DESIGN OF COMPUTER-BASED ASSESSMENT SECONDARY EDUCATION FOR UNDERSTANDING OF MATHEMATICS

by

Mark Damian Lewis

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Education in Curriculum and Instruction

Boise State University

August 2010

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Mark Damian Lewis

Thesis Title: Design of Computer-Based Assessment Secondary Education for Understanding of Mathematics

Date of Final Oral Examination: 17 May 2010

The following individuals read and discussed the thesis submitted by student Mark Damian Lewis, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Jonathan L. Brendefur, Ph.D.	Chair, Supervisory Committee
KeithW. Thiede, Ph.D.	Member, Supervisory Committee
Chareen Snelson, Ed.D.	Member, Supervisory Committee
Cheryl A. Torrez, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Jonathan L. Brendefur, Ph.D., Chair of the Supervisory Committee. The thesis was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

ABSTRACT

The purpose of this study was to evaluate the effectiveness of a computer-based assessment to reveal mathematical understanding. Relevant literature suggested that developments in cognitive science and computer-based assessments could allow the outcomes of cognitively guided instruction to be made explicit. An assessment instrument designed to make mathematical thinking explicit was developed and administered, consisting of 15 animations showing the solutions of one and two digit multiplication problems. A consistent set of five questions followed each animation. The assessment was administered to four classes of fourth grade students in two elementary schools participating in cognitively guided instruction professional development programs.

Findings showed that students, individually and as a group, preferred a limited and consistent set of strategies to solve problems and that some students may have developed increased understanding of a problem over the course of the five questions. Results also showed that the group was weakest on the concept of place value, but was able to apply strategies appropriate to particular problems. Correlations between the data from different questions suggest students vary in their understanding of components of the proposed construct of multiplication, which might otherwise be viewed as a unitary concept. Individual student strengths and weaknesses could not be determined because of the data's low reliability quotients.

For open-ended questions, smaller amounts of information in responses seemed to equate to lower levels of understanding. The assessment revealed possible instructional strategies at the group level, but refinement of the assessment will be necessary before individual student abilities can be reliably assessed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: NEED FOR THE STUDY

This dissertation details the background, creation, and implementation of a tool to study computer-based assessment of higher-order thinking in the field of mathematics, specifically multiplication for elementary-age students. A review of literature explored the intersection and history of three relevant areas: cognition, assessment, and computer-based assessment. The shaded area in Figure 1 represents the targeted knowledge that falls within the overlap of these topics. The dissertation describes the development of a new assessment tool and a methodology for evaluating this tool's efficacy. The study outlined here used technology in the form of personal computers in a 1:1 setting with students as an assessment tool to make students' thinking about mathematics explicit, which is one of the primary goals of cognitively guided instruction (CGI). The efficiency afforded by this tool allows teachers to continue instruction in a manner that best addresses student needs, whether at an individual or class level.

A desire to improve the assessment of higher-order thinking supports the increasingly cognitive orientation of instructional theory (Niemi, 1996) and aligns particularly well with the aims of cognitively guided instruction (CGI). CGI seeks to use students' own thinking processes to make their mathematical strategies and misconceptions known to themselves and their teachers. Once explicit, these processes indicate to students and teachers a path for further learning and instruction.

However, assessing those cognitive processes remains an elusive goal (Niemi, 1996). One premise of CGI is that existing knowledge, misconceptions, and the ability to use various problem-solving strategies vary among students, and that those different perspectives, when compared and combined, are the fertile grounds from which to better grow conceptual understanding. Locating students within the cognitive space of a given problem has proven to be time consuming, particularly when accomplished at a level to sufficiently reveal and individual student's needs. However, reducing the time allotted to that task may not give teachers enough detailed knowledge to provide individualized, student-centered instruction.

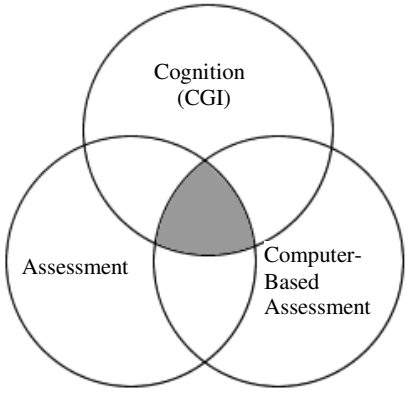


Figure 1. Relationship of knowledge base

Performing mathematics at the elementary-school level requires a number of cognitive abilities. Young students are expected to automatically recall basic

mathematical facts and strategies while also gaining a conceptual understanding that will serve as a foundation for further learning. Teaching early mathematics skills from purely algorithmic and memorization perspectives will not provide an adequate foundation for future learning (Hiebert et al., 1997). CGI research recognizes that children develop both general and domain-specific strategies for solving mathematics problems long before their formal mathematics education begins. CGI therefore seeks to build on those strategies rather than pretending that students' minds are, mathematically speaking, a blank slate (Carpenter, Fennema, Franke, Levi, & Empson, 1999).

Problem solving, another cognitive ability used in mathematics at this level, poses difficulties for assessment because it demands thinking at a higher level than that required by rote learning of algorithms. Problem solving is increasingly important as the availability of information increases in the information age. CGI uses word problems extensively because their translation into mathematical language invokes desired types of thought (Carpenter et al., 1999).

If students are expected to develop and demonstrate mathematical knowledge that goes beyond recall of facts and procedures, new types of assessments must be developed, assessments that allow and even encourage students to respond in ways that mirror or illuminate their cognitive processes. Acknowledging and identifying the highly individualistic nature of students' existing mathematical knowledge presents a challenge for assessment, whether administered by a teacher or a computer. Research has identified typical strategies used by young mathematics students when solving certain kinds of

problems, but cannot be sure to have found all possible strategies, especially those based on misconceptions. However, probable student responses for selected response items can be predicted from known strategies (Fuson, 2003).

Lipson, Faletti, and Martinez (1990) envision an assessment system that incorporates models and assesses important constructs and structures of mathematics; is responsive to individual student needs and knowledge; and paints a detailed picture of the state of a student's mathematical knowledge that is useful for instructor and student. Feedback from such an assessment will be more than a simple score. On a larger scale, compilation of individual student results would create a more complete picture of students' mathematical thinking, thereby improving the assessment system's capacity to assess individuals accurately. This cyclical feedback loop integrates mathematical knowledge and student performance, which stands in stark contrast to traditional mathematics assessments that separate the student from substantive knowledge and provide little useful feedback for the student (Kulm, 1990; Lipson et al., 1990).

Despite great promise, technology has thus far fallen short of its potential for educational use (Lesh & Kelly, 2000). Roschelle and Jackiw (2000) concur and offer the explanation that the application of technology often suffers from a problem similar to many other school reforms: the more effective it is in a given locale, the less likely it is to generalize to other situations (Glennan, Bodilly, Galegher, & Kerr, 2004). Too often, technology use lacks a solid pedagogical or theoretical background.

The purpose of this study is to evaluate a type of assessment tool that has been designed to use common, non-specialized technology to provide teachers and students with information not generally brought out by traditional assessments of mathematical performance. This information provides another measure of mathematical understanding and assists teachers' instructional efforts. The new assessment is not intended to replace performance-based, results-oriented tests of mathematical achievement but to complement them.

Delimitations of the Study

The participants in the study were a convenience sample. The students, their teachers, and schools are participants in the Developing Mathematical Thinking program run by the Initiative for Developing Mathematical Thinking (IDMT) at Boise State University, but the length of any individual's participation could not be determined. Previous mathematical training of students and their teachers was a possible factor in students' performance, but was not assessed or controlled in this study. Students not in class on testing day were not included in the study.

The assessment focused exclusively on multiplication of one and two digit positive integers. Generalizations beyond the classes tested and the subject matter of the assessment cannot be made. One section of the assessment (computer familiarity) relied on self-reporting. Relevant sections of the analysis contain additional limitations.

Research Questions

This study attempted to answer the following questions:

- What is an assessment instrument able to reveal about students' understanding of mathematic concepts related to multiplication of integers?
- What relationships are demonstrated between the results of the assessment for mathematical understanding and the assessment of demonstrated algorithmic proficiency in multiplication?
- What effect does computer familiarity have on the ability of the assessment to reveal mathematical thinking?

CHAPTER TWO: REVIEW OF LITERATURE

Three bodies of knowledge inform a computer-based effort to elicit students' thinking about mathematics: cognition and assessment, computer-based assessment, and cognitively guided instruction. The relevance to the current study and importance of each of those areas is explored below.

Cognition and Assessment

This section focuses on instructional purposes and uses of assessment as they relate to the goals of cognitively guided instruction. Although other purposes of assessment (accountability, promotion, etc.) are important, the effect of instruction cannot be overstated and must be a starting point for successful school improvement (Chappuis & Chappuis, 2002).

The terms cognition (how students think) and assessment (how can we tell what they know) were often thought of separately because of the pervasiveness of behaviorism in 20th century educational thinking (Driscoll, 2005; Saettler, 1990). Behaviorism, by definition, takes into account only stimuli and observable behaviors, omitting any explanation of cognitive mediation that connects those two end points. Successful learning from a behavioral perspective depends on observable behaviors and not on internal states (Driscoll, 2005). Mastery learning and programmed instruction are additional models of instruction built on behavioral principles (Joyce, Weil, & Calhoun,

2000). Standardized testing perpetuates the divide. Until about 20 to 30 years ago, tests were not considered capable of measuring how students actually thought; instead, tests settled for capturing the results of observable behaviors and gauging factual knowledge. However, recent advances in cognitive science and measurement have made possible the assessment of the thought processes that precede the observable results (Giordani & Soller, 2004; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002; Pellegrino, Chudowsky, & Glaser, 2001).

The current climate of accountability in public schools has resulted in greater amounts of standardized testing, which does not satisfy the call of some educators for more in-depth measurements of student knowledge. Baker and O'Neil (2002) predicted this would hasten the convergence of technology, assessment, and instruction. However, there is a more fundamental reason for such forms of assessment: the major phases of the teaching and learning process (curriculum, instruction, and assessment) function best when they are aligned with each other (English, 2000). Nearly 20 years ago, Lesh (1990) stated that assessment must be an integral part of the instructional and curricular process. Roschelle and Jackiw (2000) found justification for the addition of technology to the assessment process in the philosophies of leading 20th century educational thinkers. Links among cognition, instruction, and assessment can be found in the work of the most prominent educational thinkers of the last one hundred years. Piaget's theory of cognitive development, in which children progress from concrete to abstract thinking (Driscoll, 2005; Piaget, 1969), supports moving from using real manipulables to virtual ones. The

rich environments that technology and its multiple forms of representation provide create a rich Vygotskian environment where artifacts become tools that promote learning (Vygotsky, 1978). Student-centered environments that value the perspective of each student would please Dewey (1960).

Aligning assessment practices with the philosophies of past educational thinkers would be an academic endeavor, but change in such a large facet of contemporary education must have some rationale based on the reality today's students will face. Recent emphases on constructivist and social learning theories have created a need for assessments to do more than rate observable behaviors. The content and organization (schemas) of long-term memory provide clues about how people solve problems, a skill seen as increasingly important in today's knowledge-based society (Pellegrino et al., 2001). Assessments that make students' thinking explicit benefit learners and teachers and shifts the role of the teacher to that of a facilitator, which is in keeping with today's knowledge-based society (Lesh, Hoover, Hole, Kelly, & Post, 2000).

The National Research Council (NRC) also recommends that all levels of assessments, from informal classroom assessments to state- and nation-wide standardized tests, "work together in a system that is comprehensive, coherent, and continuous" (Pellegrino et al., 2001, p. 9). Although tests on various levels serve different purposes and therefore require different evidence, such a goal perhaps places the greatest burden on large-scale assessments if they are to become capable of eliciting knowledge at a deeper level than most currently do.

The purposes of assessment, stated or not, have become increasingly numerous: promotion, graduation, accountability, motivation, planning instruction at all levels (individual student, class, school), and making cognitive processes explicit (Airasian, 2005; Walvoord, 2004). This last purpose emphasizes the shift toward constructivism. With the recognition that student representations of knowledge differ comes the realization that their thought processes need to be made explicit and taken into account for purposes of instruction and assessment. No single testing method serves all of these purposes, but an emphasis on accountability has reduced the relative importance of the other purposes in schools today (Baker & Mayer, 1999). Increasing assessments' capacity to serve instruction will be difficult, because changing the mindset of the general public about what testing should look like and accomplish is a difficult task (Schacter, Herl, Chung, Dennis, & O'Neil Jr., 1999).

