target reliability of the scale was achieved for over 93% of the sample.

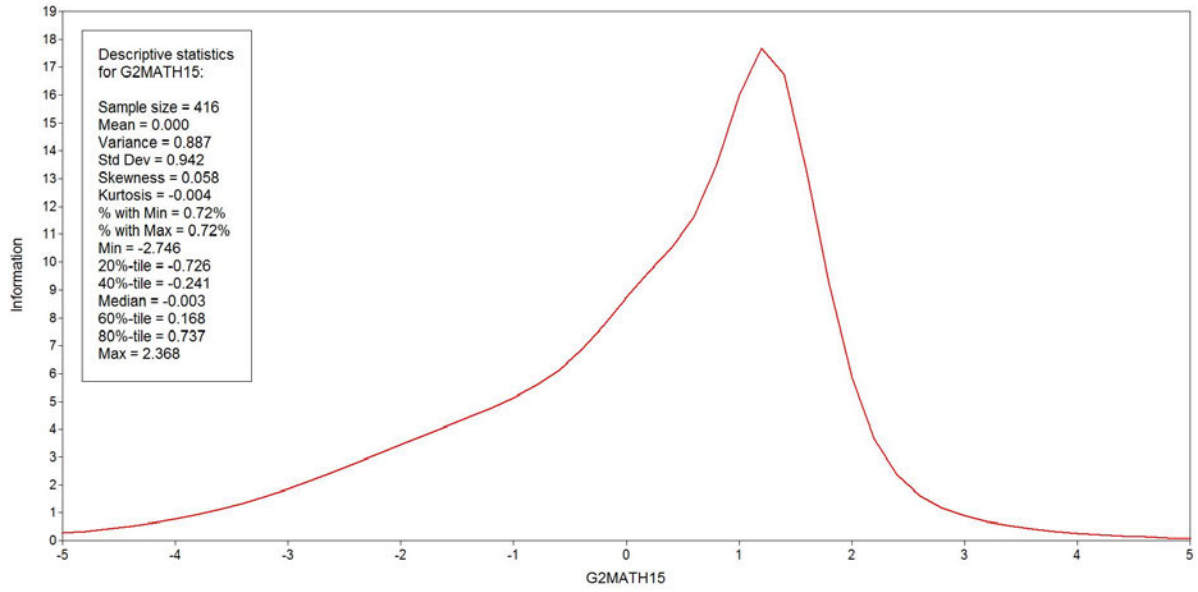


Figure 10. Grade 2 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

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 F. Interested readers can find information about the most common incorrect response to the various items in Appendix G.

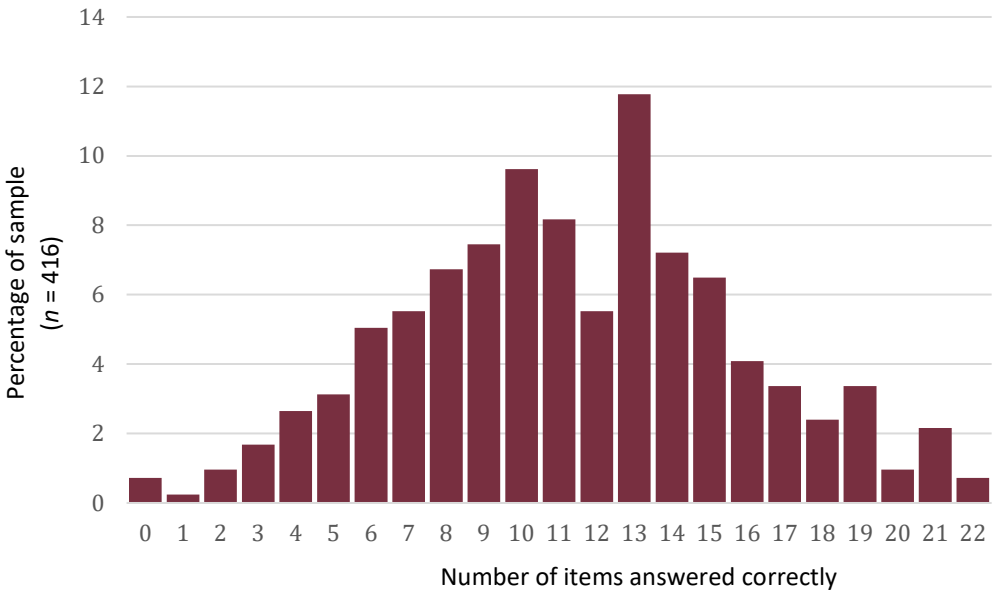


Figure 11. Distribution of the number of items individual students in the Grade 2 sample answered correctly on the complete reduced set of items.

4.6. Concurrent Validity Evaluation

4.6.1. Grade 1 MPAC concurrent validity

The correlations between the Grade 1 MPAC student interview and the ITBS were consistently moderate to large, providing evidence of concurrent validity of the student interview. See Table 26 for correlations between manifest factor scores for the interview scale and standard scores for the ITBS tests. Using correlations > .7 to indicate scale correspondence, we found a pattern of correspondence between the MPAC interview Total scale, the WP and COMP subscales, and the ITBS Math Problems (ITBS-MP) test. Correlations between the MPAC scales and the ITBS_MP ranged from .60 to .77. Although moderately sized correlations were found between the MPAC interview and the ITBS Math Computation (ITBS-MC) test (range .50 to .66), none of the correlations surpassed the .7 correspondence criterion. All correlations were statistically significant at $p < .001$.

Table 26. Correlations Among Grade 1 MPAC Interview Scales and the Iowa Tests of Basic Skills

	1	2	3	4	5	6	7	8
Grade 1 MPAC Interview								
1. Total (Math)	—							
2. Number Facts (NF)	.88	—						
3. Operations on Both Sides of the Equal Sign (OBS)	.86	.71	—					
4. Word Problems (WP)	.97	.82	.79	—				
5. Equal Sign as a Relational Symbol (ESRS)	.83	.69	.76	.76	—			
6. Computation (COMP)	1.00	.88	.86	.97	.83	—		
Iowa Test of Basic Skills (ITBS)								
7. Math Problems, Level 7	**.75**	.65	.65	**.77**	.60	**.75**	—	
8. Math Computation, Level 7	.66	.64	.50	.61	.56	.66	.60	—

Note. Grade 1 MPAC interview $n = 440$. ITBS Math Problems test $n = 1599$. ITBS Math Computation test $n = 1571$. MPAC with ITBS Math Problems correlation $n = 412$. MPAC with ITBS Math Computation correlation $n = 407$. ITBS Math Problems with ITBS Math Computation correlation $n = 1570$. All correlations were statistically significant at $p < .001$. Correlations within borders signify correlations that indicate potential concurrent validity between measures. Boldface values are concurrent validity correlations > .7, indicating ≥ .5 shared variance between measures. ITBS was Form C Level 7. The MPAC interview and ITBS were all administered spring 2015.

4.6.2. Grade 2 MPAC concurrent validity

The findings for the Grade 2 MPAC interview are nearly identical to those for the Grade 1 MPAC interview. The correlations between the Grade 2 MPAC and the ITBS were consistently moderate to large, providing evidence of concurrent validity of the student interview. See Table 27 for correlations between manifest factor scores for the MPAC scales and standard scores for the ITBS tests. Using correlations > .70 to indicate scale correspondence, we found a pattern of correspondence between the MPAC interview Total scale, the WP and COMP subscales, and the ITBS Math Problems test. Correlations between the MPAC scales and the ITBS Math Problems scores ranged from .58 to .75. Although moderately sized correlations were found between the MPAC interview and the ITBS Math Computation

scores (range .50 to .61), none of the correlations surpassed the .70 correspondence criterion. All correlations were statistically significant at $p < .001$.

Table 27. Correlations Among Grade 2 MPAC Interview Scales and the Iowa Tests of Basic Skills

	1	2	3	4	5	6	7	8
Grade 2 MPAC Interview								
1. Total	—							
2. Number Facts	.89	—						
3. Operations on Both Sides of the Equal Sign	.86	.72	—					
4. Word Problems	.95	.82	.76	—				
5. Equal Sign as a Relational Symbol	.85	.72	.74	.76	—			
6. Computation	.97	.84	.79	.91	.79	—		
Iowa Test of Basic Skills (ITBS)								
7. Math Problems, Level 8	**.73**	.63	.58	**.75**	.58	**.72**	—	
8. Math Computation, Level 8	.61	.55	.53	.56	.50	.60	.62	—

Note. Grade 2 MPAC interview $n = 416$. ITBS Math Problems test $n = 1,491$. ITBS Math Computation test $n = 1,482$. MPAC with ITBS Math Problems correlation $n = 395$. MPAC with ITBS Math Computation correlation $n = 393$. ITBS Math Problems with ITBS Math Computation correlation $n = 1,482$. All correlations were statistically significant at $p < .001$. Correlations within borders signify correlations that indicate potential concurrent validity between measures. Boldface values are concurrent validity correlations > .7, indicating ≥ .5 shared variance between measures. ITBS was Form C Level 8. The MPAC interview and ITBS were all administered spring 2015.

References[4]

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., Franke, M. L., Levi, L., & Empson, S. B. (1999). Children's mathematics: Cognitively guided instruction. Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). Children's mathematics: Cognitively guided instruction (2nd ed.). Portsmouth, NH: Heinemann.

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), 499–531.

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.

Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., and 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 on line at 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.

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

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*, 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, B., & Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis*. Mplus Web Notes (No. 17).

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

National Governors Association Center for Best Practices (NGACBP) & Council of Chief State School Officers (CCSSO). (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*, 57–74.

Revelle, W. (2016). *psych: Procedures for Personality and Psychological Research* (Version 1.6.6). Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych.

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016). Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2014. (Research Report No. 2016-04). Tallahassee, FL: Learning Systems Institute, Florida State University. DOI: 10.17125/fsu.1508174887.

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.

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

End-of-Year Student Interviews

A.1. Purpose of the Interview

The primary purpose of the student interview is to gather information on what strategies students are using to solve mathematics problems involving number, operations, and algebraic thinking. We will then use information about student strategies to look for associations between strategies students use and their overall achievement on the interview and the ITBS items as well as associations between observed strategies and treatment condition.

A.2. General Protocol

Overall, the interviews are expected to take about 40-45 minutes for each student. The interviewer script and the detailed blueprint contain important information and guidance. Please be very familiar with those two documents and ask any questions that you may have.

Based on initial pilot testing, the interviewers stated a preference for giving the students the problems one-at-a-time. For that reason, the students' pages will not be stapled. Also, to keep the notes sheet synchronized with the interview, the interviewers' pages parallel the students' pages. Also, we will only provide two markers to the student. They should both be a dark color (e.g., blue and purple). This is because the students were spending too much time switching colors. Two colors should give them the opportunity to choose, but it should decrease the likelihood that they will try to use a new color on every problem and lengthen the overall time of the interview.