Learning About Student Thinking

A great leap of faith in cognitive science is that assessments actually can do more than capture behaviors; they can reveal what and how students think. Lesh and Lehrer (2000) believe this to be true, certainly in mathematics, and that this has benefits for both teachers and students. Such assessments align well with constructivist and student-centered learning, including CGI. However, information about thinking processes is usually not gathered because it is difficult to accomplish. Simply defining and describing these processes is difficult. The fact that knowledge structures vary from student to student and that those structures are dynamic, not static (Carpenter et al., 2004),

complicates the task. Lesh states that “there is no single right or wrong way to organize a system of ideas” (1990, p. 98), and McKnight (1990, p. 172) says that “higher order thinking, even in mathematics, is not a unitary phenomenon.” In addition, students do not posses component skills in equal proportions. Those good at following rules or applying models were not always the same students as those good at creating models (Lesh et al., 2000).

A common criticism of traditional assessments, whether paper and pencil or standardized multiple choice, is that they do not make explicit the cognitive processes behind student answers. Some newer assessments attempt to simply add the cognitive component to assessments that bear many similarities to traditional tests (Hoeft et al., 2003). Hoeft et al. took a different approach by attempting to make explicit the schemas a student has for a given topic and omitting the application and “answer” components of a traditional test. The tool they used was a concept map, which depicts only concepts and relationships.

Discerning thinking indicative of learning and not of innate ability is difficult. Assessments must require more than recall of rote learning, and the cognitive processes being assessed must be sensitive to instruction (Baker & Mayer, 1999). Lesh and Kelly (2000) state that one way to find out what students know is to teach them. Prior knowledge and misconceptions, if not formally assessed before instruction, will become apparent during instruction. Such formative assessment is often informal and made

through observations of answers to questions, questions asked, attention, and facial expressions (Airasian, 2005).

Models

Models and modeling are at the heart of mathematical understanding and expert application. Roschelle and Jackiw (2000) state that modeling merges the empirical and the theoretical: the reasons behind observable mathematical behavior. Determining the models used by students is therefore a critical step toward understanding their mathematical thinking. In the real world, the ability to develop models is more important than just being able to apply them, but most teachers and textbooks do not encourage it (Fuson, 2003).

Models, including those developed by students, should be at the heart of knowledge construction and instruction, student thinking, and assessment (Lesh & Lamon, 1992). Students create and rely on internal models to process and interpret incoming information, whether the models are up to the task or not (Fuson, 2003). This is important to recognize because it explains that misconception can reflect incomplete rather than incorrect learning. Models distill experience into reusable knowledge, and because humans tend to use models, they also tend to create them. Conditions that tend to promote model development include (a) situations in which predictions based on patterns must be made, (b) explanations or justifications are required, and (c) strategies of others must be analyzed. All these conditions are present in cognitively guided instruction.

Efforts to make students mathematize problems by converting everyday situations into mathematical terms and notation have been called “model eliciting” (Lesh et al., 2000). Creating model-eliciting assessments is the result of purposeful design, but before the test can be created, the content material must be modeled (Lipson et al., 1990; Martinez & Bennett, 1992; Mislevy et al., 2002; Pellegrino et al., 2001). For complex assessments, the plan for scoring must be determined during the design stage by using an evidence-centered approach that models required knowledge and skills, the tasks that elicit them, and levels for measuring how well one meets the other. This is similar to problem-based learning, where one begins with the end in mind (Boud & Feletti, 1997). However, model-eliciting assessments try to make in-depth student knowledge available to students and instructors.

The National Research Council (Pellegrino et al., 2001) underscored another reason for eliciting models when they state that CGI and assessments based on its principles can differentiate cognitive processes behind similar if not identical uses of algorithms. Two mathematical word problems requiring the same algorithm may require different initial strategies. Students’ selection of strategies may depend on the semantics of a word problem, making selection and execution of the algorithm dependent on cognitive processes that occur earlier in the solution process. Determining students’ capabilities requires understanding the processes that occur before the algorithm is used.

Model eliciting assessments discussed so far have pre-supposed students actively participate in authentic tasks that reveal their thinking. Giordani and Soller (2004),

however, found that when working at a computer in groups, elementary age students were more likely to express and articulate their ideas about solving the problem at hand when another student had control over the mouse. This suggests students may not actually have to actively solve a problem to elicit their thinking.

Of course, not all that is revealed about students' thinking will prove to be correct. Identifying misconceptions and false assumptions should be as much a goal of assessments as identifying what is correct (Lesh, 1990). Any instruction that does not make explicit, use, and correct misconceptions in existing knowledge is likely to result in fractured and incoherent learning combined with continuing misconceptions. When eliciting students' mathematical knowledge, students should be allowed to represent their current knowledge accurately and fully, not just the parts of it that align to traditional or "correct" mathematical thinking (Lesh et al., 2000; Yeh, 2001).

Problem Solving

An assessment whose purpose is making strategies and models explicit must be engaging students in problem-solving tasks. Mayer and Wittrock (1996) define problem solving as "cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver" (p. 47). That is, if the path to a solution is known from the start, then students are applying algorithms or recalling factual knowledge. Baker and Mayer (1999) state that "problem-solving performance can be a more challenging indicator of a student's understanding" (p. 271). They go on to state that problem solving and testing for depth of understanding involve creation of mental models, which have

been discussed as critical to understanding mathematical thinking. Their criteria for computer-based assessment of problem solving include cognitive complexity, meaning that items must require students to do more than just recall material. Examining cognitive modeling, problem solving, and the type of complexity that computer-based assessments can provide demonstrates how closely linked they are.

Transfer is relevant to the assessment of student thinking because problems that might appear quite different on the surface may require the same mathematical principles to solve. A child's ability to solve one problem based on a particular concept but not another problem based on the same concept might indicate rote learning of solutions and a lack of true understanding. A learner that is not able to apply what was learned from one problem to a similar problem might be thought of as developmentally incapable of abstract thought (Driscoll, 2005), but also might be thought of as a novice in the relevant domain: lacking in the conceptual depth required to determine similarities between the two problems (Gagne, Yekovich, & Yekovich, 1993).

Alignment

If stated instructional objectives and instruction promote problem solving, modeling, and conceptual understanding, then assessments must be capable of determining how well students have mastered those objectives (Stroup & Wilensky, 2000; Yeh, 2001). Teaching higher-order thinking but assessing rote performance yields data neither valid nor useful in planning further instruction. To reverse the logic as Maslow (1966, p. 15) did when he said "if the only tool you have is a hammer, [it is

tempting] to treat everything as if it were a nail," the nature of the assessment drives the instruction, so the assessment needs to test the core of the desired objectives. When the standard form of assessment is a multiple choice or short answer type test, that tends to dictate a pedagogical approach designed to yield the ability to produce the types of answers those tests require. An algorithmic approach to solving problems can, with enough repetition, produce the skills necessary to answer those types of questions. A standard argument against standardized testing – what gets tested, gets taught – extends beyond content into the pedagogy and philosophy of instruction. Limitations in the type of assessment come from many factors: the powerful effects of teachers' own school experiences (Richardson & Placier, 2001), teacher training, available time and testing technologies, and assessment formats of external, standardized tests. Improved methods of assessment must be a component of reform of instruction (Chappuis & Chappuis, 2002).

Results of assessment tasks that demand higher-order thinking are not represented well by a single score. Stated another way, a rich learning environment creates rich assessment data (Lesh & Lamon, 1992; Stroup & Wilensky, 2000). Items are needed to produce rich data and constrain student responses as little as possible.

Authenticity

Authentic assessment sounds appealing, but researchers disagree on its importance and even its definition. Yeh (2001) stated that even multiple choice tests can be authentic. However, most multiple choice test questions do not represent authentic

problems or experiences. Short answer mathematics problems bear little relation to students and their world; they are about factors external to the student (Lesh et al., 2000). Others discount the importance of authenticity by saying that it not the most important factor in designing an assessment; that providing the required evidence is more important (Mislevy et al., 2002). Lesh and Kelly (2000) argue the limitations of authentic assessment by stating that "most students' relevant knowledge seldom develops beyond primitive levels as long as their mathematical experiences are restricted to those that occur naturally in everyday settings" (p. 203). Still others feel authentic problem solving, especially in a technology-based environment, requires the same set of skills as the 21st century workplace (Schacter et al., 1999) and that this higher level of authenticity results in better validity (Martinez & Bennett, 1992).

Lesh and Lamon (1992) described characteristics of authentic assessment items: they take at least five minutes to complete, allow demonstration of individual student understandings, are more complex than answering a specific question, and allow multiple solution paths. However, most so-called authentic problems are not really authentic (Lesh et al., 2000). Their givens are very constricting, and they are derived from existing models. They are simply application problems. The ability to develop models is more important in the real world than just being able to apply models, and most training, assessments, and textbooks do not encourage this. There is a correlation between problems that are truly real-world and those that elicit model creating. Knowing how to apply models does not necessarily transfer to being able to create them. Both are certainly

important skills, but it is too easy to confuse them. A final note from Lesh et al. about authenticity is that attempting to ground problems in reality should not be an attempt to define students' realities. The worlds of reality and theory are not immutable or exclusive.

Authenticity is pertinent to the topic of this study because the goal is to elicit students' mathematical thinking, thinking derived from their informal mathematical experiences and learning in the real world as well as their formal education. Authentic problems have larger contexts and depend more on understanding than on rote learning and algorithms.

Summative or Formative?

The line between formative and summative assessment blurs as tasks become more authentic. Identifying existing cognitive processes is formative in that it provides guidance for further instruction, but if making students proficient in a particular manner of thinking is a goal of the instruction, then it also plays a summative assessment role. The ability of any assessment (and CBA in particular) to provide individualized and nearly instantaneous feedback and to record student progress blurs the lines not only between formative and summative assessment but also between instruction and assessment. These capabilities also increase opportunities for reflection, which Lesh and Kelly (2000) describe as the usual way to induce thinking about changes in knowledge. This is not to imply that technology is the answer for the entire instructional process, but

it can play a role in each step of the instructional process. As discussed elsewhere in this paper, teachers and technology are more effective when they work together.

Many summative assessments focus on narrow and comparatively unimportant portions of the content: the results from narrow, specific, and artificial problems (Lesh, 1990). These types of assessments do not measure generalizable procedural or conceptual knowledge that benefit students in further study or outside the classroom. The blurring between instruction and assessment and between summative and formative is inevitable as tasks—taught and measured—become more authentic. Difficulty defining or separating the two is more of a problem for researchers than for teachers in classrooms.

Interpreting Data and Results

A number of factors come into play when the complex data gathered from assessments that elicit models, strategies, and other forms of higher-order thinking must be analyzed. The greater depth of information they provide comes at a cost, which is breadth of knowledge. Such assessments tend to cover a limited number of concepts and also take more time (Lesh et al., 2000). Assessments eliciting these forms of knowledge must be able to capture and interpret intermediate steps in the solution process. To do this, probable student models of knowledge and solution paths must be pre-constructed. Lipson et al. (1990) envision an assessment system that incorporates and assesses important models and structures of math; that is responsive to individual student needs and knowledge; and that can paint a detailed picture of the state of a student's mathematical knowledge, which is useful for instruction and to the student. Feedback

from such an assessment would be much more than a simple score. On a larger scale, the state of each student's knowledge would inform the assessment system of how students tend to think and serve to create a more complete picture of how students think mathematically. This cyclical feedback loop integrates mathematical knowledge and student performance, which stands in stark contrast to traditional mathematical assessments that separate the student from substantive mathematical knowledge and provide little useful feedback for the student (Lesh et al., 2000).

Assessing Understanding

Determining whether or not students have arrived at a correct or defensible answer for a mathematics problem is a fairly objective judgment at the elementary level of mathematics. For example, there is only one reasonable answer that could be expected from an elementary school student given the problem 5 x 6 = ? This problem could be posed as either a selected- or constructed-type item. As previously noted, the focus of mathematical education has shifted from that type of judgment to understanding and developing the mathematical thinking of students. A constructed response item that requires only a final answer is not necessarily the best means of exposing the thinking and selection of strategies that get students from the start to the conclusion of a problem. Lesh et al. (2000) state that the type of problem should be dictated by the desired type of information. Problems requiring rational, finite answers are unlikely to elicit the freest thinking. Giordani and Soller (2004) go a step further by saying that students are most

likely to express their thoughts about a problem when they do not have control over the solution process.

As an example, McClain, Cobb, Gravemeijer, and Estes (1999) demonstrated how students in a first grade classroom benefited from having to explain their thinking, and showed how those types of activities blur the distinction between instruction or development and assessment, a distinction that may have a stronger basis in instructional design than in cognitive science. Yeh found that forcing students to express and defend their thinking improved their critical thinking because "they frequently realized the need to modify their claims and reasons, ultimately resulting in stronger arguments and improved reasoning" (2001, p. 16).