There are some instructions to interviewers on the interview script regarding how to read the questions. In general, the interviewer will speak the words in bold typeface. There are going to be a great many things to consider as we conduct these interviews, and there will be many times when you will find that you have to use your best discretion and clinical judgment regarding what to do and how to record data on what students did.

In order for our data set to be meaningful, there must be some standardization of the procedures so that we can compare results in a meaningful way. To that end, we should follow the interviewer script with fidelity, and we should ask questions of the student only for the purpose of getting the level of detail required to accurately and reliably record the key features in how the student solved the problem. Ultimately, we need to come, as a group, to a consistent understanding of how to interpret and code observed student strategies. It is our hope that you can take advantage of this opportunity to learn about student thinking, learn about the interview process, learn about the process of attaining inter-rater reliability in measures involving human observers, and make important contributions to furthering knowledge of teaching and learning in mathematics. For those of us who fully take part in this experience, the process will be a great learning experience.

We are interested in observing and recording how students solved the problems in the questionnaire. For our purposes, there is no reason to ask students to prove their solutions or to solve problems in a different way. We are also not interested in whether a student could have solved it in another way. Please attempt to refrain from asking questions that ask students to prove their solutions or solve the

problem in a different way. We are looking for that "go to" strategy that the student actually used to solve the problem in that moment.

We should always avoid interrupting a student who is thinking about how to solve a problem. Given that we should not interrupt a student's thinking, we should keep the tempo high enough that students are responding in the moment. In other words, we don't want to give too much time for reflection or re-analysis of a problem or solution method. We want to pose the problem, observe the student solving it, ask questions to clarify their thinking processes, and move to the next problem.

We are most interested in how the student solved the problem. From the beginning of the interview, we want to teach students that we want them to explain how they solved the problem. Be considerate of how you praise students. If you praise them, praise them indirectly by saying "Thank you. I understand just how you did that one." That will help them to understand their role in the interview and will help to reinforce the message that we are interested in how they solve or think about mathematics problems.

In many cases, the codes will be sufficient to describe student solution strategies. In others, the codes will not be sufficient. Furthermore, it may take considerable experience to become fluent with using our coding system. For these reasons and more, it is recommended that you take copious notes during the interview and then review and reflect upon the interview after the student is gone in order to code your notes and enter the data.

As an end goal, we want to try to code the students' strategies in real time. That will not be attainable in the first dozen or more interviews that you do. It may not be attainable in every case, regardless of how much experience you have as an interviewer. Take plenty of notes and use the time you will have between interviews to review those notes, understand what the student was thinking, and determine the data that will be entered to the computer. If you start to see things that multiple students do that are not described by the codes, please notify Rob Schoen (rschoen@lsi.fsu.edu).

A.3. Inter-rater Reliability and Video Consent

When allowable, student interviews will be video and audio recorded using the plug-in webcam device on the laptop used by the interviewer. These videos will be transferred from that original computer onto an external hard drive at the school site at the end of each day of interviews. The videos will be sampled at random, and another observer will record data. Those data will be entered along with the data entered by the interviewer, and the level of agreement will be used to estimate inter-rater reliability.

When the other coder does not agree or has any other questions about the interview, you can expect to hear from the other coder, especially early in the process. The goal of this communication is to maximize our confidence in the data. For instance, there may have just been a data entry error. Or, it might be that we need to redefine a code or that not all interviewers are using the codes consistently.

In the cases that we do not have parental consent to video record students, two interviewers will record data in real time for the student. In these cases, it is preferred that only one interviewer do the talking. The other adult will be introduced to the student as an "observer" who is also recording data under the auspices of studying how students think about math problems. After the interview, the interviewer and the observer will briefly discuss their observations and their codes for each item in the interview. In these discussions, the observers can change their minds if they think their original codes were inaccurate. The goal will be to have 100% agreement on how to code the student's strategies. In some

cases, the interviewer and observer will not agree on the final code, and that is okay. Both sets of data will be entered into the computer. In any case, these post-interview discussions will serve as an important opportunity for the interviewers to have ongoing discussion about student thinking and interview protocol.

A.4. On Incorrect Answers

This will likely be the greatest challenge in recording data. As a group, we have decided that it is important to have data describing how students solve problems, irrespective of whether they arrived at correct answers. In general, attempt to clearly identify the answer provided by the student, and then record data explaining how a student arrived at that answer. The prescribed categories and codes will be sufficient to describe many of the ways that students solve the problems, but they will not be sufficient to describe every way.

The coding categories come from a body of literature that is focused on the myriad correct ways that students solve problems. There is much less depth of coverage in the literature about ways that students arrive at incorrect answers. To the extent that we can, we will use the same codes as a means to explain how a student arrived at an answer, irrespective of whether the answer is predetermined as correct. When those codes are insufficient, please use the notes section to describe the reasoning process of the student. In some cases, the student will be guessing or simply doesn't care. In most cases, however, there will be some logic to the student's response, so we will try to understand that logic and record it. This may be an area where this project ultimately makes an important contribution to scholarly knowledge that may be useful in teaching and learning.

A.5. On Materials

Students will not have tools available during the Counting section of the interview, but the tools will become available for the Word Problems section and continue to be available in the Equations and Calculations section.

All students will have the same tools available to use. We will invite the student to use tools at the beginning of the interview, but we will not prompt them to use any specific tools during the course of the interview (except when it is required for us to understand how they solved the problem, such as asking them to show us how they secretly used their fingers).

The following tools will be available for students to use:

1. Paper and markers (exactly two dark colors of marker)
2. Base Ten blocks (approximately 45 unit cubes, 12 tens rods, and two hundred flats)
3. Snap Cubes (approximately 130 cubes, separated into units)
4. Fingers and toes
5. Ceiling tiles, pegboard holes, hair beads, and other objects in the environment

We will not provide additional tools such as rulers, number lines, hundreds charts, etc. If students ask for these tools, you should respond along the lines of "I'm sorry, I didn't bring one of those today." If they must have one, they could use the markers and paper to create one.

A.6. On Verbalization of Numbers

There are a few cases of three-digit numbers in the interview. There are several accepted ways of saying numbers such as 105. Colloquially, people may say "one-oh-five," "one hundred and five," or "one hundred five." The interviewers should use the latter style, as it is acknowledged as the proper way to say the number. In other words, we speak it according to place value, and there is no "and" in the number.

A.7. On Multiple Strategies

There are no instances in this interview when we ask students to try to solve problems in more than one way. We are also stopping short of asking students to prove that their method or answer is valid. In this interview, we aim to determine the primary way that the student used to solve the problem. In almost all cases, it will be reasonably clear that the student used a single strategy.

Students may offer to show more than one way to solve a problem. If they offer more than one way, ask the student to show you just one way. Try to refrain from asking them to make a value judgment (such as "show me the best way" or "show me your favorite way"). There may be instances when students use more than one way. This is a place where you will need to use clinical judgment to determine which way the student really used to arrive at the answer. As we continue to interview students together, we will have some specific instances to discuss, and we will strive to come to a consistent way to handle these situations and record the most appropriate data.

A.8. On Interviewing

In general, the students will be excited to have the chance to be interviewed. Almost all of them will be agreeable to the "rules of the game" that we define. Thus, it is important for the interviewer to establish a clear focus, routine, and pace from the beginning. You will want to introduce yourself, be amicable, explain what you want to do, and get down to business. Do not chit-chat at the beginning of the interview.

There will be a balance to strike in the interview. You need to set the tempo, but you must also allow the student sufficient time to think. As a rule, do not interrupt their thinking or writing to ask questions unless you have an urgent reason to interrupt them.

Students should not be expected to be good at explaining their thoughts verbally. In many cases, it will be necessary to coach the student to talk about his or her thinking. It may be helpful to give them feedback when they do manage to make their thinking clear by saying "I see just how you did that."

A.9. On Observation and Inference

We strive to observe how students solved the problems. Ultimately, this process necessarily involves some inference. In general, we will trust what students say and do at face value. That is sometimes easier to do than other times.

A very important part of the observation and interview process is the observation of behaviors such as the movement of fingers, lips moving silently, the student's writing (and the chronological order of the writing), angle of their head or gaze, facial expressions, gestures, pauses, etc. Watch carefully, and learn to attend to these details. These behaviors will give you cues to ask follow-up questions and will help

you to make decisions about what they were thinking and what strategies they used to solve the problem.

Related to the question of observation or inference, we will accept what the student tells us as truth. For instance, suppose a student produced an answer of "seven" to the computational question "What is 15 minus 8" almost immediately. While you might be tempted to infer that they knew the number fact, we will always follow up with the question "How did you get seven?" If the student responds that they counted in their head, "14, 13, 12, 11, 10, 9, 8," then you will record that the student used a counting strategy (counting backward from 14), because that is what the student told you happened in his or her mind. (Note, this is why we do not ask them to show us another way or to prove their answer; that makes it much more difficult to decide how they really arrived at their answer the first time.)

Don't talk too much. The student should do most of the talking, not you. Spend your time observing behaviors and listening to what they say.

A.10. On Long Pauses

Sometimes, the student will be thinking or working for what feels like a very long time. If it begins to feel awkward, pretend to be busy with something else to reduce the awkwardness, but keep watching carefully. Most likely, if the student is drawing sixty-some circles during this time, the student is working and thinking hard, so the student is not feeling awkward, and the awkwardness is all in your head. In order to feel less awkward, you might start taking notes on the sequence of things the student is writing at that moment; avoid turning your attention to other matters, such as past or future items.

In either case, this would be a good time to check the video feed to make sure it is capturing the story that is unfolding in front of you. If the student is silently thinking and you feel uncomfortable, you might turn your attention to the laptop to watch the facial expressions that way if it feels less awkward.

If the student is silently thinking and then you pick up on facial expressions, fidgeting, or changes in eye tracking that indicate the student is no longer thinking (or is stuck), then you may ask them to explain what they were thinking about or ask if they want to hear the problem again. Ultimately, if they cannot solve the problem, permit them to move to the next problem. It is okay if they do not know how to solve every problem.

A.11. On Awareness of Your Own Personal Bias

In the spirit of good science and ethical evaluation, it is very important that we all attempt to be unbiased in our interview protocols and interpretation of student strategies. You may want the study to have positive effects in favor of CGI; you may want something else. The goal of this particular endeavor is to generate accurate data. Please be careful not to allow your personal bias interfere with getting good data. If you are afraid that you are biasing the data, please speak with the project managers. We will look for a solution to avoid that bias and still allow you to contribute to the end goals.

Along those lines, we will not be identifying which schools are in the various treatment conditions. We also ask that you do not ask teachers, students, or others involved for information about their role in the research study.