A synthesis of the above ideas indicates a path toward a method of eliciting students' mathematical thinking: remove the mathematical solution as the student output for the problem and relocate the process control away from the student. Student responses would be purely reactive and conceptually based. In a CBA environment, the computer is the logical source of control, and also the means of collecting student responses.

Computer-Based Assessment

New definitions of learning have refined purposes and methods of assessment, but the late 20th century saw another major innovation in assessment with the proliferation of personal computers. They have been used to conduct assessments since early on in their existence, but it is worth noting that the 1971 edition of the *Educational Measurement* handbook did not cover the topic of computerized testing (Thorndike, 1971). Although

technology is the assessment medium of the current study, the National Research Council (NRC) warns “technology will not in and of itself improve educational assessment.” However, the NRC does go on to state that technology can “enhance the linkages among cognition, observation, and interpretation” (Pellegrino et al., 2001, p. 9).

Baker and Mayer (1999) believe computer-based assessment (CBA), whose origins stem from a desire for efficiency, is the future of assessment; others believe the inevitable trends toward lower cost and ease of use might make technology a force for true change in education, which has resisted large scale change for so long (Baker & O'Neil Jr., 2002). Many researchers agree with this statement for a variety of reasons. One is the technological capacity to display information in multiple, more realistic representations by using photographs, animation, photographs, interactivity, and increased user control (Bransford, Brown, & Cocking, 2000; Lesh, 1990; Lipson et al., 1990; Pellegrino et al., 2001; Scalise & Gifford, 2006). These all provide better representations of the real world and embrace student variability. Beyond the possibilities for tests themselves, McKnight (1990) believes this also supports an increasing need for graphic literacy in our information-based society. In addition, modeling of the content to be tested and of student knowledge aligns with recent work in cognitive psychology. When well designed, CBA interfaces can and should be unobtrusive (Chung & Baker, 2003). This is a critical aspect of any test attempting to illuminate student thinking. Johnson and Green (2006) found students resort to mental calculation when doing so is

easier than working out the problem on paper. The same will likely hold true in other testing media, making ease of use a positive factor for eliciting thinking.

The most compelling reason for using CBA is its capacity to capture a more in-depth picture of student knowledge (Lipson et al., 1990; Mislevy, 2004; Pellegrino et al., 2001). CBA has primarily been used in mathematics and science but also to assess writing. It is the natural form of assessment to use with computer-assisted instruction (Lipson et al., 1990) and intelligent tutoring systems (Bransford et al., 2000), and can detect previously unknown patterns and relationships in student knowledge and performance (Bransford et al., 2000). CBA has the potential to perform this analysis much faster than can humans (Chung & Baker, 2003). In regard to accuracy of scoring, computer scoring of questions on the GRE was found to be highly correlated with human scoring, especially in algebra (Martinez & Bennett, 1992). Some newer methods of converting evidence from these types of assessments into usable information require technical skills in modeling and statistics that most educators do not possess. Technology can help bridge this gap. If the instruction is not preparing students for these types of tests, however, the tests will lack validity because they will not measure and support judgments about what was taught.

Computer-based assessments or components of them can be reusable (Baker & O'Neil Jr., 2002) or even generative. This is critical because such assessments require a lot of development time. Twenty years ago, developing a computer-based assessment took 200 times as long as the instruction for that assessment. Developing computer-based

assessment is still time intensive, though that ratio has since lessened (Anderson, Boyle, & Reiser, 1985). As domain specificity of an assessment increased, however, designing reusable tests or components becomes more difficult.

The design of computer-based assessments that are useful outside of their original setting encounters problems similar to research designs that must balance validity and reliability with generalizaibility. Measuring problem-solving ability in a specific domain requires a knowledge base and assessment formats that are difficult to apply in other domains or situations (Bransford et al., 2000). Such assessments must rely on analysis, identification, and use of non-domain specific knowledge, strategies, and assessment structures whenever possible if they are to become practical. Baker and O'Neil (2002) argue that only such careful analysis can produce computer-based assessments that are both valid and practical.

Previous Applications of Computer-Based Assessment

Many technology-based immersive, multimedia, and collaborative simulation and learning environments have positive effects on student learning, but most of these technologies did not originally incorporate assessment. The need for scaffolding and feedback fueled the integration of assessment, as did the need for accountability and documentation of effects of such systems. Such authentic and immersive environments aligned poorly with most standardized test formats, creating the need for assessments that could elicit the cognitive skills such environments endeavor to instill in students. Previous assessments using computer-based problem solving include domains as

divergent as architecture (Katz, Martinez, Sheehan, & Tatsuoka, 1993), and dental hygiene (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

One style of CBA relies on recording locations and effect of mouse clicks as students solve a problem. This is referred to as a click-through interface. A problem-solving environment using this type of interface was developed at UCLA in the early 1990s. The Interactive Multimedia Exercises (IMMEX) tracks what users click on in a non-restrictive environment. Each click provides more information, some relevant, some not; but also costs the student a bit of the currency provided for each solution attempt for that problem. The amount of currency provided is enough to provide students some freedom in deriving their solution path, but not enough to arrive at a solution by trial and error using every available option. Students must ultimately choose from a long list of possible final answers, and guesses also cost the student a bit of currency. The list of possible answers is long enough to discourage guessing. The recorded click stream and the final answer together provide ample evidence of how a student solved the problem ("IMMEX," 2007).

This interface was also used by Chung and Baker (2003) on college freshmen solving a design problem. All information and processes were available for students to select with a mouse click, and previous analysis of the domain allowed the researchers to capture assessment solution paths. This study demonstrated that although the amount of interaction was low, the analysis was cost and time efficient.

The ability of a click-through interface to capture the steps students take in their paths to a solution is not dependent on the level or content of the problems. Two students who both selected the correct answer in the end may have had very different solutions paths. Researchers have found strong correlations between desired cognitive processes and successful problem solutions (Chung, de Vries, Cheak, Stevens, & Bewley, 2002). That is, solutions paths pre-defined as desirable generally produced correct answers. Misconceptions do have to be deduced from the selection of answers; they are demonstrated by the students' selections that have been captured in the mouse-click data.

Concept maps are another tool that has been used to capture student representations of "a domain's conceptual structure" (Pellegrino et al., 2001, p. 265). In these activities, students do not actually solve problems but create a concept map to represent their thinking in a domain. One study using this method created a computer-scoring system (O'Neil Jr. & Klein, 1997). After an initial training period, the computer-based method was as effective as a pencil and paper version of the same task and also assessed collaborative skills.

In some cases of computer-based problem solving and assessment, two nearly separate systems handle the problem presentation and the assessment. The Adventures of Jasper Woodbury problem-solving series from the Cognition and Technology Group at Vanderbilt University (CTGV) is one such case. Assessment data and interpretation was handled by a separate, web-based program called Scientific and Mathematical Arenas for Refining Thinking (SMART). One feature of the SMART web site showed videos of

students explaining solutions that deliberately contained incorrect statements. The students watching the videos had to provide "feedback" to the students in the videos. This created cognitive dissonance in students who understood the problems correctly while also creating opportunities for them to demonstrate that knowledge.

Scaffolding and Feedback

Scaffolding is generally thought of as a component of instruction but is also present in assessment. Azevedo (2005) found scaffolding necessary for changes in thinking to occur. Technology enables scaffolding and feedback at a speed and in quantities not possible in teacher-mediated instruction, which increases a student's zone of proximal development (ZPD) (Vygotsky, 1978). Vygotsky's theory of ZPD describes the difference between what a child can do by him or herself and what the child can do with the assistance of an adult or more advanced peer. Vygotsky hypothesized that measuring just the former did not provide a full picture of the child's intelligence or learning. This relates to the concept of scaffolding because any assistance provided will be most effective if it is within the child's ZPD (Siegler & Alibali, 2005).

Despite Vygotsky's theories, most assessments attempt to limit scaffolding. Many selected response items provide clues that influence students' responses. While most questions in the assessment instrument developed for this study are selected responses, the responses do not contain mathematical content: most are of the yes or no variety.

Technology can also vary the rate and type of feedback (Baker & Mayer, 1999). This is particularly useful in formative assessment. Computers allow better detection of

previously unseen patterns and relationships in students' thinking. The positive effects of increased feedback in computer-based learning has been documented in intelligent tutoring systems (Pellegrino et al., 2001).

The NRC details a study comparing the effects of practice and feedback from a teacher to that of an intelligent tutor (Kulm, 1990). The teacher's feedback was more accurate than the intelligent tutor because the immediacy and volume of feedback from the intelligent tutor made the students' needs explicit. In this study, the teacher and intelligent tutor formed a complementary system able to meet the needs of the student better than either one alone. Another benefit of receiving feedback from another, non-judgmental source, such as a computer, might be to reduce the pressure a student feels to perform well (Lesh, 1990).

The socially constructivist nature of CGI and the scaffolding it offers create challenges for aligning assessment with instruction. This is compounded by increased opportunities for collaboration provided by technology (Bransford et al., 2000). The validity of supposedly authentic tasks is reduced by the absence of the social components present in the instruction and by the fact that problem solving in the real world is often a collaborative process.

Effects of Test Mode

Numerous researchers have compared computer-based to paper-and-pencil tests. A recent meta-analysis (Wang, Jiao, Young, Brooks, & Olson, 2007) of mathematics studies reviewed the results of three other previous meta-analysis studies (Bergstrom,

1992; Kim, 1999; Mead & Drasgow, 1993) that did not focus on mathematics and were largely targeted at adult learners or secondary students. None of these studies found a significant effect for test mode.

The meta-analysis by Wang et al. (2007) set stringent criteria for inclusion of studies. The samples had to be English speaking, drawn from K-12 classrooms, and have a minimum within-groups sample size of 25. The studies also had to present or have gathered the data necessary to calculate effect sizes. Finally, the studies had to directly compare results from the two tests modes. These criteria greatly reduced the number of studies in the meta-analysis from 312 after the initial literature review to 44. Most of the included studies were published in 2004. It may be worth noting the variables Wang et al. found did not have an effect when comparing computer-based and paper-and-pencil testing. These include study design, grade level, sample size, type of test, computer delivery method, and practice. The type of computerized test (linear versus computer-adaptive) was a significant factor, with linear tests showing greater differences in the comparisons to paper-and-pencil tests (PPT).

The study by Sandene et al. (2005) of the National Association of Educational Progress (NAEP) noted test mode effects were larger for constructed response items than for multiple choice items. However, Martinez and Bennett (1992) found computer scoring of algebra problems matched that of human scorers. This relied on a complete pre-evaluation of both correct and incorrect solutions, mirroring the process that goes into creating selected response items.

Differential Item Functioning

Differential item functioning (DIF), which occurs when groups perform differently on an item after controlling for ability (Gierl, 2004), was found in CBA by a number of studies. Studies by Johnson and Green (2006), and Poggio, Glasnapp, Young, and Poggio (2005) did not find a significant effect for test mode, but did note differences at the item level that they could not conclusively explain. Both offered question content as a possible factor. Gu, Drake, and Wolfe (2006) attempted to identify sources of DIF in tests of college students on quantitative items. Over one-third of the items (38%) showed DIF. Question content was again cited as a factor, as was the mathematical notation used. Page formatting and methods of responding to questions were not found to be factors. Items containing DIF-producing content were noted as easier or harder by mode, but no researcher attempted to relate the differences to content validity.

Pommerich (2004) maintains the purpose of research into test mode effects is to ensure that variability in scores is due to differences in content knowledge and not effects of testing on a computer. This would seem to assume that paper-and-pencil tests do not contain factors that contribute to test mode effects and that any effects found are due to deficiencies (or strengths) of CBT. The goals in development of CBT might better be defined as high degrees of validity and reliability, not a lack of mode effects when compared to another form of testing.

Computer Familiarity

Researchers have frequently examined computer familiarity as a possible confounding factor in studies dealing with computer-based instruction or assessment, hypothesizing that the quantity and quality of students' previous computer experiences can affect their performance. However, the effect of students' familiarity with computers on computer-based assessment is not conclusive. Some studies found no effect (Clariana & Wallace, 2002; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004; Wang et al., 2007), but the NAEP study (Sandene et al., 2005) found familiarity might be a factor. This conclusion was bolstered by the fact that students who took the test on notebook computers supplied by NAEP facilitators scored significantly lower than those who took the tests on their schools' computers. Despite suggestions that CBA produced lower scores (possibly just a matter of calibration) and that a lack of familiarity with computers possibly lowered scores (Sandene et al., 2005), the authors concluded that the use of CBA could shorten the development cycle. As far as Educational Testing Service was concerned, apparently the prospect of greater efficiency outweighed the possible downsides of CBA.