One of the ways you might introduce bias is by having a different behavior depending upon whether students give correct or incorrect answers. Because you are interested in teaching and learning of

mathematics, it is probably in your nature to want the students to generate correct answers. Try to respond in the same ways whether they provide incorrect answers or correct answers. For our purposes, both cases are very interesting, so always demonstrate to the student that you are interested in how they got their answer.

A.12. On Questioning Strategies

It is very important that you refrain from correcting the student or teaching mathematics. Keep in mind that this interview is strictly an assessment. This is not the time to teach.

In this type of situation, "The golden rule is to avoid suggestions" (Piaget, 1976). Ginsburg (1997) gives the following example: Don't say "Did you get the answer by ….?" Rather, ask "How did you get that answer?" or "What did you do to get that answer?"

A.13. On Awareness of Your Body Language and Cues to the Student

The student is very excited to have the complete attention of an adult, and almost all children of this age want to please adults. As a result, you will need to practice awareness of your facial expressions and sitting position. Leaning forward or making eye contact too strongly may scare the child. You will also want to be aware of cultural norms that may make children more comfortable or less comfortable in the interview situation. To the extent that you are able or comfortable, attempt to identify the cultural background of the child you are interviewing and provide accommodations in your own behavior that will neutralize the differences so that the data are unbiased.

For instance, some students may be taught at home that they should look adults in the eye and speak boldly when they are talking to them. Other students may be taught at home that they should not speak boldly to adults or to look them too strongly in the eye. Try to read the preference of the child you are interviewing and be respectful of the background of the student in order to avoid causing the child to feel anxiety. We want them to be comfortable so that they can think clearly and communicate well.

Do not show disapproval of methods they are using. The perception of boredom or other lack of interest may be interpreted as disapproval. Show interest in what the student is doing as a way to coach the student to show his or her thinking process.

A.14. An Incomplete List of Suggested (and not suggested) Questions

The following questions are suggested.

1. How did you get [that answer]?
2. What did you do in your head?
3. Can you show me (or tell me) how you got [that answer]?
4. Did you use your fingers to keep track? Can you show me what you did with your fingers?
5. How did you know when to [stop counting]?

The following questions would not be appropriate for this interview:

1. Can you show me another way?
2. How do you know [that you can count from that one]?
3. Why did you use your fingers?
4. Why did you use the base ten blocks on this one?
5. Why did you use 8+8 and not 8+7?
6. Could you use the base ten blocks to show me your thinking?

Appendix B—Grade 1 Interview Script

Interviewer: _____ **Date : _____**

Student Name: _____ **Grade: _____**

School: _____ **Teacher: _____**

Start Time: _____

GRADE 1 STUDENT INTERVIEW SCRIPT

Hi, my name is _____ and I am from _____.

Before we begin, I'd like to make sure that I have your name correct. Will you please tell me your name? *[verify the child's name]*

What grade are you in? *[verify the grade level]*

I asked you to come here and talk with me, because I'm interested in learning about how students solve math problems.

I brought some math problems with me that I am going to ask you to solve. Sometimes, I may write things down or ask you questions about your answer. Other times I may not. This doesn't mean you are right or wrong. It just means I need some more information about how you were thinking about the problem or needed to record what you said. Also, since I don't know how you solve problems as well as your teacher does, I might need to ask a lot of questions.

Is it okay with you if I ask you some questions about math? *[If the child does not assent, terminate the interview and instead interview the student selected as alternate for this student.]*

[If the parent or guardian consents to video, read the following. If not, skip the video question and make certain that there is an observer present to record data on the interview.]

I would like to record our conversation with this video camera just to be sure I don't miss anything while we are talking. Is it okay with you if I video our conversation?

[If the child does not agree, contact Amanda. She will provide an additional interviewer to record data on the interview and you can then proceed with the interview without video.]

You don't have to solve all of the problems if you don't want to. You can choose. You can ask me questions or change your mind about this interview at any time. If you don't want to do it, just let me know. Okay? Are you ready to start?

Number Facts:

For each of these items, the interviewer should cover the items below with a blank sheet of paper and read each expression aloud at a normal pace to the student. The expressions should be read using the word "plus" for the addition symbol, and the word "minus" for the subtraction symbol.

"We are going to start with some number fact problems. First, I am going to read each problem aloud to you. Then I want you to solve the problem mentally, using your brain. If you want to use your fingers, that is fine too. If you do choose to use your fingers, please put them where I can see them and count out loud so I can see and hear how you are thinking about the problem. Also, I am going to use this folder to cover up the problems you haven't done yet and I will move it up after you do each problem."

NF1

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] [redacted] [redacted] Other Derived Fact (Explain):	Recall: [redacted] [redacted]
Fingers		

NF2

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] [redacted] [redacted] Other Derived Fact (Explain):	Recall: [redacted] [redacted]
Fingers		

NF3

[redacted] _____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

NF4

[redacted] _____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	
MERCY (Move to Subtraction NF items)		

NF5

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] Other Derived Fact (Explain):	Recall: [redacted]
Fingers		
MERCY (Move to Subtraction NF items)		

NF6

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] Other Derived Fact (Explain):	Recall: [redacted]
Fingers		

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

NF7

_____ **NR**_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____		Recall: 
Fingers	Other Derived Fact (Explain):	

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

NF8

_____ **NR**_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____		Recall: 
Fingers	Other Derived Fact (Explain):	

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

NF9

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	[redacted] Other Derived Fact (Explain):	
MERCY (Move to Solving Equations items)		

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

NF10

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	[redacted] Other Derived Fact (Explain):	
MERCY (Move to Solving Equations items)		

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

Solving Equations

For this section the interviewer should read each equation to the student exactly as it is written. The subtraction symbol should be read as "minus," and the addition symbol should be read as "plus." The equal sign should be read as "equals," and for the blank line the interviewer should read it as, "what number."

Below are two examples of how each equation should be read.

[redacted] + ☐ = [redacted] *should be read as, "[redacted] plus what number equals [redacted]."*

☐ − [redacted] = [redacted] *should be read as, "what number minus [redacted] equals [redacted]."*

For the next part, I want you to tell me the number that goes in the box to make the equation or number sentence correct.

SE1

[redacted]

NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

SE2

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		

SE3

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		

SE4

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		
MERCY	(Move to Word Problem items)	

SE5

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		
MERCY	(Move to Word Problem items)	

Word Problems:

"**For the next part, I am going to ask you to solve some story problems. If you want, you can solve them mentally, with just your brain, or you can use your fingers or any of these tools to help you.** *[present the paper, markers, base ten blocks, and snap cubes and check for understanding of each tool.]* **These are snap cubes, and they can be broken apart. These are base ten blocks. Some of them are single blocks and others are grouped together. They don't come apart like the snap cubes. Just like before, if you use your fingers, I would like you to keep them where I can see them and count out loud so I can see and hear how you are thinking about the problem.**

I am going to read each story problem aloud. If you want me to read it again, you just have to ask. I will read it as many times as you want. I have the problem on my paper, and you have the same words on your paper. You can also use this marker to write or draw on the paper if that helps you solve the problem.

Are you ready for the first one?"

[Read each word problem aloud at least one time. If the child asks questions about the problem, you can read it again, but always read the entire problem; do not read only part of the problem or answer questions about "how many" without first reading the problem in its entirety. If a child appears stuck for an extended time, you might want to ask them whether they are thinking about the problem and if they can tell you what they are thinking. The goal here is to document how a child really arrived at their answer. The goal is not to ask the child to prove it or show it in a different way. To that end, the standard question will be **"How did you get [say the student's answer]?"** *Be sure not to inflect when you get say aloud their answer. This can lead the student to think that their answer is correct or incorrect. When the child gives an explanation that provides the information you require, say,* **"Okay, I see just how you got that answer. Thank you."**

WP1

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Join All	Join To	Separate From	Separate To	Matching
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Incrementing: [redacted]				
	Other (Explain):				
Addition Standard Algorithm	Counting	Buggy Algorithm [redacted]			
	Fact Recall	[redacted]			Other (Explain)
	Representation (circle one)	Vertical	Non-Vertical	Mental	
Subtraction Standard Algorithm	Attempted to use the [redacted]				
	Used buggy [redacted] algorithm (Use coding options above)				
Recall	[redacted]				
DNS (Explain)					

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

WP2 [redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Grouping	Join All	Join To		
	Separate From	Separate To	Matching		
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	[redacted] _____	[redacted] _____			
	[redacted]	Partitioning Number Strategy: Incrementing (Explain)			
	[redacted] _____				
	Other (Explain):				
Recall	[redacted]	[redacted]	[redacted]	[redacted]	
DNS (Explain)					

WP3

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Join All	Join To	Separate From	Separate To	Matching
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Incrementing: [redacted]			Derived Fact [redacted]	
	Other (Explain):				
Addition Standard Algorithm		Buggy Algorithm			
	Counting	[redacted]			
	Fact Recall	[redacted]			Other (Explain)
	Representation (circle one)	Vertical	Non-Vertical	Mental	
Subtraction Standard Algorithm	Attempted to use [redacted]				
	Used buggy [redacted] algorithm (Use coding options above)				
Recall	[redacted]				
DNS (Explain)					

___ Additive ___ Subtractive

___Take-away Model

___Difference Model

___Other

WP4

[redacted]

Response: _____ **NR_____**

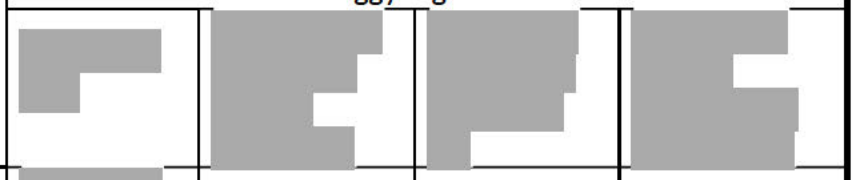
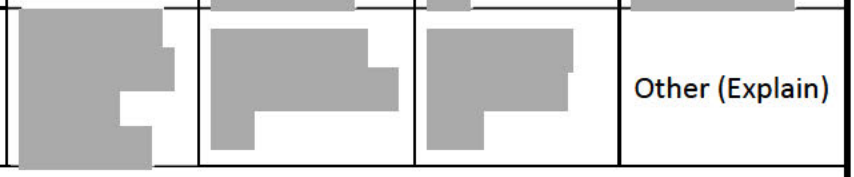
How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Measurement	Join All		Join To	
	Separate From	Separate To		Matching	
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Place Value Explanation: [redacted] [redacted] Place Value Explanation (Circle response): [redacted] [redacted] [redacted]			Repeated Addition by _____ Repeated subtraction by _____	
	Other (Explain):				
Standard Algorithm	Add	Counting	Buggy Algorithm [redacted]	[redacted]	[redacted]
	Sub	Fact Recall	[redacted]	[redacted]	[redacted] Other (Explain)
	Representation (circle one)	Vertical	Non-Vertical	Mental	
Recall	[redacted]	[redacted]	[redacted]	[redacted]	
DNS (Explain)					
MERCY	(Move to True/False items)				