Issues with Technology

The desire to assess students' higher-order thinking in mathematics is not new. Kulm (1990) reports on efforts by the National Assessment of Educational Progress (NAEP) in the 1980s to develop a conceptual framework for tests that would assess higher-order thinking in mathematics and science. They encountered difficulties

separating mathematical thinking from other skills required to complete the questions and identifying age-appropriate benchmarks for mathematical thinking. However, demands for accountability and the ever-present financial constraints on public education have ensured the status quo on most forms of assessment. Instead of increasing their capacity for individualization, these forces have reduced standards to the lowest common denominator. Demands for high validity and reliability, while not inherently negative forces, have magnified the effects of the previously mentioned factors. Tests produced under such conditions cannot be expected to allow for individualized or multiple forms of demonstrating knowledge.

As with any assessment medium, CBA has potential pitfalls. Difficulties with language may inhibit valid assessment of the target skills (Baker & O'Neil Jr., 2002), and the interface may not be as transparent to the test takers as with a paper test. In a study by Chung and Baker (2003), college freshmen using a mouse-based testing interface reported having to navigate by clicking was somewhat bothersome. The researchers interpreted this to mean items in the test worth clicking were things the students felt were truly worth pursuing, but younger students may lack the cognitive development to make that assumption valid for their age.

Summary

Issues Ketterlin-Geller (2005) encountered when designing a CBA in mathematics for third graders who required assistive technologies are instructive when designing similar assessments for a general population. She had to accommodate not only

cognitively variable solution paths, but also variable abilities of perception and physical dexterity. Practice with the computer was also required to ensure adequate familiarity and lessen effects of the medium. All the learner characteristics and technological factors must be planned from the beginning of the design stage of the assessment. This produces far better results than retrofitting the test later.

Computer-based assessment, although no panacea for the challenges of learning or assessment, has already become a fixture in education. Its benefits include efficient gathering of comparable data from large populations, automated or even instant grading, acceptance of varied response formats, and the ability to produce individualized feedback and scaffolding. Computer-based assessments making student thinking explicit can shorten the assessment, feedback, and revision portions of the instructional cycle. Ideally, enough iterations of a given assessment will yield models of the content that can be fed back into the assessment, eventually leading to real-time feedback in an adaptive format assessment. Issues with differential item functioning occur at rates similar to those encountered with paper-and-pencil tests. These findings are general in that they have not focused on any particular content area. The following section examines cognition and assessment as they specifically relate to mathematical understanding, especially in young students.

Mathematics, Cognition, and Assessment

Mathematics has received a great deal of attention with regards to cognition and assessment because of its prominent position in school curricula, its importance in

economically important fields such as science and engineering, and the relatively quantifiable nature of computation. However, to mathematicians, computation represents a very limited portion of their discipline. They are more concerned with solving problems and understanding patterns. If students are to gain knowledge of and appreciation for concepts of mathematics that go beyond computation, they must have a curriculum, instruction, and assessment that reflect those broader goals. They also must sense that their teacher's knowledge of and disposition towards mathematics align with that mindset (Bransford et al., 2000).

Meaningful mathematical knowledge relies on models that learners develop, use, and refine in an iterative process (Lesh, 1990). However, models are individualistic and segmented in young learners. Lesh reminds us of this and also that knowledge is local by stating that "knowledge exists in pieces" (p. 84). It is also situated, and simultaneously coded in multiple forms, including language (written and spoken), mathematical notation, internal models, diagrams, and with manipulatives. Finally, he states that the purpose of assessment is "to probe the nature of the interpreting model to determine its degree of accuracy, complexity, completeness, flexibility, and stability" (p. 86).

The importance of manipulatives has carried over from the classroom into computer-based mathematics instruction and assessments. Manipulatives have been shown to be effective, but they are not all this is needed. Manipulatives do not inherently posses mathematical concepts any more than a digit does. They are helpful to the extent that a child has constructed a mathematical idea and related that idea to the

manipulatives. Once that connection is made, use of manipulatives may allow children to construct further mathematical meaning. Without that meaning, manipulatives may be as mysterious to children as the numbers or concepts they are supposed to represent, and fail to create the desired bridge between the abstract and the real world (Clements & McMillen, 1996; Lesh & Lamon, 1992).

Virtual manipulatives have no more claims to inherent possession of mathematical concepts than do physical ones, but they do have other potential advantages. To students, they might be just as real and, more importantly, as meaningful. They are flexible in a cognitive sense, and can more closely represent mathematical processes than can physical objects. For example, "breaking" a virtual 10-rod into 10 unit pieces is a more accurate depiction of that process than is trading in a physical 10-rod for 10 unit pieces (Clements & McMillen, 1996).

Configurations and processes involving virtual manipulatives can be saved and replayed for either individuals or a class, and can provide feedback in ways blocks cannot. "Certain computer manipulatives help students view a mathematical object not just as one instance but as a representative of an entire class of objects" (Clements & McMillen, 1996, p. 274). Virtual manipulatives are yet another way to represent content, which may reach students that other methods have not. Their use should precede teaching of algorithms and be treated as a means or tool, not as an end in themselves.

Clements and McMillen (1996) recommend giving students adequate time to work with virtual manipulatives and not forcing any particular type, use, or method on

students. They need the freedom to allow their own ways of thinking to come through. Given that freedom, computer-based mathematics problems can become very real to students.

Lipson et al. (1990) describe numerous criteria and characteristics of computer-based tests that allow students to demonstrate the higher order thinking and problem-solving abilities that recent mathematics curricula are demanding. Such tests should elicit numerous facets of student knowledge:

- What prior knowledge does a problem stimulate?
- How does the student represent the problem?
- How does context affect student response?
- What algorithms does the student use?
- How is the student reasoning?
- Does the student use an estimate to check the answer?
- How does the student handle roadblocks?
- What non-school mathematics skills does the student bring to the problem?
- What general knowledge does the student bring to the problem?

Such tests should allow the use of constructed response items, which are easier to score in mathematics than in other areas, rather than multiple choice items whenever possible. Constructed response items force students to think rather than rely on clues from the answers in a multiple choice item. This would encourage all students to behave in a manner similar to high-ability students, who more readily bring existing knowledge

to bear on a problem and save other clues, such as possible answers, for checking their work (Snow, 1987).

In assessments of mathematical achievement, truly authentic problems are less likely to elicit rote, algorithm-based strategies to problem solving. Lipson et al. (1990) demonstrate how standard test questions that attempt to make explicit the underlying concept of a problem are likely to be cut from traditional tests because such items do not effectively discriminate high-achieving students from low when compared to simpler test items. Whatever inferences can be made from test items that elicit more than rote level thinking must be followed up with additional items. This creates a rich description of a student's capabilities, which can be used for instructional purposes. Computers make this level of analysis possible.

Computer-based assessment in mathematics has great potential for evaluating the outcomes of cognitively guided instruction. The next section explores the basic principles and processes of CGI.

Cognitively Guided Instruction

Cognitively guided instruction (CGI) is an approach to teaching mathematics that is transmitted primarily through professional development and predicated on the notion that children enter elementary school with considerable yet informal knowledge of mathematics. Prior to formal instruction in mathematics, children can solve problems involving the basic operations of addition, subtraction, multiplication, and division (Carpenter et al., 1999). This existing knowledge should form the basis of development

of more formal mathematics and not be disregarded as irrelevant. The basic principles of this approach were explored by Carpenter (1986), although others were also exploring the relationship between informal and formal mathematical knowledge in children (Hiebert, 1986). Carpenter did not use the phrase *cognitively guided instruction* in his 1986 chapter, but was using it by publication of the first full-scale study based on those ideas (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989).

CGI falls within a social constructivist perspective of learning in which prior knowledge forms the basis for internal development of new knowledge. Students are encouraged to use and explain their own methods for solving problems, whether informal or formal. Students learn from each other's ideas, and seeing the work of other students can produce cognitive dissonance that helps correct misconceptions. Lesh and Kelly support this aspect of CGI by stating that "ways of thinking tend to be externalized in a group" (2000, p. 214). However, the primary goal of CGI is to increase mathematical understanding in individual students, not of groups.

Through guidance from teachers and observing how other students approach the same problems, students' mathematical skills progress through a number of largely predictable stages of increasing formality and abstraction. Although their original, informal mathematics skills are used as a starting point, more advanced mathematical thinking generally will not develop further without some type of formal instruction (Carpenter et al., 1989).

The most basic strategy of informal mathematics for children in early elementary school is modeling, in which students use their fingers or other manipulatives to physically model the action described in the problem. This type of modeling remains at the core of mathematical understanding for some time but becomes increasingly complex and abstract. From modeling with physical objects, most progress to counting, which is an abstraction of modeling, and begin to develop number sense, which is more efficient than physical representations. This leads to counting strategies, beginning with counting on (beginning with the first quantity stated in the problem) and leading to counting on from the larger number, which is more efficient. Once students can use numbers in an abstract sense, they begin to acquire number facts such as doubles (e.g. 6+6=12) and complementary numbers (pairs of numbers that add up to 10). These number facts can be either memorized through repetition or spontaneously derived. Students generally require formal instruction to advance to the next stage, which involves place value and the meaning of a base-10 system. Finally, working with multi-digit numbers requires place-value based decomposition of numbers (Carpenter, Fennema, & Franke, 1996; Dehaene, 1997; Fuson, 2003).

For teachers to build upon students' existing knowledge, they need a general understanding of these typical stages of early mathematical development and to learn how their particular students are thinking. In a CGI classroom, teachers seek any and all solution paths by asking students to describe and demonstrate their strategies. In this regard, CGI performs the role reversal typical of constructivist environments: instead of

trying to get students to understand teachers' methods and explanations, teachers strive to understand their students' methods. Misconceptions, once uncovered, are not opportunities for corrections but serve only to identify the boundaries of existing knowledge and a starting point for further progress (Carpenter et al., 1996). Misconceptions are explored from a mathematical point of view, not in terms of correctness or in a teacher-centered manner (Hiebert et al., 1997), which would be inconsistent with a socially constructivist perspective.

Implementation of CGI is accomplished through the teachers and depends on extensive and ongoing professional development. The professional development is time intensive and requires multiple sessions with follow ups before teachers are comfortable enough to integrate the new instructional methods in their classrooms. Students do not receive direct instruction in CGI; the emphasis is on the knowledge, skills, and dispositions of the teachers (Carpenter et al., 1999; Carpenter & Franke, 2004; Fennema et al., 1996). The aspects of CGI dealing with professional development are not directly germane to this study of assessing students' mathematical thinking and will therefore not be addressed further.

CGI offers teachers a means to bridge content knowledge and general pedagogical knowledge, creating math-specific pedagogical content knowledge (PCK) (Carpenter et al., 1996; Fennema et al., 1996; Shulman, 1986). It also improves mathematical content knowledge, which is a common weak spot for elementary educators. More importantly, it "provides teachers a coherent basis for identifying what is difficult and what is easy for

students and for dealing with the common errors they make" (Carpenter et al., 1996, p. 14). These skills are critical for effective implementation of CGI in the classroom.

Carpenter et al. (1996) were not the first or only researchers focusing on teachers' concepts of mathematics, mathematics instruction, or student knowledge. However, their work differs in that the core of CGI is the merging of those three bodies of knowledge (Carpenter et al., 1996). Instead of the inevitable struggles resulting from two sides (teachers and students) approaching the problem of instruction from their own perspective, teachers using CGI learn to understand and work from the students' perspectives. This approach connects students' formal study of mathematics with their previous experiences and does not invalidate the informal mathematical skills they bring to the classroom.

Results from the first complete implementation of CGI (Carpenter et al., 1989) showed that CGI teachers taught more problem solving when compared to a control group, spent more time eliciting students' strategies, and expressed positive attitudes about CGI. Students who received CGI performed slightly better on achievement measures and reported higher confidence in their mathematical abilities.

Conclusion

The literature on computer-based assessment provides ample evidence that its validity and reliability are equal or nearly equal to that of paper-and-pencil test. It can also record and interpret the cognitive processes of test takers. Comparisons of scores between CBA and PPT have shown little difference, although the cause for the fact that

some items function differently based on delivery mode has not been determined. The effectiveness, validity, and reliability of CBA have not been shown to be significantly lower than PPT, and the efficiency for capturing cognitive processes may be higher than existing methods such as videotaping and think-alouds.

Despite (or perhaps because of) the current emphasis on standardized testing, mathematics educators are moving away from algorithmic and solution-based assessments towards development and expression of mathematical thinking. However, current methods of this type of assessment are not cost or time efficient.