WP5

Response: _____ NR_____

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Join All	Join To	Separate From	Separate To	Matching
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Incrementing 		Compensation 		
	Other (Explain):				
Standard Algorithm	Add	Counting	Buggy Algorithm 		
	Sub	Fact Recall			Other (Explain)
	Representation (circle one)		Vertical	Non-Vertical	Mental
DNS (Explain)					
MERCY	(Move to True/False items)				

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

WP6 [redacted]

Response: _____ NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td colspan="4">Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td colspan="4">Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="8">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="8">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="8">Incrementing [redacted]</td></tr>
<tr><td colspan="4">Compensation [redacted]</td><td colspan="4">Combining Tens and Ones [redacted]</td></tr>
<tr><td colspan="8">Other (Explain):</td></tr>
<tr><td rowspan="4">Standard Algorithm</td><td colspan="2" rowspan="2">Add</td><td colspan="2" rowspan="2">Counting</td><td colspan="4">Buggy Algorithm</td></tr>
<tr><td colspan="2">[redacted]</td><td colspan="2">[redacted]</td></tr>
<tr><td colspan="2" rowspan="2">Sub</td><td colspan="2" rowspan="2">Fact Recall</td><td colspan="2">[redacted]</td><td colspan="2">[redacted]</td></tr>
<tr><td colspan="2">Other (Explain)</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td colspan="2">Vertical</td><td colspan="2">Non-Vertical</td><td colspan="2">Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="8"></td></tr>
<tr><td>MERCY</td><td colspan="8">(Move to True/False items)</td></tr>
</table>

WP7

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
_____ORQSS					
_____Counting					
_____Ad Hoc					
_____Recalled Fact					
_____Standard Algorithm Choose: _____Addition _____Subtraction _____Multiplication					
_____Other (Explain):					
DNS (Explain)					
MERCY	(Move to True/False items)				

*Check a box when a major category is observed (check all that apply).

Equations: True/False

For each of the items in this section of the interview, the interviewer should show the student the equation and ask the student to read each equation aloud. The interviewer should ensure that the student has correctly read the equation exactly as it is written, from left to right. If the student is unable to do so after two tries, the interviewer should read it to the student and then ask the student to read it back.

After the student correctly reads each equation, the interviewer asks, "Is that equation true or not true?"

After the student tells if it is true or not true, the interviewer asks, "What makes this equation [true or not true]?"

For this part, I am going to ask you a few questions about equations or number sentences. For these, I will first ask you to read the equation out loud to me. Then, if the equation is correct as it is written, I want you to say 'The equation is True.' If the equation is not correct as it is written, I want you to say 'The equation is Not True' or 'The equation is False.'

After you tell me whether the equation is true or not true, I am going to ask you to tell me how you decided on your answer. For these problems, I am going to use the folder again so I can cover up the ones we haven't done yet.

Are you ready?"

TF1

Please read this equation just as it is written. *[Pause for student to read equation.]*

Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True	
	Other (Explain):
Not True	
	Other (Explain):
DNS (Explain)	

TF2

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True		
	Other (Explain):	
Not True		
	Other (Explain):	
DNS (Explain)		

TF3

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True	
	Other (Explain):
Not True	
	Other (Explain):
DNS (Explain)	

TF4

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True		
	Other (Explain):	
Not True			
	Other (Explain):	
DNS (Explain)		

TF5

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF6

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF7

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF8

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True		
	Other (Explain):	
Not True		
	Other (Explain):	
DNS (Explain)		

Multi-Digit Computation:

For this section the interviewer should read each equation to the student exactly as it is written. The subtraction symbol should be read as "minus," and the addition symbol should be read as "plus." The equal sign should be read as "equals," and for the blank line the interviewer should read it as, "what number."

Below are two examples of how each equation should be read.

[ ] + [ ] = [ ] *should be read as, "* [ ] *plus what number equals* [ ] *."*

[ ] – [ ] = [ ] *should be read as, "what number minus* [ ] *equals* [ ] *."*

"For the next three problems, you are going to need the marker. I am going to read an equation aloud to you, and your job is to write the number that goes in the box to make the equation correct. To solve these problems, you can use the base ten blocks, snap cubes, marker and paper, fingers, or you can solve it mentally with just your brain. But no matter how you solve it, I want you to write the number in the box that makes the equation correct.

Are you ready to do that?"

MDC1

NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="2">Incrementing</td><td colspan="3" rowspan="2">Compensation</td></tr>
<tr><td colspan="2">Place Value Explanation</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td rowspan="2">Add</td><td rowspan="2">Counting</td><td colspan="4">Buggy Algorithm</td></tr>
<tr><td></td><td></td><td></td><td rowspan="2">Other (Explain)</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">Representation (circle one)</td><td colspan="4">Vertical Non-Vertical Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

MDC2

[redacted]

NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">Incrementing
[redacted]</td><td colspan="2">Combining Tens and Ones
[redacted]</td></tr>
<tr><td colspan="3">Compensation
[redacted]</td><td colspan="2">[redacted]</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="3" rowspan="2">Buggy Algorithm
[redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td>Other (Explain)</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td colspan="3">Vertical Non-Vertical Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

MDC3

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
_____					
_____					
_____					
_____					
_____Other (Explain)					
DNS (Explain)					

We only have one more problem. For this last one, I want to see if you can solve it without using any of the manipulatives or the marker. Are you ready for the last problem?

MDC4

[redacted]

NR____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td colspan="2">Fingers</td><td colspan="4">Other</td></tr>
<tr><td>Counting</td><td colspan="6">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">Combining Tens and Ones [redacted]</td><td colspan="3">Incrementing [redacted]</td></tr>
<tr><td colspan="6">Compensation [redacted]</td></tr>
<tr><td colspan="6">Other (Explain):</td></tr>
<tr><td rowspan="2">Mental Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="4">Buggy Algorithm [redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td>[redacted]</td><td>[redacted]</td><td>[redacted]</td><td>Other (Explain)</td></tr>
<tr><td>DNS (Explain)</td><td colspan="6"></td></tr>
</table>

Appendix C—Grade 2 Interview Script

Interviewer: _____ **Date : _____**

Student Name: _____ **Grade: _____**

School: _____ **Teacher: _____**

Start Time: _____

GRADE 2 STUDENT INTERVIEW SCRIPT

Hi, my name is _____ and I am from _____.

Before we begin, I'd like to make sure that I have your name correct. Will you please tell me your name? *[verify the child's name]*

What grade are you in? *[verify the grade level]*

I asked you to come here and talk with me, because I'm interested in learning about how students solve math problems.

I brought some math problems with me that I am going to ask you to solve. Sometimes, I may write things down or ask you questions about your answer. Other times I may not. This doesn't mean you are right or wrong. It just means I need some more information about how you were thinking about the problem or needed to record what you said. Also, since I don't know how you solve problems as well as your teacher does, I might need to ask a lot of questions.

Is it okay with you if I ask you some questions about math? *[If the child does not assent, terminate the interview and instead interview the student selected as alternate for this student.]*

[If the parent or guardian consents to video, read the following. If not, skip the video question and make certain that there is an observer present to record data on the interview.]

I would like to record our conversation with this video camera just to be sure I don't miss anything while we are talking. Is it okay with you if I video our conversation?

[If the child does not agree, contact Amanda. She will provide an additional interviewer to record data on the interview and you can then proceed with the interview without video.]

You don't have to solve all of the problems if you don't want to. You can choose. You can ask me questions or change your mind about this interview at any time. If you don't want to do it, just let me know. Okay? Are you ready to start?

Number Facts:

For each of these items, the interviewer should cover the items below with a blank sheet of paper and read each expression aloud at a normal pace to the student. The expressions should be read using the word "plus" for the addition symbol, and the word "minus" for the subtraction symbol.

"We are going to start with some number fact problems. First, I am going to read each problem aloud to you. Then I want you to solve the problem mentally, using your brain. If you want to use your fingers, that is fine too. If you do choose to use your fingers, please put them where I can see them and count out loud so I can see and hear how you are thinking about the problem. Also, I am going to use this folder to cover up the problems you haven't done yet and I will move it up after you do each problem."

NF1

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

NF2

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

NF3

[redacted] _____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

NF4

[redacted] _____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	
MERCY (Move to Subtraction NF items)		

NF5

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] Other Derived Fact (Explain):	Recall: [redacted] [redacted]
Fingers		
MERCY (Move to Subtraction NF items)		

NF6

[redacted] _____ **NR_____**

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted] Other Derived Fact (Explain):	Recall: [redacted]
Fingers		

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

NF7

_____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	 Other Derived Fact (Explain):	Recall: 
Fingers		

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

NF8

_____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	 Other Derived Fact (Explain):	Recall: 
Fingers		

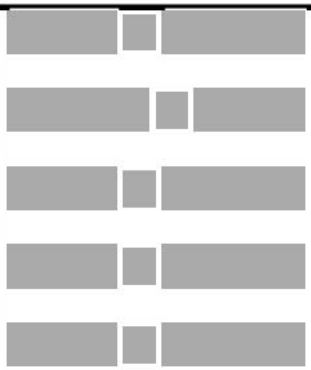
___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

NF9

_____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____		Recall: 
Fingers	Other Derived Fact (Explain):	
MERCY (Move to Solving Equations items)		

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

NF10

_____ NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____		Recall: 
Fingers	Other Derived Fact (Explain):	
MERCY (Move to Solving Equations items)		

___ **Additive** ___ **Subtractive**

___Take-away Model
___Difference Model
___Other

Solving Equations

For this section the interviewer should read each equation to the student exactly as it is written. The subtraction symbol should be read as "minus," and the addition symbol should be read as "plus." The equal sign should be read as "equals," and for the blank line the interviewer should read it as, "what number."

Below are two examples of how each equation should be read.

[redacted] + ☐ = [redacted] *should be read as, "[redacted] plus what number equals [redacted]."*

☐ − [redacted] = [redacted] *should be read as, "what number minus [redacted] equals [redacted]."*

For the next part, I want you to tell me the number that goes in the box to make the equation or number sentence correct.

SE1

[redacted]

NR_____

How did you get [say the student's answer]?

Count forward/backward from _____ to _____ by _____	[redacted]	Recall: [redacted]
Fingers	Other Derived Fact (Explain):	

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

SE2

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		

SE3

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		

SE4

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		
MERCY	(Move to Word Problem items)	

SE5

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Other
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____ [redacted]		
_____Other (Explain):		
DNS (Explain)		
MERCY	(Move to Word Problem items)	

Word Problems:

"**For the next part, I am going to ask you to solve some story problems. If you want, you can solve them mentally, with just your brain, or you can use your fingers or any of these tools to help you.** *[present the paper, markers, base ten blocks, and snap cubes and check for understanding of each tool.]* **These are snap cubes, and they can be broken apart. These are base ten blocks. Some of them are single blocks and others are grouped together. They don't come apart like the snap cubes. Just like before, if you use your fingers, I would like you to keep them where I can see them and count out loud so I can see and hear how you are thinking about the problem.**

I am going to read each story problem aloud. If you want me to read it again, you just have to ask. I will read it as many times as you want. I have the problem on my paper, and you have the same words on your paper. You can also use this marker to write or draw on the paper if that helps you solve the problem.

Are you ready for the first one?"

[Read each word problem aloud at least one time. If the child asks questions about the problem, you can read it again, but always read the entire problem; do not read only part of the problem or answer questions about "how many" without first reading the problem in its entirety. If a child appears stuck for an extended time, you might want to ask them whether they are thinking about the problem and if they can tell you what they are thinking. The goal here is to document how a child really arrived at their answer. The goal is not to ask the child to prove it or show it in a different way. To that end, the standard question will be **"How did you get [say the student's answer]?"** *Be sure not to inflect when you get say aloud their answer. This can lead the student to think that their answer is correct or incorrect. When the child gives an explanation that provides the information you require, say,* **"Okay, I see just how you got that answer. Thank you."**

WP1

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Join All	Join To	Separate From	Separate To	Matching
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Incrementing: [redacted]				
	Other (Explain):				
Addition Standard Algorithm	Counting	Buggy Algorithm [redacted]			
	Fact Recall	[redacted]			Other (Explain)
	Representation (circle one)	Vertical	Non-Vertical	Mental	
Subtraction Standard Algorithm	Attempted to use [redacted]				
	Used buggy [redacted] algorithm (Use coding options above)				
Recall	[redacted]				
DNS (Explain)					

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

WP2

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
ORQSS	Join All	Join To	Separate From	Separate To	Matching
	Other (Explain):				
Counting	Count forward/backward from _____ to _____ by _____				
Ad Hoc	Incrementing: [redacted]			Derived Fact [redacted]	
	Other (Explain):				
Addition Standard Algorithm	Counting	Buggy Algorithm [redacted]			
	Fact Recall	[redacted]			Other (Explain)
	Representation (circle one)	Vertical	Non-Vertical	Mental	
Subtraction Standard Algorithm	Attempted to use [redacted]				
	Used buggy [redacted] algorithm (Use coding options above)				
Recall	[redacted]				
DNS (Explain)					

___ Additive **___ Subtractive**

___Take-away Model

___Difference Model

___Other

WP3 [redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

<table>
<tr><th>Tools</th><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="3">ORQSS</td><td colspan="2">Grouping</td><td colspan="2">Join All</td><td>Join To</td></tr>
<tr><td colspan="2">Separate From</td><td colspan="2">Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">[redacted] _____</td><td colspan="2">[redacted] _____</td></tr>
<tr><td colspan="3">[redacted]</td><td colspan="2">Partitioning Number Strategy: Incrementing (Explain)</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="3" rowspan="2">Buggy Algorithm [redacted]</td><td rowspan="2"></td></tr>
<tr><td>Sub</td><td>Fact Recall</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td>Vertical</td><td>Non-Vertical</td><td>Mental</td><td>Other (Explain)</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

WP4

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

<table>
<tr><th>Tools</th><th colspan="2">Fingers</th><th>Base Ten</th><th>Snap Cubes</th><th>Drawings</th><th>Other</th></tr>
<tr><td rowspan="3">ORQSS</td><td colspan="2">Measurement</td><td colspan="2">Join All</td><td colspan="2">Join To</td></tr>
<tr><td colspan="2">Separate From</td><td colspan="2">Separate To</td><td colspan="2">Matching</td></tr>
<tr><td colspan="6">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="6">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="2">Ad Hoc</td><td colspan="3">Place Value Explanation:
[redacted]
[redacted]
Place Value Explanation (Circle response):
[redacted]</td><td colspan="3">Repeated Addition by _____
Repeated subtraction by _____</td></tr>
<tr><td colspan="6">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="4">Buggy Algorithm
[redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td colspan="3">[redacted]</td><td>Other (Explain)</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td>Vertical</td><td>Non-Vertical</td><td>Mental</td><td></td></tr>
<tr><td>Recall</td><td colspan="2">[redacted]</td><td>[redacted]</td><td>[redacted]</td><td>[redacted]</td><td></td></tr>
<tr><td>DNS (Explain)</td><td colspan="6"></td></tr>
<tr><td>MERCY</td><td colspan="6">(Move to True/False items)</td></tr>
</table>

WP5

[redacted]

Response: _____ **NR_____**

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="5">Incrementing
[redacted]</td></tr>
<tr><td colspan="5">Compensation
[redacted]</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td colspan="2" rowspan="2">Add
Sub</td><td rowspan="2">Counting
Fact Recall</td><td colspan="3">Buggy Algorithm
[redacted]</td></tr>
<tr><td>Other (Explain)</td></tr>
<tr><td colspan="6">Representation (circle one) Vertical Non-Vertical Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
<tr><td>MERCY</td><td colspan="5">(Move to True/False items)</td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

WP6

Response: _____ **NR_____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
_____ORQSS					
_____Counting					
_____Ad Hoc					
_____Recalled Fact					
_____Standard Algorithm Choose: _____Addition _____Subtraction _____Multiplication					
_____Other (Explain):					
DNS (Explain)					
MERCY	(Move to True/False items)				

*Check a box when a major category is observed (check all that apply).

WP7

Response: _____ **NR____**

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
_____ORQSS					
_____Counting					
_____Ad Hoc					
_____Recalled Fact					
_____Standard Algorithm Choose: _____Addition _____Subtraction _____Multiplication					
_____Other (Explain):					
DNS (Explain)					
MERCY	(Move to True/False items)				

*Check a box when a major category is observed (check all that apply).

Equations: True/False

For each of the items in this section of the interview, the interviewer should show the student the equation and ask the student to read each equation aloud. The interviewer should ensure that the student has correctly read the equation exactly as it is written, from left to right. If the student is unable to do so after two tries, the interviewer should read it to the student and then ask the student to read it back.

After the student correctly reads each equation, the interviewer asks, "Is that equation true or not true?"

After the student tells if it is true or not true, the interviewer asks, "What makes this equation [true or not true]?"

For this part, I am going to ask you a few questions about equations or number sentences. For these, I will first ask you to read the equation out loud to me. Then, if the equation is correct as it is written, I want you to say 'The equation is True.' If the equation is not correct as it is written, I want you to say 'The equation is Not True' or 'The equation is False.'

After you tell me whether the equation is true or not true, I am going to ask you to tell me how you decided on your answer. For these problems, I am going to use the folder again so I can cover up the ones we haven't done yet.

Are you ready?"

TF1

Please read this equation just as it is written. *[Pause for student to read equation.]*

Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True	
	Other (Explain):
Not True	
	Other (Explain):
DNS (Explain)	

TF2

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True		
	Other (Explain):	
Not True		
	Other (Explain):	
DNS (Explain)		

TF3

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True	
	Other (Explain):
Not True	
	Other (Explain):
DNS (Explain)	

TF4

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True		
	Other (Explain):	
Not True			
	Other (Explain):	
DNS (Explain)		

TF5

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF6

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF7

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

NR_____

What makes this equation [true or not true]?

True			
	Other (Explain):		
Not True			
	Other (Explain):		
DNS (Explain)			

TF8

Please read this equation just as it is written. *[Pause for student to read equation.]*
Is that equation true or not true?

[redacted] **NR____**

What makes this equation [true or not true]?

True	[redacted]	
	Other (Explain):	
Not True	[redacted]	[redacted]
	[redacted]	
	Other (Explain):	
DNS (Explain)		

Multi-Digit Computation:

For this section the interviewer should read each equation to the student exactly as it is written. The subtraction symbol should be read as "minus," and the addition symbol should be read as "plus." The equal sign should be read as "equals," and for the blank line the interviewer should read it as, "what number."

Below are two examples of how each equation should be read.

[ ] + [ ] = [ ] *should be read as,* [ ] *plus what number equals* [ ]*."*

[ ] – [ ] = [ ] *should be read as, "what number minus* [ ] *equals* [ ]*."*

"For the next four problems, you are going to need the marker. I am going to read an equation aloud to you, and your job is to write the number that goes in the box to make the equation correct. To solve these problems, you can use the base ten blocks, snap cubes, marker and paper, fingers, or you can solve it mentally with just your brain. But no matter how you solve it, I want you to write the number in the box that makes the equation correct.

Are you ready to do that?"

MDC1

NR_____

How did you get [say the student's answer]?

<table>
<tr><th>Tools</th><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="2">Incrementing</td><td colspan="3" rowspan="2">Compensation</td></tr>
<tr><td colspan="2">Place Value Explanation</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td rowspan="2">Add</td><td rowspan="2">Counting</td><td colspan="4">Buggy Algorithm</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td></td><td></td><td></td><td>Other (Explain)</td></tr>
<tr><td colspan="6">Representation (circle one) Vertical Non-Vertical Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

MDC2

[redacted]

NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">Incrementing
[redacted]</td><td colspan="2">Combining Tens and Ones
[redacted]</td></tr>
<tr><td colspan="3">Compensation
[redacted]</td><td colspan="2">[redacted]</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="3">Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="3" rowspan="2">Buggy Algorithm
[redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td>Other (Explain)</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td colspan="3">Vertical Non-Vertical Mental</td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

MDC3

[redacted]

NR_____

How did you get [say the student's answer]?

Tools	Fingers	Base Ten	Snap Cubes	Drawings	Other
_____[redacted]					
_____[redacted]					
_____[redacted]					
_____[redacted]					
_____Other (Explain)					
DNS (Explain)					

MDC4 [redacted]

NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td>Fingers</td><td>Base Ten</td><td>Snap Cubes</td><td>Drawings</td><td>Other</td></tr>
<tr><td rowspan="2">ORQSS</td><td>Join All</td><td>Join To</td><td>Separate From</td><td>Separate To</td><td>Matching</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td>Counting</td><td colspan="5">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">Incrementing [redacted]</td><td colspan="2">Combining Tens and Ones [redacted]</td></tr>
<tr><td colspan="3">Compensation [redacted]</td><td colspan="2">[redacted]</td></tr>
<tr><td colspan="5">Other (Explain):</td></tr>
<tr><td rowspan="4">Standard Algorithm</td><td rowspan="2">Add</td><td rowspan="2">Counting</td><td colspan="3">Buggy Algorithm</td><td rowspan="2">[redacted]</td></tr>
<tr><td>[redacted]</td><td>[redacted]</td><td>[redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td>[redacted]</td><td>[redacted]</td><td>[redacted]</td><td>Other (Explain)</td></tr>
<tr><td colspan="2">Representation (circle one)</td><td>Vertical</td><td>Non-Vertical</td><td>Mental</td><td></td></tr>
<tr><td>DNS (Explain)</td><td colspan="5"></td></tr>
</table>

___ Additive ___ Subtractive

___Take-away Model
___Difference Model
___Other

We only have one more problem. For this last one, I want to see if you can solve it without using any of the manipulatives or the marker. Are you ready for the last problem?