Eliciting mathematical thinking (as opposed to concrete solutions) is best accomplished with items purposefully created for that goal. Requiring final solutions can narrow students' thinking and hamper chances of eliciting, even provoking, the desired type of responses: those that reveal how students are thinking. One method of drawing out student responses is to put students in situations in which they are not in control of the solution process. CBA that meet these criteria and elicit the desired information about students' mathematical thinking are possible to create, although they may simply shift the inefficiencies of administration (think-alouds, videotaping, clinical interviews) and interpretation to the development phase.

Assessment should produce some instructional benefit, and items that elicit student thinking do so in two ways: they provide teachers with a roadmap of what students do and do not understand, and the students' act of recording their thinking in some fashion encourages refinement of that thinking.

Evaluation of the results of such assessments requires some sort of standard against which responses can be compared. One such standard could be results of performance-based assessments. This might produce some conclusions about transfer or application, but would not serve as a direct evaluation of cognitive processes. Another means of comparison could be with videotapes, think-alouds, and clinical interviews. This comparison might initially be used to validate such assessments, but beyond that would defeat the benefits of efficiency CBA can provide. Ultimately, the point of comparison for anything not directly observable is a construct or model. As previously described, the construct or model must be developed in the initial stages of instructional planning so instruction and assessment can align with the stated goals (Ketterlin-Geller, 2005). This is similar to an objective test, except that instructional objectives are typically described using an observable behavior (Dick, Carey, & Carey, 2005), which is not the case in assessments of cognitive processes.

Developing a construct—even for a well-researched area such as children's mathematics education—is no easy task. No model is likely to satisfy all researchers or schools of thought. However, if assumptions and components of a construct are transparent, assessments relying on them can used defensibly by practitioners and researchers alike.

Directions for Study

Previous studies and existing literature do not fully answer the question of whether computer-based tests could effectively assess the mathematical skills cognitively

guided instruction seeks to develop in elementary school students. Therefore, the next chapter proposes a study whose participants are within that age range and are receiving CGI instruction in mathematics.

Although some concerns over differential item functioning and computer familiarity remain, computer-based assessment has a sufficient record of research and implementation to warrant its use. It is efficient and has been shown to be capable of eliciting cognitive processes and is therefore appropriate for a study using web-based animations of solutions to a well-defined set of problems targeted to a specific age group. These animated items, which modeled multiple examples of both successful and incorrect strategies, attempted to elicit students' mathematical thinking and understanding of predefined mathematical concepts. Lack of direct control over the strategies and process focused the student on expression of agreement or disagreement with the process they were viewing. Although the main purpose of the activities was assessment, it also incorporated elements of instruction due to its recursive nature. Outcomes evaluated included the effectiveness of the tool to elicit students' mathematical thinking and the relationship between those results and objective measures of students' mathematical problem-solving skills.

CHAPTER THREE: METHODOLOGY

The study described in this section was planned to explore a gap in existing literature: to determine if a computer-based assessment can reveal mathematical thinking in primary school students in ways useful for teaching and learning. The assessment focused on a narrow area of mathematical skills (multiplication of integers) and students in a single grade (fourth). Specifically, this study attempted to answer the following questions:

- What is an assessment instrument able to reveal about students' understanding of mathematic concepts related to multiplication of integers?
- What relationships are demonstrated between the results of the assessment for mathematical understanding and the assessment of demonstrated algorithmic proficiency in multiplication?
- What effect does computer familiarity have on the ability of the assessment to reveal mathematical thinking?

Participants

Participants were drawn from fourth grade classrooms in Lincoln and Van Buren Elementary Schools, two of six elementary schools in the Caldwell, Idaho school district. Both schools are Title 1 eligible. Caldwell is located is southwestern Idaho, about 20 miles west of the capital city Boise. Caldwell's population in 2008 was 42,331, which represents a 63% increase over 2000. Rapid growth has produced a relatively young

population: the median age is 4.5 years younger than the median for the state of Idaho. Median and per capita incomes are also substantially lower than the remainder of the state. The area's only substantial minority population is Hispanic ("Caldwell, Idaho," 2009b).

The schools met all No Child Left Behind (NCLB) mathematics goals for the 2008-2009 school year. However, despite geographic proximity and many similarities, the two schools display some differences. Table 1 contains demographic information for the schools (Brendefur, Strother, & Bunning, 2009); Table 2 displays results from the spring 2009 Idaho Standards Achievement Tests in mathematics, whose scores are used to determine adequate yearly progress (APY) for NCLB. Noteworthy differences include the fact that these two schools outperformed the other four elementary schools in the district in mathematics and that the disparity in performance between white and Hispanic students, significant in Lincoln Elementary, the district, and the state; was very small at Van Buren Elementary ("Caldwell, Idaho," 2009a; "Statistics," 2009).

Table 1

Demographics of Treatment Schools

Caldwell School District (2008-2009)		
Characteristic	Lincoln	Van Buren
Enrollment	531	536
Faculty	25	21
Math Endorsement	0	0
Racial/Ethnic	White: 54%	White: 33%
	Latino: 45%	Latino: 65%
E.L.L.	25%	37%
Migrant	3%	3%
Languages	English, Spanish	English, Spanish
Low-income	94%	82%
Free/reduced lunch	76%	88%
Title 1	Yes	Yes

Table 2

NCLB 2008-2009 Report Cards for Treatment Schools in Mathematics

	Lincoln	Van Buren	District	State
School, (%) Proficient/Advanced,	81.91	82.71	69.27	81.57
4th Grade, (%) Proficient/Advanced	76.92	90.66		
White	94.29	90.90	77.95	84.60
Hispanic	51.85	89.79	62.42	66.56

Treatment

Lincoln and Van Buren Elementary Schools have been treatment schools in the Initiative for Developing Mathematical Thinking (IDMT) project for 5 and 2 years, respectively. As treatment schools, all teachers responsible for teaching math have received training in cognitively guided instruction during intensive week-long training sessions each summer. Training sessions are run by personnel from the IDMT, including its director, Dr. Jonathan Brendefur. IDMT personnel also visit project schools to observe mathematics instruction, conduct follow-up workshops, and advise teachers during the school year. Because of turnover in teacher and student populations, there is no way to control or guarantee how many years a given teacher has participated in the program or how long a given student has received cognitively guided instruction in mathematics.

Measures

The assessment instrument was a web-based survey consisting of two sections, which gathered data regarding mathematical understanding and proficiency in multiplication. Data for both sections of the assessment was gathered using Qualtrics survey software, which is commercial survey software available to faculty and students of Boise State University through a university-wide licensing agreement. The assessment was available only to those provided the URL by the researcher. Data was and remains accessible only to the researcher through a secure (HTTPS) login.

The first section consisted of 15 animated solutions to multiplication problems. Animations allowed the students to follow the solution process step by step. The problems contained pairs of one and two-digit numbers presented with and without context. Each problem solution was followed by three or four multiple choice questions and one constructed response item that asked for students' reactions to the strategies and errors (if any) in the solutions. The fourth multiple choice question was displayed only when the students indicated the presence of a mistake in the third question. The animated solutions demonstrated constructs and ideas critical to a mathematical understanding of multiplication. Approximately one-half of the problems (7 of 15) contained errors. See Appendix A for a complete description of the assessment items: a list of the problems and their determining characteristics, a chart detailing the strategies used to solve each problem and the mathematical concepts demonstrated, and a list of the numbers used in the problems.

The need for each concept to be assessed multiple times dictated the number of items contained in the assessment. A single item does not yield reliable data about a student's true ability on the assessed concept. Three to five items per concept or topic are therefore required to produce a reliable measure (Airasian, 2005; Oosterhof, Conrad, & Ely, 2008).

Problems in the first section were divided approximately equally between those presented with and without context: some problems were embedded in word problems while others were already represented in equation or number sentence format. Niemi, Vallone, and Vendlinski (2006) found context was an important factor not just in solving problems but in assessing problem solving in sixth graders. Ginsburg, Klein, and Starkey state "the likelihood that children will solve a word problem is influenced by the degree of interest they find in the content" (1998, p. 422), which dictates that the context of problems be made as relevant and contemporary as possible. Similarly, Siegler and Alibali (2005) found that "unfamiliar contexts often lead children not to apply procedures that they use successfully in other contexts" (p. 393) and that children in the United States use more sophisticated mathematical strategies in a school setting than, for example, when playing a game requiring simple mathematical computations. This speaks to the aforementioned disconnect between "school math" and everyday mathematical situations.

Problems involving higher single digit numbers take longer for both adults and children than problems containing lower single digit numbers. This may be due to

inherently greater complexity or because such problems are drilled less frequently (Dehaene, 1997). Accordingly, numbers used in the problems were selected based on the following criteria:

- Diversity of numbers: numbers repeat as little as possible (within the confines of subsequent criteria) so that previous problems provide as few clues as possible about later problems.
- Maximum three-digit products: Products of two-digit by two-digit numbers are less than 1000. This kept the difficulty appropriate for the age group.
- Reliance on known number facts: problems do not require single-digit number facts in which both digits are above five. This should reduce the time spent performing and analyzing the lower cognitive levels (number facts) of the problems.

The second section assessed proficiency in multiplication. It contained five items requiring students to solve multiplication problems with characteristics similar to those of the problems presented in the first section. Students were not required to demonstrate solution paths or strategies; they were only required to provide an answer.

Mathematical Construct of Multiplication

A construct of understanding, especially one that is to be assessed in a large-scale standardized method, must be predefined (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003; Niemi et al., 2006). This study attempted to elicit mathematical understanding of multiplication as described below and summarized in Table 3.

Concepts

Place value is an important concept in any mathematical procedure involving multi-digit numbers. It means the value of a digit depends on its location relative to the decimal point in a number, even if the decimal point is only implied. For the purposes of the current study, the work of Fosnot and Dolk (2001) and Carpenter et al. (1999) define the role of place value in multiplication for elementary school students.

Many different levels and branches of mathematics use the distributive property, but its primary use by elementary school students is combined with decomposing numbers – even single digit numbers – to break a problem into a group of partial products. When decomposing multi-digit numbers, place value is once again an important concept. Caliandro (2000) and Fuson (2003) defined and described uses of the distributive property, decomposition, and place value by elementary students.

Table 3

Concepts for Multiplication

Concept	Definition	Method of eliciting understanding
Place value	The value represented by a digit depends on its location	Recognition of errors in place value
Distributive property	A x (B + C) = AB + AC	Decomposition using "friendly" numbers

table continues

Table 3 (continued)

Concept	Definition	Method of eliciting understanding
Communicative property	A x B = B x A	Accepting solutions with the order of terms reverse from expected/typical
Flexibility	Knowing there are multiple solution paths that will produce correct results	Testing acceptance of multiple and sometimes unconventional solution paths
What is multiplication/ Use of multiplication	Recognizing situations in which multiplication is appropriate, and that multiplication is a summative process	Identifying problems for which multiplication is an appropriate function

The commutative property, which states that the order of two terms in multiplication does not change the product, cannot be violated by commission. This property simplifies procedures, such as putting the larger number on the top in the standard multiplication algorithm regardless of which number appears first in the problem or what the numbers represent. An example in which the larger number was placed on the bottom in the algorithmic solution could differentiate between those who recognized the solution's inefficiency from those who believed the solution incorrect.

Cognitive flexibility serves as an instructional method and a goal in CGI (Carpenter et al., 1999). As a goal, it is indicative of a high level of conceptual understanding and abstract thought. Such abilities allow students to grasp the mathematical similarities among seemingly different problems and to select solution strategies appropriate for a given problem (Caliandro, 2000; Wilhelmi, Godino, & Lacasta, 2007). Acceptance of a variety of solution strategies would therefore be one indicator of understanding of the mathematical concepts related to multiplication. However, such acceptance could also indicate merely familiarity with the presented strategies.

The concept of multiplication is both so commonplace and abstract that not a single resource consulted for this study actually defined it. While the previously mentioned characteristics define aspects of multiplication, the best way to show an understanding of it as a whole may be to recognize situations for which multiplication is the appropriate process. To be sure problems include the student's decision of whether to us multiplication or some other process, items must be presented in pre-mathematized as word problems, diagrams, or other non-mathematical format.

Strategies

Elementary school students use a variety of strategies to solve problems. These strategies may be acquired through modeling, formal instruction, or from other sources. The ability to use varying strategies appropriate for the problem is one indication of understanding beyond the procedural level. Recognizing alternate strategies either for

their own sake or for checking an answer is another indication that students have a good grasp of the mathematical aspects of a problem. The instrument designed for this study used these strategies, with and without errors, to elicit students' thinking. Compiled from a number of the resources cited in this study, Table 4 contains a list and explanation of strategies commonly used by elementary school students. See Appendix B for a list of the resources consulted to compile this list. No listed strategy is exclusive to any single source.