MDC5

[redacted]

NR_____

How did you get [say the student's answer]?

<table>
<tr><td>Tools</td><td colspan="3">Fingers</td><td colspan="4">Other</td></tr>
<tr><td>Counting</td><td colspan="7">Count forward/backward from _____ to _____ by _____</td></tr>
<tr><td rowspan="3">Ad Hoc</td><td colspan="3">Combining Tens and Ones [redacted]</td><td colspan="4">Incrementing [redacted]</td></tr>
<tr><td colspan="7">Compensation [redacted]</td></tr>
<tr><td colspan="7">Other (Explain):</td></tr>
<tr><td rowspan="2">Mental Standard Algorithm</td><td>Add</td><td>Counting</td><td colspan="5">Buggy Algorithm [redacted]</td></tr>
<tr><td>Sub</td><td>Fact Recall</td><td colspan="4">[redacted]</td><td>Other (Explain)</td></tr>
<tr><td>DNS (Explain)</td><td colspan="7"></td></tr>
</table>

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—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 recorded 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 under the substrategy *separate from*.

When the student used an *ORQSS*-type strategy, we used the following names for 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 in *Children's Mathematics: Cognitively Guided Instruction* (Carpenter et al., 1999). Additional information on how the student counted the set representing the answer was also recorded.

ORQSS Fingers—When students determined solutions to the Number Fact items by employing a *direct modeling* (Carpenter et al. 1999) strategy using their fingers as the tools, we coded the strategy as *ORQSS ringers.* Because the students did not have access to manipulatives for the Number Fact items, we did not initially include *ORQSS* as an option within the coding scheme, but because of the higher than anticipated number of students using this strategy during the pilot interviews, we decided to include this option.

Counting—We used the *counting* code when the student employed a strategy where at least one of the quantities in the problem was not represented physically. For these items, we used the coding descriptions developed by Carpenter et al. (2015), which include *counting all, counting on from first, counting on from larger, counting on to, counting down,* and *counting down to.* For a full description of these codes, see Carpenter et al. (2015).

Recalled Fact—When the student stated that the answer to the problem was recalled from memory, we code the strategy as *recalled fact.* Examples could include the fact presented or an application of the commutative property. We also used this code for those students who recalled an addition fact to solve a subtraction problem, such as using the knowledge that [redacted] to solve [redacted].

Derived Fact—When the student stated that the answer was derived from another known fact, we coded the strategy as *derived fact. Derived facts* are used when the student combines known quantities when a specific fact is not known at a recall level. An example would be a student's first decomposing one of the addends to determine a sum of ten and then adding 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," historically applied to these strategies because they were not included among those that students were instructed to use. Over the past few decades, these strategies have been added to textbooks (including those used in the schools in our analytic sample) and are now taught directly to students by teachers (who also expect students to know the strategies by name). As a result, the boundary between invented and instructed strategies may no longer be clear. We therefore did not assume or attempt to determine whether the strategy was invented or instructed.

Within the *ad hoc* strategy, we coded (where applicable) whether students used 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 are 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
- The following so-called buggy algorithm applications when the student used an incorrect variation of the algorithm:
    - 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 anything to be zero
    - "Borrowed" without subtracting adjacent ten

We also noticed use by a number of students during the pilot testing of alternate notations of the standard algorithm. These were typically done mentally or when the student attempted to perform the steps of the United States standard algorithm without rewriting in vertical notation. Below are the criteria provided to interviewers about how to code an individual student's response.

Standard Algorithm (Standard Notation):

- The problem is rewritten vertically, aligned by place value.
- The student works right to left, and each result is recorded below on the same line, within each column (allow for errors).
- Each step attempts to obtain the same type of result (sum/difference) from each place value at a time.
- Regrouping/borrowing occurs when necessary in correct solutions. Incorrect solutions allow for misuse (or omission) of regrouping principles.

Standard Algorithm (Alternate Notation: Mental or Nonvertical):

- The student works right to left, aligned by place value. Concurrent results do not change previous results (i.e., the result from the tens column does not change the result from the ones column).
- Each step attempts to obtain the same type of result (sum/difference) from each place value at a time.
- Regrouping/borrowing occurs when necessary in correct solutions. Incorrect solutions allow for misuse (or omission) of regrouping principles. Regrouping/borrowing is either explicitly stated or implied though what they are saying or doing.
- Where the result of an addition problem involves regrouping, students who merely add the full sum of the ones to the sum of the tens violate the criteria for working with only one place value at a time and are not employing an application of the standard algorithm.

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

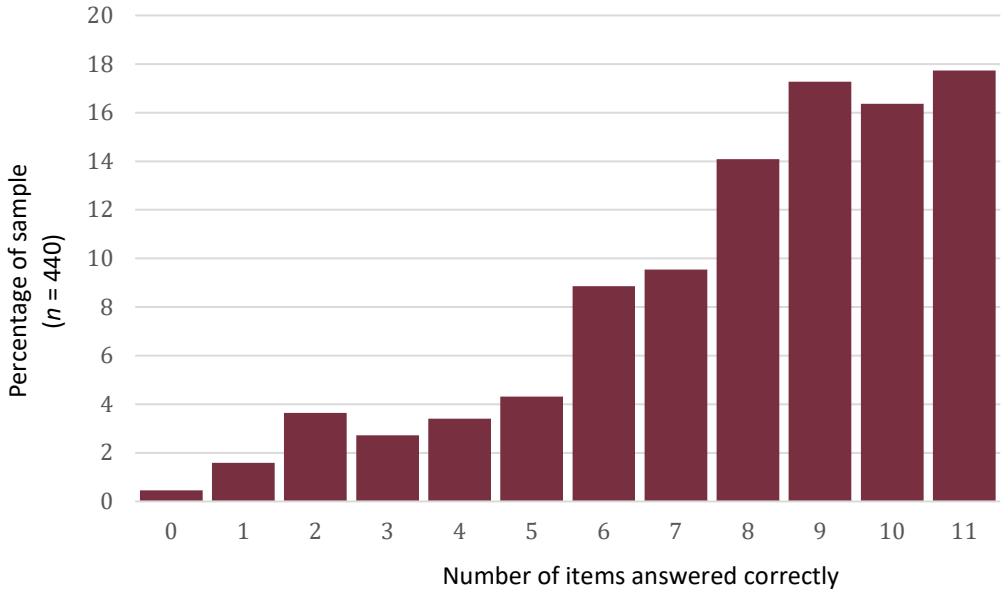


Figure 12. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Number Facts factor.

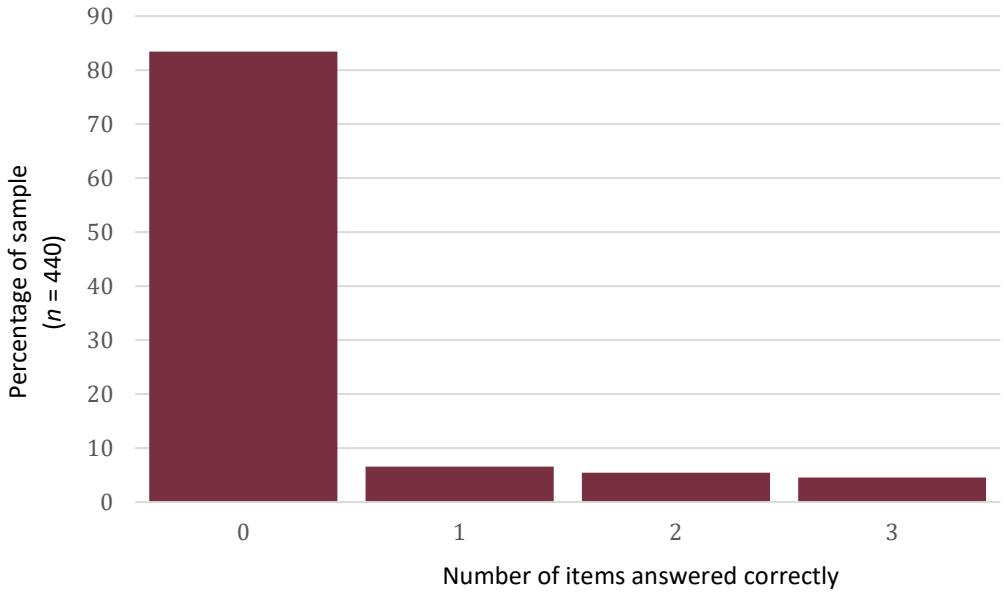


Figure 13. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Operations on Both Sides of the Equal sign factor

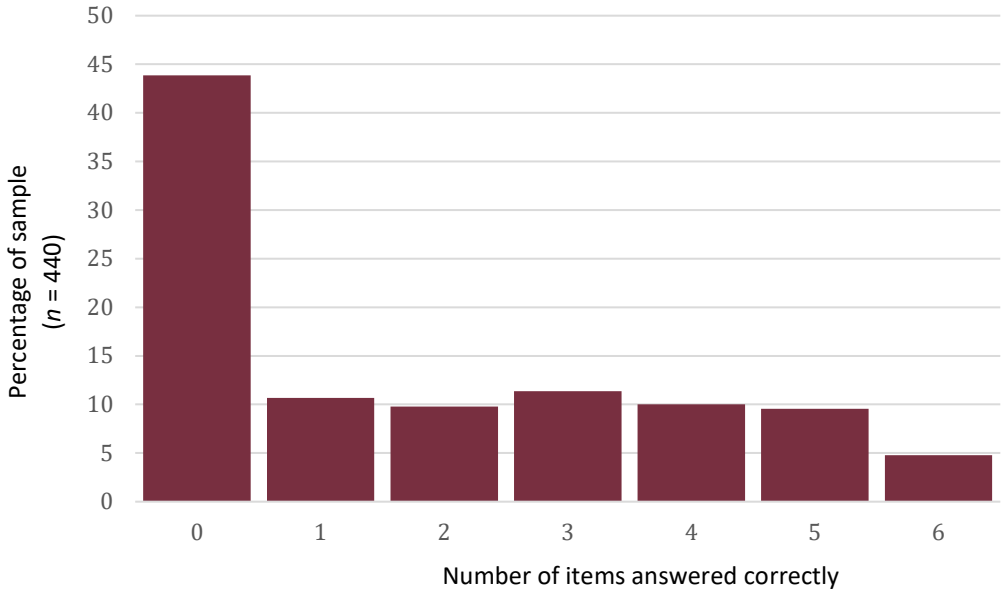


Figure 14. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Word Problems factor.

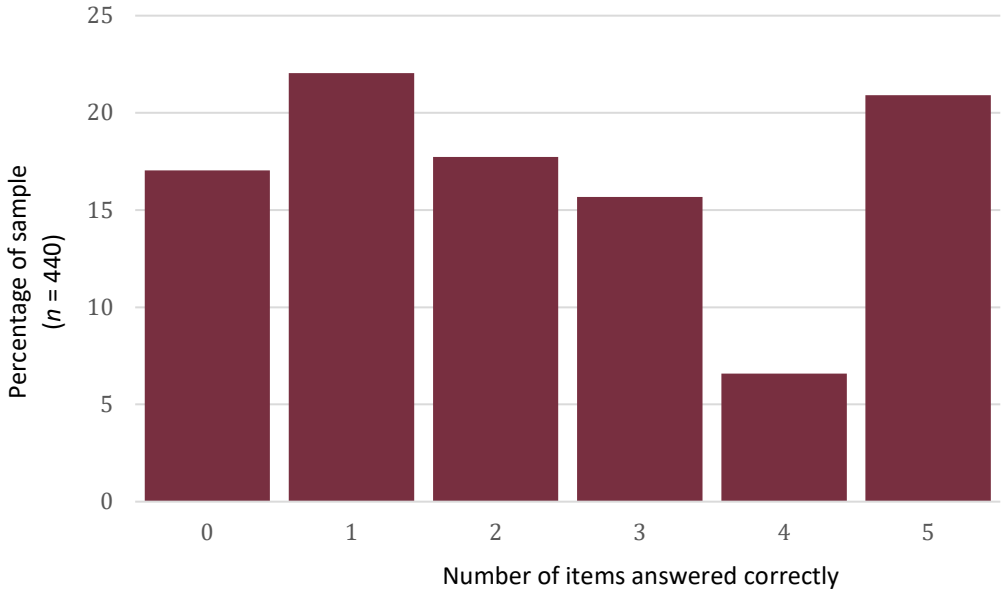


Figure 15. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Equal sign as a Relational Symbol factor.

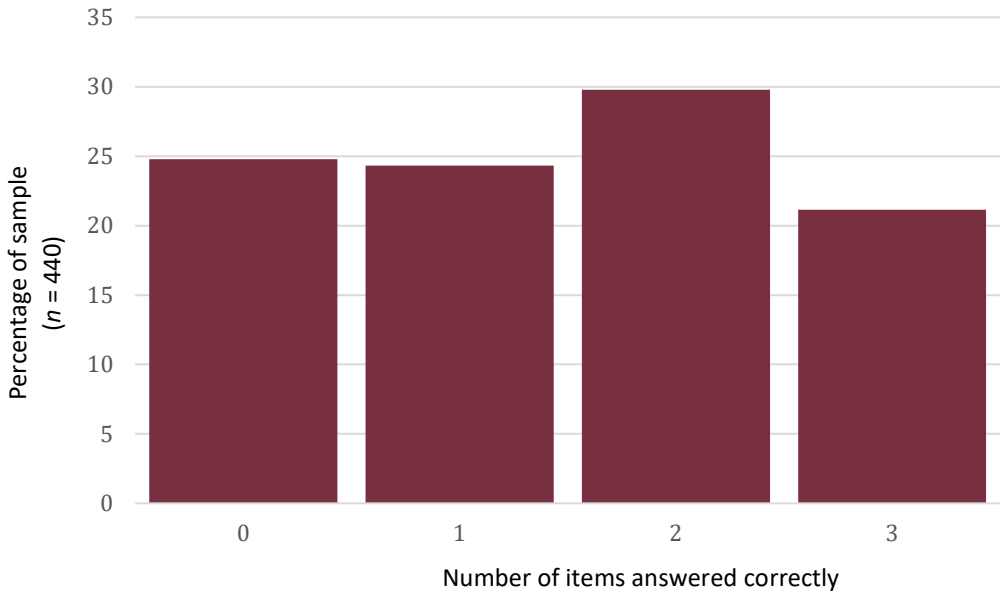


Figure 16. Distribution of the numbers of items individual students in the Grade 1 sample answered correctly within the Computation factor.

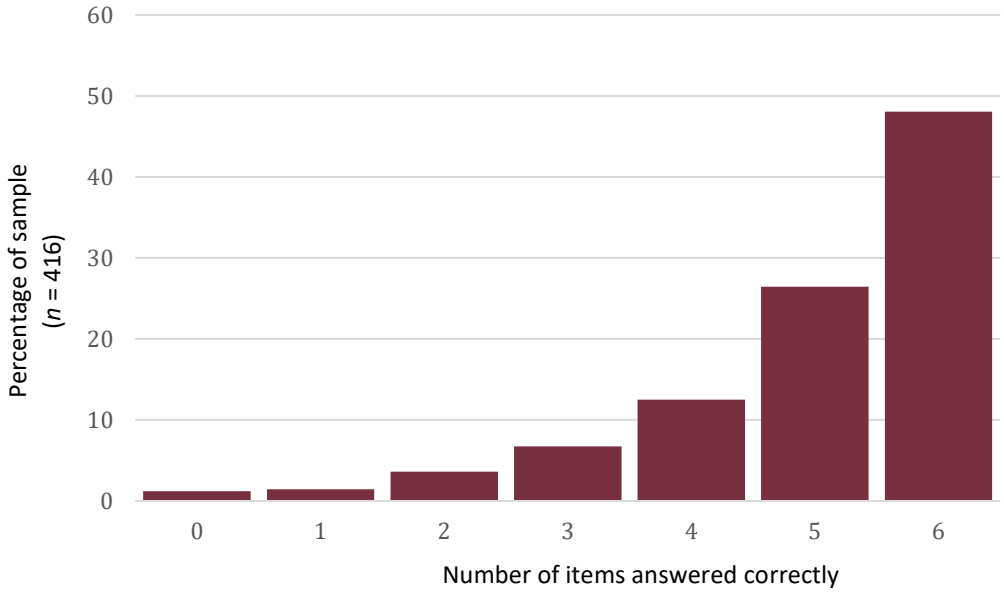


Figure 17. Distribution of the number of items individual students in the Grade 2 sample answered correctly within the Number Facts factor.

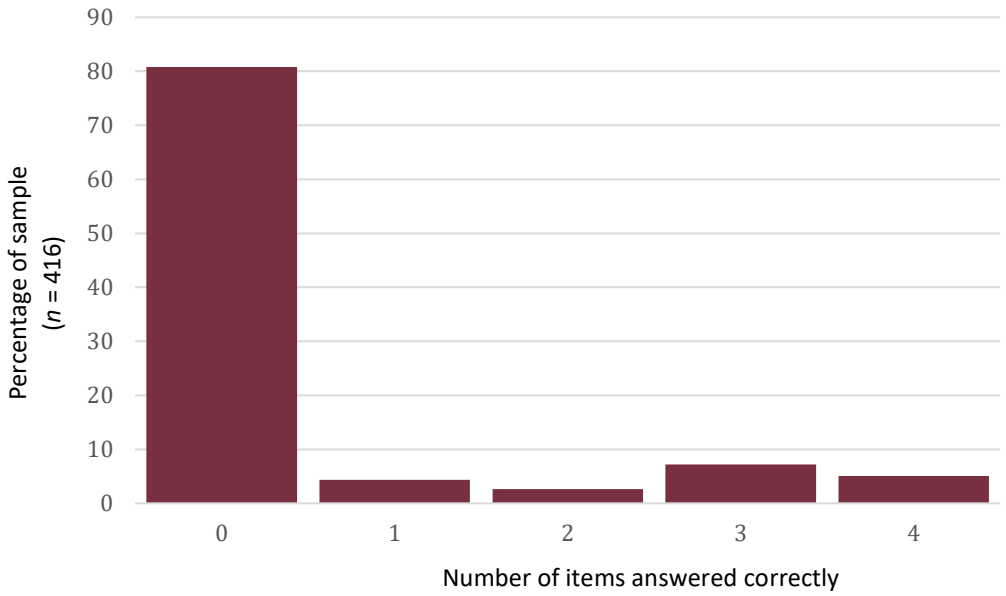


Figure 18. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Operations on Both Sides of the Equal sign factor.

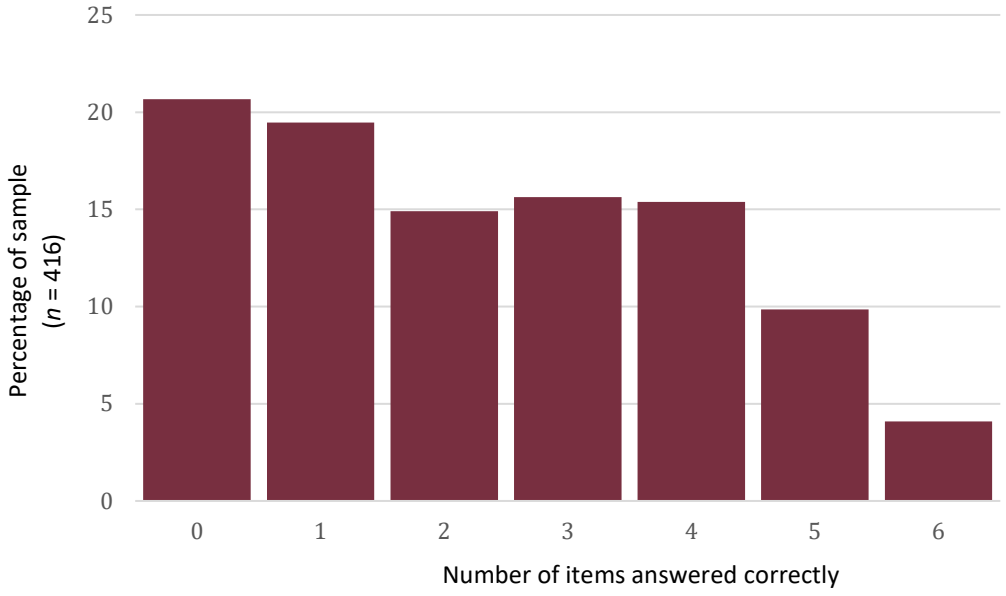


Figure 19. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Word Problems factor.