Counting, which is a common early strategy used by students learning multiplication (Dehaene, 1997; Fosnot & Dolk, 2001; Ginsburg et al., 1998), was not included because fourth graders (the target population of this study) typically have moved on to more abstract concepts and methods, and because most of the problems used in this assessment use numbers too large to make counting a practical strategy. Single-digit multiplication is used almost exclusively in the context of the presented solutions because it represents the simplest level of number facts. The solutions do not contain erroneous number facts with the exception of the problem containing two single digit numbers, so responses to number fact concepts are not solicited.

Table 4

Strategies for Multiplication

Strategy	Description
Repeated addition	Adding the value of multiplicand to itself the number of times represented by the multiplier
FOIL	The process of adding the four partial sums (first, outside, inside, last) generated by placing two 2-part terms next to each other. May be done in an algebraic sense or as a result of decomposing numbers
Multiplying by 10 (Zero trick)	A shortcut for multiplying by 10 by adding a zero to the right end of a whole number or shifting the decimal point one digit to the right.
Friendly numbers	Using known number facts of nearby numbers and then compensating for the difference(s) between the actual and friendly numbers
Halving and Doubling	An extension of the distributive property in which factors of two are moved from one number to another in a multiplication problem
Algorithm	A step-by-step procedure for solving a type of problem
Area/Arrays	Representing the product of two numbers in rows and columns whose length is each one of the numbers
Decomposition	Breaking down a factor in a multiplication problem as the sum of two or more numbers. May be done by place value or by using friendly numbers.

Sample

This study used a convenience sampling: students were selected from two participating treatment schools in the Developing Mathematical Thinking project in Idaho. The goal for sample size was a total of 80-100 students from two classes from each school. The actual sample size was 86 participants. Boise State University IRB approval for the Developing Mathematical Thinking project covers data collection in the schools noted above. To control for possible interaction effects between the conceptual and procedural portions of the assessment, half the students took the conceptual assessment first and the other half began with the procedural assessment. All other aspects of the survey (content, instructions, and interface) were identical for all participants.

Data Collection

The assessment was administered by a research assistant in the IDMT project at Boise State University. Four classes (n=86) took the assessment in their respective schools, one class per day, within the span of one week in December 2009. To make sure the technology worked and to be able to answer questions from students, the research assistant took the assessment prior to administration. His was the first data record, which was deleted. The research assistant reported computer problems for a number of students in one session, but students were able to complete the assessment and save their data in all but two cases. Without personally identifiable information, students who completed the assessment but experienced technical difficulties could not be identified from the

data. Therefore, any possible effect of the technical difficulties on their performance could not be determined. The average duration was 68 minutes per student.

To control for the effects of one section of the assessment on the other section, the order of the two sections (understanding and performance) was reversed after two sessions, resulting in a roughly equal division into two groups by order of the sections. Assignment to the two groups was not truly random, so that process did not fully control for existing differences that may exist between the classes.

CHAPTER FOUR: RESULTS

Data gathered from administration of the constructed assessment instrument required a extensive treatment before they could be used to answer the research questions: (a) What was the assessment instrument able to reveal about students' understanding of mathematic concepts related to multiplication of integers?, (b) What relationships were demonstrated between the results of the assessment for mathematical understanding and the assessment of demonstrated algorithmic proficiency in multiplication?, and (c) What effect did computer familiarity have on the ability of the assessment to reveal mathematical thinking?

This chapter discusses the process of converting qualitative data into usable formats, coding qualitative data, and the analysis of the data as it pertains to the research questions. Quantitative data from the multiple choice questions in the animation section of the assessment were analyzed first, followed by qualitative data from the constructed response questions. Finally, data from those two sections were examined as a whole and then compared with data from the performance and familiarity sections. An outline of the analyses described in this chapter is presented in Table 5.

Table 5

Outline of Procedures and Outcomes in Data Analysis

Procedure	Outcome
Quantitative data preparation	Correct /incorrect coding of all responses to multiple choice questions
Compiled data from familiarity questions	Combined computer familiarity score
Unanswered question frequencies	Participants whose responses were excluded from certain analyses
Coding of qualitative data	Frequencies of nine identified response characteristics
Counts and averages of correct answers	Performance by question and by problem
Reliability quotients calculated	Reliability of data evaluated
Correct answers totaled by strategy and concept calculated	Performance strengths of identified strategies and concepts
Response patterns determined	Progression of understanding
Patterns of error identification determined	Relationship between error identification and other skills revealed
Suggested strategies analyzed	Students' preferred strategies
Compared length of response with understanding	Importance of length of response
Understanding compared with proficiency	Moderate positive correlation established
Familiarity scale compared with other sections	Computer familiarity determined not to be a factor
Influence of understanding and proficiency sections on each other compared	Effect of order of sections inconclusive due to non-random assignment

Data Preparation

The data were downloaded in comma separated values format (.csv) and prepared in Microsoft Excel for analysis. Each participant was coded for section order and answers to the first four questions for each solved problem in the animation section and for the five performance questions were coded dichotomously: correct or incorrect. In the animation (understanding) section, the same given response might be correct for a question pertaining to a problem containing an error in the solution but incorrect for a problem without errors. In other words, a response of "yes" could be correct for a problem containing an error but incorrect for a problem that did not. The coding process took this into account.

The questions pertaining to familiarity with computers and computer testing were coded so that greater familiarity (as defined by greater and more recent use for testing use or having a computer in the home) resulted in a higher familiarity score as detailed in Table 6. Participants received a composite familiarity score ranging from 0 to 7. The mean composite score was 4.6, with a standard deviation of 1.2. In rural areas and in schools with socioeconomic profiles similar to these schools, significant numbers of families do not have computers in the home, and schools do not have sufficient resources to make up this "digital divide" (Thorsen, 2009).

Table 6

Computer Familiarity Questions

Question 1: How many tests have you taken on a computer?

Response	Points	No. of Responses
None	0	0
1 or 2	1	15
3 to 5	2	22
> 5	3	50

Question 2: When was the last test you took on a computer?

Response	Points	No. of Responses
> 2 year ago	0	11
> 1 month ago	1	40
< 1 month ago	2	21
< 1 week ago	3	14

Question 3: Is there a computer you can use at your home?

Response	Points	No. of Responses
No	0	21
Yes	1	65

Unanswered Questions (UAQ)

A number of responses were missing throughout the assessment. Overall, 550 of 4472 (12.3%) of questions in the animation section that should have been answered were not answered. However, this includes 264 questions not answered because students

answered a previous question incorrectly and therefore were not presented with the follow-up question. Subtracting those questions from the unanswered category brings the percentage of unanswered questions down to 6.4%. Of the multiple choice questions related to the animations, none were unanswered more than 7% of the time.

There are several possible reasons for questions not being answered. The assessment did not require students to answer any question before moving on the next question, and students may have been reluctant to answer a question on which they were unsure of the answer. Time was definitely a factor: the response rate for first problem's questions was 97.7%; this decreased over the course of the assessment, dropping to 84.6% by the last problem. Possible causes include time pressures or decreasing motivation. On the constructed response questions, students who agreed with the solution strategy or did not see any errors and students who felt they lacked an appropriate response may not have felt compelled to respond.

To check whether one type of question (of the first three for each problem) went unanswered at a different rate, a Chi-square test was run on the frequency answered. The first three questions for each problem were not answered at statistically significantly different rates, χ^2 (2, N = 86) = .34, $p > .05$.

Response rates for some participants in some sections were low enough to cause concern. When a response rate for a student dropped below the thresholds described below, that participant's data for that section were not used for comparisons across sections and broader statistical analyses. All data were retained, however, and used in

qualitative analyses and in tests where SPSS could maximize the amount of data used by applying pair-wise comparisons rather than list-wise deletions.

The rationales for cut-off points for unanswered questions are described below by section.

1. Familiarity: All but two participants answered all three questions; two left one question unanswered. It was not necessary to omit any participants from analysis based on UAQ in this section.
2. Performance: 80 of 86 participants answered all five performance questions; two left one UAQ; four left four or five UAQ. These last four participants were omitted from analysis as described above.
3. Animation section (multiple choice questions): The distribution of unanswered questions provided a cut off for the maximum number of UAQ allowed in this section. There is a gap between six and nine UAQ. (No participants had seven or eight UAQ.) Because a UAQ occurred on a random basis (other than varying positively with elapsed time), data from participants with up to six UAQ were retained and data from participants with nine or more UAQ in this section were dropped from certain analyses as described above.
4. Constructed Response: The 15 solved problems were divided between seven containing a deliberate mistake in the solution process and eight without such mistakes. These responses were the most time consuming for participants to

complete, and the response rate also dropped as the test progressed. The amount of useful data contained in given responses varied greatly. When data from this section were analyzed in relation to data from other sections of the test, participants must have responded to at least 10 of the 15 problems. While this is only two-thirds, some blank responses could be interpreted not as the question having been unanswered but as the student having nothing to say. Qualitatively, any response that yields information about the understanding of the problem was included.

In problems not containing a mistake, some participants indicated that there was one and were then presented with the fourth question asking them to identify the step containing the mistake. Responses to such instances of the fourth question were not coded for quantitative analysis because it is not possible to quantify the response to a question for which there is no correct answer.

Coding of Constructed Response Questions

All the responses to two constructed response questions were mined for codes (Glesne, 1999). A preliminary set of 11 possible codes was compiled from a review of approximately 120 responses, and were divided into five groups. Responses within a group (each comprising 1 to 3 codes) were generally mutually exclusive, but a code could be used from as many groups as applicable. The same responses and preliminary codes were sent to a second coder (a graduate student), and the results compared to the researcher's. Inter-rater reliability was moderately consistent (65%), but two codes

regarding understanding (codes 1 and 2 in the original set of codes) were found to be difficult to interpret and apply consistently. Consider two responses to the second problem (which did not contain a mistake): (a) “I would fix it by doing another stradagie,” and (b) “You could make it more simple in less steps.” Determining whether the students did or did not understood the method used in the animated solution or if they just would have preferred another method does not seem possible. As a result, the two problematic codes were eliminated and the remaining codes were rearranged to better align with the wording of the question by moving the codes regarding mistakes from between the other two groups. Interpretation guidelines were also clarified. See Appendix D for the full set of codes, their evolution, and instructions for their application used by both coders. The researcher and second coder each coded a second group of responses, some each from questions containing and not containing a mistake. The second round of coding produced agreement on 114 of 128 responses, or 89.1%. The researcher and second coder then each coded half of the remaining responses, using the revised code list and revised set of guidelines. The second coder expressed uncertainty about 19 specific responses in the second round of coding, but only two responses required recoding by the researcher.

Subsequent analysis and discussion refers to and differentiates between the five questions asked after each problem was presented. To simplify, each question in its entirety (15 presentations each) is referred to by uppercase letters A through E as noted in

Table 7. Graphics of the 15 problems with the presented solutions are contained in Appendix D.

Table 7

Question Types

Question type	Question text
A	Did my solution work?
B	Is my answer correct?
C	Did I make any mistakes?
D	Can you tell me which step I made a mistake in?
E	Please tell me how you would have solved this problem or how I could fix any mistakes I made.

Analysis of Data from Multiple Choice Questions

Students received the first three questions for each problem 15 times (once per problem). The mean scores and standard deviations for the first three question types are shown in Table 8. The numbers of correct responses to the first three questions for all animated problem are shown in Table 9. The differences among the totals by question is significant as confirmed by a Chi-square test, χ^2 (28, N = 80) = 554.915, $p < .05$. The third (C) question was answered correctly more than either of the first two, and the second (B) question was answered correctly more often than the first question for problems with mistakes. Because these three questions were always presented in the

same order, it is not possible to determine whether these and progressions noted are due to inherent differences in the questions or to students becoming more familiar with each problem as the over the course of three questions.

Table 8

Mean Scores by Question Type

Question type	Question text	M	SD
A	Did my solution work?	7.77	3.01
B	Is my answer correct?	8.11	3.05
C	Did I make any mistakes?	9.36	2.28

Table 9

Correct Responses Counts by Question Type

Problem	A	B	C	Total
1	30	35	46	111
2	50	53	51	154
3	69	72	72	213
4	49	59	58	166
5	34	44	43	121

table continues

Table 9 (continued)

Problem	A	B	C	Total
6	29	34	40	103
7	54	49	60	163
8	16	21	22	59
9	47	39	53	139
10	42	39	56	137
11	27	32	36	95
12	43	39	50	132
13	54	53	60	167
14	26	32	30	88
15	42	38	58	138
Problems w/o mistake	401	382	460	1243
Problems with mistake	211	257	275	743
All problems	612	639	735	1986

Reliability

Internal reliability, a necessary condition for validity (Airasian, 2005; Gay & Airasian, 2003), was tested at multiple levels using SPSS to calculate Cronbach's Alpha. Coefficients for question types by mistake/no mistake and across all questions by type are presented in Table 10.