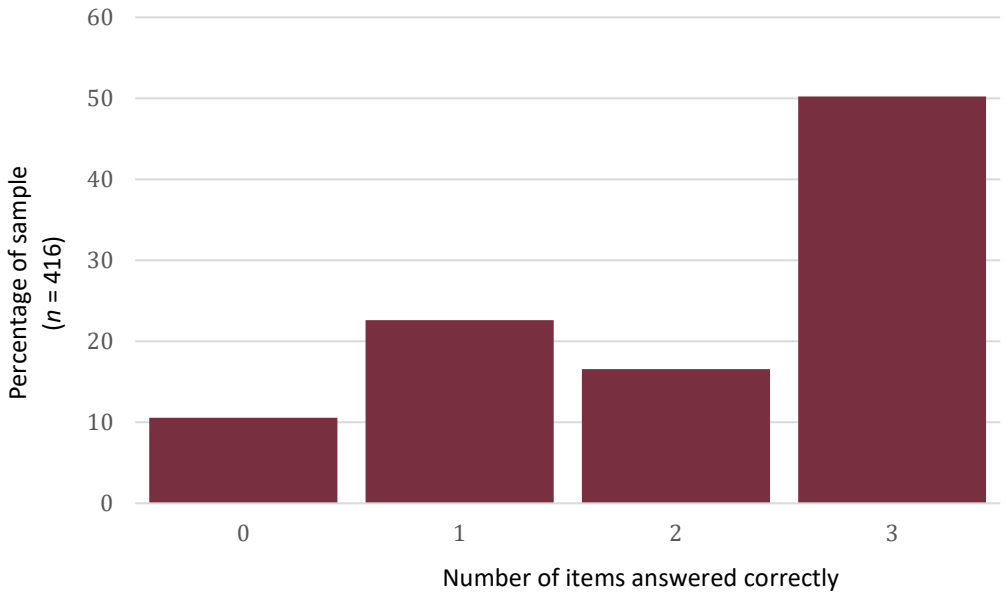


Figure 20. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Equal sign as a Relational Symbol factor.

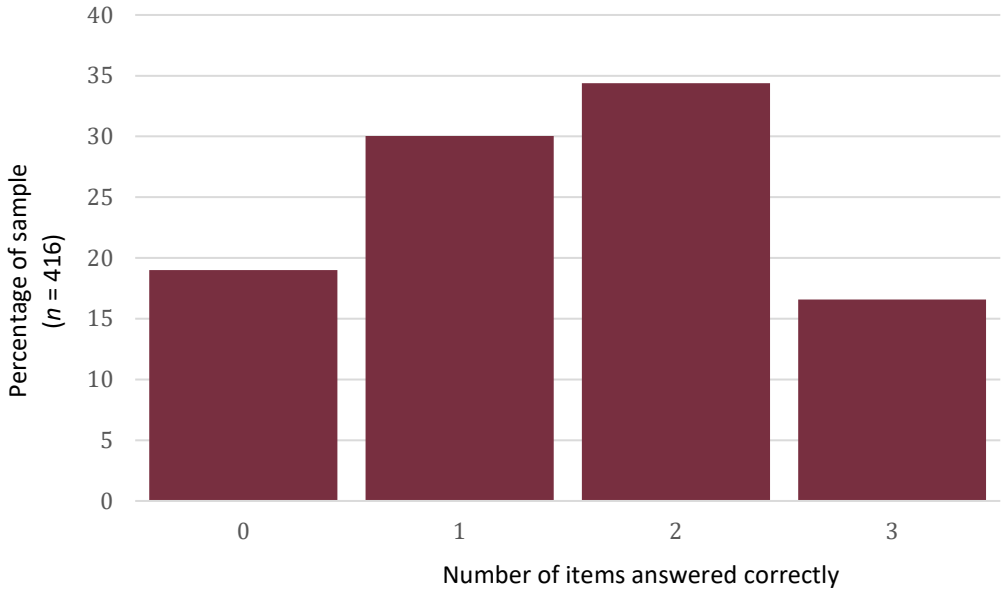


Figure 21. Distribution of the numbers of items individual students in the Grade 2 sample answered correctly within the Computation factor.

Appendix G—Most Common Student Responses by Item

Table 28. Proportion of Grade 1 Student Responses by Item

			Correct response	Most common incorrect responses			
Item	Factor	Item description	Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
NF1	NF		(.98)	(<.01)	(<.01)	(<.01)	(<.10)
NF2	NF		(.81)	(.04)	(.04)	DNS (.03)	(.02)
NF3	NF		(.83)	(.04)	(.03)	(.03)	(.02)
NF4	NF		(.76)	(.05)	(.04)	(.02)	(.02)
NF5	NF		(.75)	DNS (.090	(.04)	(.03)	(.02)
NF6	NF		(.93)	(.02)	(.02)	(.02)	(.01)
NF7	NF		(.63)	DNS (.10)	(.08)	(.05)	(.04)
NF8	NF		(.56)	(.09)	(.07)	DNS (.07)	(.06)
NF9	NF		(.48)	DNS (.13)	(.10)	(.05)	(.05)
NF10	NF		(.56)	DNS (.17)	(.07)	(.06)	(.02)
SE1	NF		(.69)	(.07)	DNS (.06)	(.03)	(.03)
SE2	OBS		(.15)	(.36)	(.14)	DNS (.13)	(.04)
SE3	OBS		(.09)	(.41)	(.21)	DNS (.07)	(.05)
SE4	OBS		(.07)	DNS (.43)	(.13)	(.04)	(.03)
SE5	—		(.06)	DNS (.72)	(.05)	(.02)	(.02)
WP1	—		(.74)	(.08)	(.05)	(.02)	(.02)
WP2	WP		(.32)	(.31)	(.04)	(.04)	DNS (.04)
WP3	WP		(.41)	(.20)	(.10)	DNS (.06)	(.03)
WP4	WP		(.33)	(.16)	DNS (.23)	(.06)	(.04)
WP5	WP		(.34)	DNS (.34)	(.05)	(.04)	(.03)
WP6	WP		(.26)	DNS (.47)	(.03)	(.03)	(.03)
WP7	WP		(.15)	DNS (.53)	(.08)	(.02)	(.02)
TF1	—		True (.93)	False (.06)	DNS (.01)		
TF2	—		False (.97)	True (.01)	DNS (.02)		
TF3	ESRS		True (.68)	False (.29)	DNS (.03)		
TF4	ESRS		True (.53)	False (.44)	DNS (.03)		
TF5	ESRS		True (.51)	False (.45)	DNS (.04)		
TF6	—		False (.92)	True (.06)	DNS (.02)		
TF7	ESRS		True (.30)	False (.62)	DNS (.08)		
TF8	ESRS		True (.33)	False (.63)	DNS (.04)		
MDC1	COMP		(.51)	(.15)	DNS (.06)	(.05)	(.03)
MDC2	COMP		(.43)	(.10)	DNS (.07)	(.06)	(.04)
MDC3	—		(.27)	DNS (.13)	(.09)	(.04)	(.04)
MDC4	COMP		(.54)	DNS (.12)	(.03)	(.02)	(.02)

Table 29. Proportion of Grade 2 Student Responses by Item

			Correct response	Most common incorrect responses			
Item	Factor	Item description	Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
NF1	—		(<1.00)	DNS (<.01)			
NF2	—		(.94)	(.02)	(.02)	(<.01)	(<.01)
NF3	—		(.94)	(.03)	(.02)	(.01)	(<.01)
NF4	—		(.93)	(.02)	(.01)	(.01)	(.01)
NF5	NF		(.93)	(.02)	(.01)	(.01)	(.01)
NF6	—		(.99)	(<.01)	(<.01)	DNS (<.01)	
NF7	NF		(.85)	(.02)	(.02)	DNS (.02)	(.02)
NF8	NF		(.75)	(.12)	DNS (.03)	(.02)	(.02)
NF9	NF		(.77)	(.07)	DNS (.07)	(.02)	(.02)
NF10	NF		(.83)	(.03)	(.03)	DNS (.04)	(.01)
SE1	NF		(.87)	(.04)	(.02)	DNS (.02)	(.01)
SE2	OBS		(.19)	(.45)	(.22)	DNS (.08)	(.01)
SE3	OBS		(.14)	(.48)	(.28)	(.02)	(.02)
SE4	OBS		(.08)	(.27)	DNS (.18)	(.13)	(.04)
SE5	OBS		(.11)	DNS (.62)	(.10)	(.03)	(.03)
WP1	—		(.84)	(.05)	(.02)	(.01)	(.01)
WP2	WP		(.69)	(.13)	(.03)	(.03)	(.03)
WP3	WP		(.38)	(.12)	DNS (.04)	(.03)	(.02)
WP4	WP		(.42)	DNS (.08)	(.08)	(.07)	(.07)
WP5	WP		(.41)	DNS (.19)	(.09)	(.06)	(.04)
WP6	WP		(.31)	DNS (.38)	(.05)	(.03)	(.02)
WP7	WP		(.11)	DNS (.45)	(.03)	(.03)	(.02)
TF1	—		True (.98)	False (.01)	DNS (.01)		
TF2	—		False (.98)	True (.01)	DNS (.01)		
TF3	ESRS		True (.79)	False (.18)	DNS (.03)		
TF4	ESRS		True (.64)	False (.34)	DNS (.02)		
TF5	ESRS		True (.63)	False (.36)	DNS (.01)		
TF6	—		False (.97)	True (.02)	DNS (.01)		
TF7	ESRS		True (.37)	False (.58)	DNS (.05)		
TF8	ESRS		True (.41)	False (.56)	DNS (.03)		
MDC1	COMP		(.70)	(.09)	(.03)	(.02)	DNS (.02)
MDC2	—		(.63)	(.08)	(.07)	(.04)	(.02)
MDC3	—		(.47)	DNS (.06)	(.06)	(.06)	(.05)
MDC4	COMP		(.24)	(.13)	DNS (.11)	(.06)	(.04)
MDC5	COMP		(.55)	(.06)	(.04)	(.04)	(.04)

Appendix H – 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. *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., 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., & 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.

Carpenter, T. P., & Moser, J. M. (1979). *An Investigation of the Learning of Addition and Subtraction* (Theoretical paper No. 79). Madison, Wisconsin: Wisconsin Research and Development Center for Individualized Schooling.

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

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.

Discovery Education Assessment (2010). Discovery Education's Common Core Mathematics Grade 1 and Grade 2 Interim Benchmark Assessment. Silver Spring, MD: Discovery Education.

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.* 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.* University of Wisconsin-Madison, Master's paper.

Koehler, J. L. (2004). *Learning to Think Relationally: Thinking Relationally to Learn.* University of Wisconsin—Madison.

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*, 361–371.

Light, G. S. (1980). = [equals 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., and 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.

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

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). CGI student achievement in region VI evaluation findings. *Newsletter of the Comprehensive Center-Region VI, 5*(2), Fall.

Tamburino, J. L. (1980). *An Analysis of the Modeling Processes Used by Kindergarten Children in Solving Simple Addition and Subtraction Story Problems.* Master's thesis, University of Pittsburgh.

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.* Doctoral dissertation, Utah State University.