Table 10

Reliability Coefficients Across all Problems and by Mistake

	Question type	Alpha
All questions by type	A	0.71
	B	0.71
	C	0.27
Problems with mistakes	A	0.73
	B	0.74
	C	0.72
Problems without mistakes	A	0.69
	B	0.68
	C	0.46

Reliability was similar for question types A and B in all three cases, but differed drastically for C type questions. Reliability for C type questions was very different for problems with and without mistakes and lowest of all when combined. This would imply that knowing a problem does not contain a mistake is different than being able to say a problem does contain a mistake. Students received the C type question (Did I make any mistakes?) once for each problem, for a total of 567 times. They answered it correctly 280 times, or 49.3%. This means they only had 280 possible times to answer the fourth (D) question (Can you tell me which step I made a mistake in?). Of those 280 times, they

correctly identified the step containing the mistake 195 times, or 69.6%. This could support the idea that the question measures two different tasks or abilities. However, students could simply be better at finding a mistake when they know or believe one exists than determining whether there is a mistake.

Performance

Percentages of correct answers for the sample across all multiple choice questions by concept and by strategy are shown in Table 11. Percentages are more helpful than counts in this case because the number of problems a strategy or concept pertained to varied from two to ten.

Table 11

Percentages of Correct Answers by Strategy and Concept

Strategies		Concepts	
Repeated addition	63.4%	Place value	49.5 %
FOIL	33.7	Distributive property	52.8
Zero trick (x10)	44.8	Communicative property	58.6
Friendly numbers	56.7	Flexibility	60.5
Halving & doubling	62.1	Concept of multiplication	59.0

table continues

Table 11 (continued)

Strategies	Concepts
Algorithm	59.9
Area/grid	60.9
Decomposition	49.3

Performance at the group level by concept and strategy reveals strengths and weaknesses. The group was weak on the concept of place value, which is of particular importance when multiplying two digit integers, and fared relatively poorly on problems that used decomposition as a strategy. The two problems that used the FOIL method both contained an error, but one was procedural (not multiplying the inner and outer pairs), while the other error pertained to place value.

Strategies on which students performed well included repeated addition and area or grid. This relates to a strong showing on the concept of multiplication, because those two strategies embody the concept of multiplication. A strong score on flexibility shows that the students were generally not confused by the unusual applications of some strategies, such as making the larger number the multiplier when use the traditional algorithm.

Response Patterns

The pattern of responses for each group of A, B, and C questions by problem is shown in Table 12. The ones and zeros in the 3-digit response pattern represent correct

(1) or incorrect (0) response to the first three questions (A, B, and C, respectively). For example, “111” indicates correct responses on all three questions; “010” indicates incorrect responses on the A and C questions and a correct answer on the B question.

Table 12

Response Pattern by Problem

	Response Pattern							
Problem	000	001	010	100	011	101	110	111
1	26	10	2	1	8	3	0	24
2	15	3	3	4	5	1	2	42
3	1	2	1	0	4	1	3	64
4	13	2	1	0	8	1	1	46
5	23	2	0	2	9	1	1	28
6	29	7	1	2	6	0	1	24
7	10	5	1	1	2	6	1	44
8	47	6	1	1	4	0	2	12
9	17	6	0	2	3	8	1	34
10	14	10	1	1	4	9	1	28
11	28	5	1	0	7	3	1	21
12	16	8	3	2	1	4	1	31
13	6	8	1	2	2	4	2	45
14	33	4	3	0	6	3	5	17
15	9	17	1	2	1	7	3	27

table continues

Table 12 (continued)

Problem	000	001	010	100	011	101	110	111
Mistake	199	36	9	6	48	11	11	172
No mistake	88	59	11	14	22	40	14	315
Total	287	95	20	20	70	51	25	487

Students responded either all correctly or all incorrectly on 73% of the problems. This consistency is further strengthened by correlations among correct responses to the first three questions: r=.79 (A to C), .71 (B to C), and .71 (A to C). However, which end of the spectrum the majority ended up on depended on whether the problem contained a mistake or not. For problems without a mistake, students overwhelmingly answered them all correctly: they thought the solution worked, the answer was correct, and there were no mistakes. For problems with a mistake, the most common response pattern by a smaller margin was to answer all three questions incorrectly, which means they thought the solution worked, the answer was correct, and there were no mistakes. Across all problems, they thought everything was fine 65% of the time. Student response patterns showed increasing understanding over the course of the problem (patterns 001 and 011) 16% of the time, but decreasing (100 or 110) or inconclusive patterns (101 or 010) 4% and 7% of the time, respectively.

Across all students, the average number of correct answers increased from the first to the second question in each group, and from the second to the third. It is difficult

to determine if this was due to the nature of the three questions or to students understanding each problem better as they thought about it more.

Error Identification

Table 13 shows performance patterns on the first, second, and fourth questions for students who correctly answered the third question (Did I make any mistakes?) for problems that contained a mistake. For problems that contained an error, students answered the third question correctly 257 of 584 times (44%). Table 13 breaks down the 257 correct responses on the third question. Among students who knew a mistake had been made, most (66%) had answered the first two questions correctly: they indicated the solution did not work and the answer was not correct. Within that group, a large majority (81%) were able to identify the step in the solution that contained the mistake. However, 13% of students who indicated the solution contained a mistake had previously indicated the solution worked and the answer was correct. Students who answered one of the first two questions correctly were able to select the step with the mistake correctly less often than not only the students who answered both of the first two questions correctly but also less often than students who answered both of the first two questions incorrectly.

Overall, students who said there was a mistake were able to identify the step containing the error on 73.5% of the time. Although students who answered the two previous questions correctly performed better than others, all groups were able to identify the step with the mistake at least 50% of the time. Of note is the fact that students who

missed both the first two questions were more likely to identify the step containing the mistake than students who missed one of the first two questions.

Table 13

Performance Patterns on the First, Second, and Fourth Question for Students who Correctly Answered the Third Question

Results on questions A and B	No. of students (%)	Picked step containing mistake?	No. of students (percent of pattern)	Percent of total
Both correct	169 (66%)	Picked	137 (81%)	53
		Missed	32 (19%)	12
Only A correct	10 (4%)	Picked	5 (50%)	2
		Missed	5 (50%)	2
Only B correct	45 (18%)	Picked	26 (58%)	10
		Missed	19 (42%)	7
Neither correct	33 (13%)	Picked	21 (64%)	8
		Missed	12 (36%)	5

Note. Percentages do not add up to 100 because of rounding.

Students' Constructed Responses

For questions that did not contain mistakes, the relevant part of the fifth question for each problem was "Please tell me how you would have solved this problem." Some students reported and even showed how to correct mistakes that did not exist, mostly relating to some unconventional ways in which some solutions represented place values.

For example, problem 12 (see Figure 2) used a mostly conventional algorithm, with the ones column in the fourth row left blank instead of containing the implied zero. Several students suggested moving the 16 to the right, despite having seen a problem three problems previously that showed the implied zero in a similar situation. No students suggested writing in the implied zero.

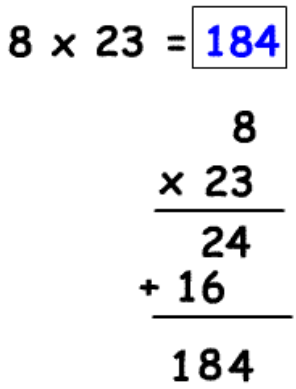


Figure 2. Problem 12 featuring missing implied zero after "16"

The most common response to this question was some form of agreement with the solution shown, despite the unconventional strategies used. After agreement, the most common pattern of responses was repeatedly suggested the same one or two strategies, with arrays, repeated addition, and decomposition being the most common. There were several suggestions for using the "regular," "traditional," or "original" method (interpreted as meaning the standard algorithm) for solving problems that were not solved

with some form of the standard algorithm. Only two students suggested more than two different methods or strategies.

Note that when responding to that same question for problems that did contain mistakes, students had already noted (or missed) the mistake. The most common response to a mistake was to show a different method to solve the problem or to re-do the problem using the given method but correcting the mistake. Students gave direct explanations of what they corrected much less frequently. Whether using corrected solutions or suggesting a different strategy, most students tended to rely on one or two strategies, mostly decomposition and number facts, often in combination.

Strategies

Students' reactions to and selection of strategies are important components of understanding. Table 14 shows the total number of times each code was applied to problems with and without mistakes. The number of codes adds up to greater than the total number of questions because multiple codes applied to some responses.

Table 14

Code Counts by Problem Type

Code	Without mistakes	%	With mistakes	%
Agreed with strategy used	140	28%	89	19%
Disagreed with strategy used	73	14%	79	17%
Suggested different general strategy	103	20%	72	16%
Suggested different, correct specific strategy	59	12%	59	13%
Suggested different, incorrect specific strategy	18	4%	27	6%
Noted error correctly, but no explanation or correction	0	0%	11	2%
Error noted and correctly explained	0	0%	45	10%
Error noted by incorrectly explained	0	0%	7	2%
Incomplete/undecipherable/meaningless response	112	22%	72	16%
Totals	505		461	

In problems without mistakes, students noted agreement with the strategy twice as often as they expressed disagreement, while in problems with mistakes, expressions of agreement and disagreement were almost even. That the students agreed more with successful strategies than unsuccessful ones indicates some level of understanding. The number of times students suggested different strategies, whether correct or incorrect, whether general or specific, were roughly even between problems with and without mistakes. This must be viewed in two ways. First, they suggested different strategies just

as often when nothing needed fixing. Second, they did not suggest different strategies any more often when something did need fixing. Students' desire to solve the problems using their choice of strategy was not affected by the presence or absence of a mistake. Many, but not all (as we will see later), had clear ideas about how to solve the problems.

In problems with mistakes, students suggested new strategies more than twice as often as they explained what was wrong with the given strategy. The question asked for either and few did both. The disparity could be because the part of the question that asked them how they would have solved the problem was first, so more students answered that part. Another explanation could be that even though they knew something was wrong, they could not explain why. To discount the first explanation would require either splitting the question into two questions or random assignments of the original question with the order of the two parts reversed.

Patterns of responses across the four multiple-choice questions in the animation section were able to reveal subtle (but not significant) differences in the abilities those questions measured, and varying levels of those abilities in students. Those abilities are useful to students when they perform multiplication or other mathematical operations, but of course are also portions of the construct of understanding multiplication. To make this information usable for instruction and feedback, reliability of the items will have to be improved and the construct of multiplication used will have to be operationalized in more definitively assessable ways.

Students were familiar with the problems by the time they saw each final question. They had watched an animation of a solution (perhaps multiple times) and had answered three or four questions that asked them to think about the problem in different ways, and each question was accompanied by a graphic showing all steps of the solution. Familiarity should have given them the freedom to say what they wanted. As noted earlier, some did not respond to the final question of a problem, and this increased as the assessment progressed. However, many responded without having anything to add (at least mathematically). For example, two responses were “I don’t know” and “I have nothing to say.” Since not responding was an option, it is difficult to blame meaningless responses on laziness or feeling rushed. The conclusion that they did not know how to fix the problem or how they would have solved it themselves is difficult to escape.

Participants showed definite preferences for certain strategies. Table 15 shows the counts of suggested strategies by question and divided by whether questions contained a mistake.

Table 15

Suggested Strategies by Question

	Problems Without Mistakes							
Strategy Suggested	5 x 28	3 x 4	12 x 14	18 x 25	16 x 12	8 x 23	4 x 17	15 x 29
Repeat add	5	12	1		1	2	4	1
Decomposition	20	7	10	6	6	6	2	5
Add zeros	1							
Array/area/grid	3	4	8	5	8	6	7	8
Arrow	1							
Traditional algorithm			1		1			
Counting		9						
Tree		1	1	1	1	2		1
Ratio table			1		1			
FOIL					1			

	Problems With Mistakes						
	23 x 7	6 x 7	12 x 27	25 x 23	13 x 16	20 x 30	19 x 4
Repeat add	2	4	1	1	1	1	5
Decomposition	15	13	8	3	5	5	5
Add zeros						1	
Array/area/grid	5	6	3	6	8	5	7
Arrow	1						
Traditional algorithm	2			1	1		

table continues

Table 15 (continued)

	Problems With Mistakes						
	23 x 7	6 x 7	12 x 27	25 x 23	13 x 16	20 x 30	19 x 4
Counting		5					1
Tree			1				1
Ratio table	1	1					1
FOIL					3		
Number facts		3					

At individual and group levels, students relied heavily on two strategies: decomposition and array/area/grid. Note that the numbers in Table 15 combine the response of students that described the strategy with those who simply stated the strategy they would use. Students often referred to decomposition, for example, by stating "I would break the numbers up," but some would detail the process: "10x18=180+10x18=180+5x18=90" (actual response for the problem 18 x 25). The graphic nature of the second most commonly suggested strategy (array/area/grid) made it impossible to suggest by anything more than name because the computer would only accept text. Despite this limitation, the frequency with which some students suggested using arrays called into question their true understanding of them or their ability to use them.

Looking at which strategies students applied to which problems is informative. Repeated addition was suggested most often where one or both of the factors were a single digit. This indicates an understanding of the essence of multiplication and is a

more appropriate strategy for these problems than for when both numbers are larger or double digits. Students suggested counting as a strategy only on the two problems that contained both single digit factors. These are the best instances to use that particular strategy, although it is earlier on the developmental timeline of mathematical understanding because of its low level of abstraction.

Only three students reported using number facts, and all on the same problem (6 x 7). That problem was directly preceded by the "3 x 4" problem, but no student stated "I just know that 3 x 4 = 12," even though it is probably fair to say that more 4th grade students know "3 x 4 = 12" as a number fact than know "6 x 7 = 42." This discrepancy may be because the "3 x 4" problem was solved correctly, but 6 x 7 was not. With the correct answer to "3 x 4" displayed, students may not have realized they would have known it anyway without using the surprising number of different strategies (five) they suggested.

Understanding as a Function of Length of Constructed Response

For a problem containing a mistake, did giving only the correct answer mean the student corrected the error in the given strategy or did it mean she used a different strategy? Combining the constructed responses with students' answers to the multiple-choice questions provided suggestive but not conclusive answers. For example, Table 16 shows all the responses by students who explicitly stated that 600 was the correct answer for problem 11. Others answered the multiple-choice questions correctly or explained the mistake, but did not state the correct answer.

Table 16

Differences in Understanding Reflected in Longer and Shorter Answers

	Did my solution work?	Is my answer correct?	Did I make any mistakes?	Can you tell me which step I made a mistake in? (step 3)	Please tell me how you would have solved this problem or how I could fix any mistakes I made.
Student 1	No	No	Yes	Step 3	i whould brak it up 20 10 10 10 20 x 10 200 20 x 10 200 10 x 10 100 10 x 10 100 = 600
Student 2	Yes	Yes	No	-- [a]	20x30=600
Student 3	Yes	Yes	No	-- [a]	20x30=600 +20x30=600
Student 4	Yes	Yes	No	-- [a]	20 times 30 eaquals 600
Student 5	No	No	Yes	Step 3	i put 20x30 it was 600 and then i did oxo it was o and my answer was 600.
Student 6	No	No	Yes	Step 2	it =600
Student 7	No	No	Yes	Step 3	2x3=6 add two zeros anser 600

[a] Student did not see this question because of response of "no" to previous question.

The question for each problem that asked whether the computer made a mistake is subject to Type I and Type II error on the student's part. These possibilities are important not from a statistical point of view but because they reveal different characteristics about the responder. The former may indicate a lack of flexibility with varied solution methods or a reliance on known procedures. The latter, which in this case would be not identifying an existing error, could mean a lack of understanding of the underlying relevant concept.

Students who stated the correct answer showed one of two distinct patterns on the previous three or four questions for this problem. They were, with one exception, consistent: they answered all the multiple-choice questions correctly or missed them all. This particular subset of responses shows none of the inconsistencies that a quarter of overall responses. Four of the students apparently understood what was wrong with the problem the entire time; the other three did not see any problem until they had to offer an explanation, when they solved it themselves. It cannot be determined whether the latter three would have revised their earlier answers after calculating the correct answer.

The determination to be made here is whether these two groups of students (four and three) have different levels of understanding of the concepts and strategies involved in this particular problem. The constructed responses of three of the four students who answered the MCQs correctly are more detailed. The fourth student, whose constructed response was not detailed, also incorrectly identified the step containing the error (question four). Looking at the constructed responses alone, it would be overreaching to say that students who did not provide an explanation of their correct answer to the

problem lacked understanding. However, when combined with the multiple-choice questions, it appears that providing only the correct answer to the problem demonstrates a lack of understanding, not just an omission. This inconsistency between the responses on the MCQs and the construction responses could indicate an inability to bridge the gap between what they know is correct (from solving the problems correctly themselves) and what looks acceptable but cannot be correct. Taking this one step further, if the multiple-choice questions are the determining factor in deciding whether a student understands what is going on in a problem, those questions could be said to demonstrate understanding on their own.

Relationship Between Understanding and Proficiency

The correlation between scores on the two sections (understanding and performance) was $r(75) = .46$, demonstrating that there is considerable but far from total overlap between these abilities. This is also evident anecdotally by looking at the performance scores of students who earned a high total score in the understanding section (see Figure 3). While some performed well in the performance section, several scored 2 or 3 (of 5) on the performance section, and some demonstrated limited familiarity with a variety of strategies. In general, however, students with high performance scores were slightly more likely to score well on the understanding section and students who answered one or none of the performance questions correctly were more likely to have lower scores in the understanding section. The assessment instrument was able to

differentiate between conceptual understanding and the ability to multiply two integers accurately.

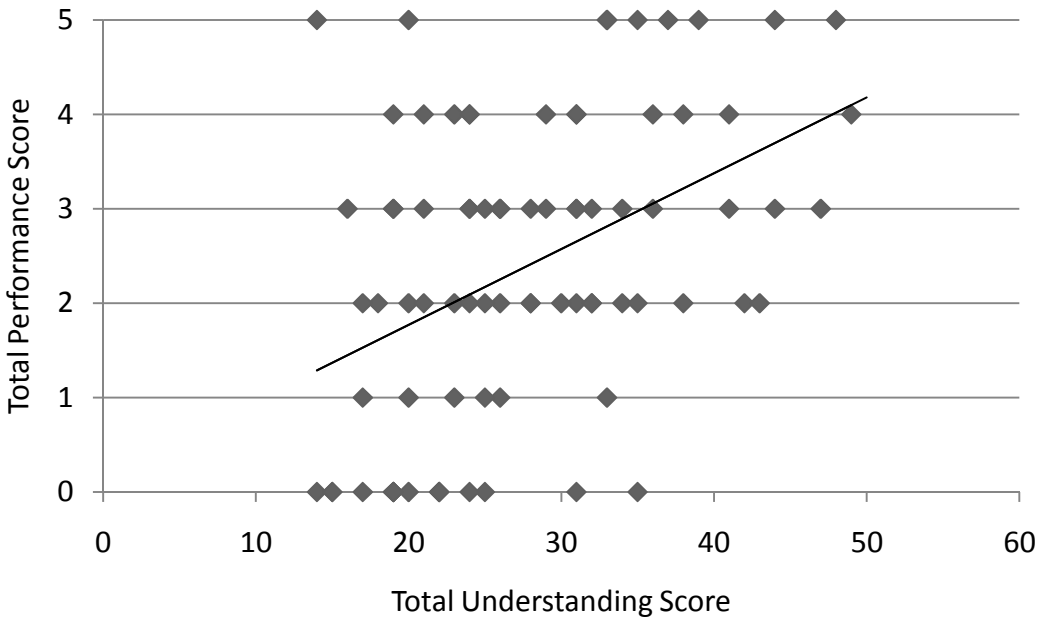


Figure 3. Scatter plot comparing students' total scores on understanding and performance sections

Relationship Between Computer Familiarity, Understanding, and Proficiency

Scores on the familiarity scale did not correlate strongly to scores on either performance or understanding, $r(75) = .03$ and $r(75) = .11$, respectively. Although a few students did indicate having little experience taking tests on computers or not having a computer at home, their computer skills are demonstrably adequate to complete this

assessment without familiarity becoming a factor. Some of the spelling and grammar issues previously noted may be the result of low keyboarding skills, which become as much an impediment to interpretation as it is to students.

Effect of Order

As previously mentioned, two of the four groups received the animation section questions before the performance question, while the section order was reversed for the other two groups. Since this was administered at the class level, assignment to those two groups (by order) is not truly random. As can be seen in Table 17, the groups performed differently on the two sections.

Table 17

Mean Scores by Section and Section Order

	Mean animation score (*SD*)	Mean performance score (*SD*)
Animation first	27.38 (9.12)	2.00 (1.81)
Performance first	28.31 (8.32)	2.83 (1.24)

The group that received the performance questions first performed better on both sections, but a one-way ANOVA indicated the difference on the animation section was not significant, $F(1, 77) = 0.30$, $p > .05$. The difference in the performance scores was

significant, $F(1, 79) = 5.04$, $p < .05$. Without truly random assignment, it is not possible to say the difference was a result of reversing the order of the two sections.

Comparison across Test Sections

For the purpose of comparisons across test sections, a total score was calculated for each student for each section. For the performance section, the score was a count of the correct answers on the five multiplication questions, yielding a score from 0 to 5. The score for the understanding section was a total of the correct responses to the first three questions for each problem, which produced scores up to 45 (15 problems x 3 questions per problem). The fourth problem for each question was not included because it was not presented if the student did not report a mistake in question three. The score for the section on computer familiarity has already been described. Correlations were calculated among all three sections. The understanding and performance sections correlated at $r = .46$. The computer familiarity section correlated with the animation section at $r = .11$, and with the performance section at $r = .03$.

Summary

Analysis of the data showed that the assessment was able to reveal student thinking in a variety of ways. Students' responses were noticeably different on questions with and without mistakes on both multiple-choice and constructed-response questions. Although they generally relied on a small number of strategies, they did suggest strategies appropriate to specific problems in some cases. The amount of detail or

information they gave in their constructed response was indicative of their level of understanding. They showed individual preferences for strategies, but generally accepted the presented strategy.

The relationship between scores on the understanding and performance sections indicated that these two characteristics correlate moderately but far from totally. The relationship between computer familiarity and both other sections were weak enough to discount familiarity as a significant factor in students' performance on the test.

Having shown what the assessment revealed and some of its limitations, the next chapter concludes the dissertation by discussing implications for teaching mathematics and how the assessment could be improved and applied in a true educational setting.

CHAPTER FIVE: DISCUSSION

This study has been a narrowly focused attempt to assess understanding of mathematics (specifically multiplication) rather than performance. This final chapter restates the need for the study and key points of relevant literature, summarizes its methodology and results, and discusses the implications of its results.

Summary of Purpose and Literature

Assessment of understanding is important for two reasons: first, true understanding is what will help students master successive levels of mathematics they will soon encounter; and second, this type of knowledge is one of the goals of cognitively guided instruction with which assessment should align. The research questions were (a) what was the assessment instrument able to reveal about students' understanding of mathematic concepts related to multiplication of integers? (b) what relationships were demonstrated between the results of the assessment for mathematical understanding and the assessment of demonstrated algorithmic proficiency in multiplication? and (c) What effect did computer familiarity have on the ability of the assessment to reveal mathematical thinking?

The literature of cognition and assessment is rich with descriptions of knowledge structures and modeling, authenticity, and uses of formative assessment. However, existing methods of revealing how students think are time consuming to administer and

evaluate. Computer-based assessment (CBA) can be used to overcome some of those difficulties. CBA has its own set of issues when compared with other forms of testing, but also has a long and established history and practices to mitigate its limitations. Many standardized tests are computer-based, so students are used to taking tests on computers. Ideally, computers could provide real-time feedback useful for learning and instruction.

Cognitively guided instruction is not a specific method for instruction but a general approach based on increased mathematical knowledge in teachers that recognizes and uses students' informal mathematical knowledge in a socially constructivist setting. Although CGI is transmitted to teachers by professional development, its ultimate goal is to improve mathematical understanding in students, thus forming a conceptual and physical setting for this study.

The general concept of the assessment (non-performance based) created for this study came from the idea that the demands of performance inhibit students' opportunity and ability to express their thoughts on the topic (Giordani & Soller, 2004; Lesh & Lehrer, 2000; McClain et al., 1999; Yeh, 2001). The assessment therefore solved problems for the students and captured their thoughts and reactions.

Summary of Methodology

Participants in the study were fourth grade students in four classes from two elementary schools participating in a larger CGI program run by the Initiative for Developing Mathematical Thinking (IDMT) at Boise State University. Although the transmission of CGI is primarily through teachers, the ultimate beneficiaries are of course

students. However, the purpose this study was to evaluate the assessment instrument, not the efficacy of CGI or the professional development.

The teachers at the elementary schools have participated in the IDMT professional development program between two and five years. Both schools have significant Latino populations, Title 1 status, and have experienced rapid growth in recent years. Despite those challenges, both schools met all No Child Left Behind (NCLB) mathematics goals for the 2008-2009 school year.

The assessment consisted of three sections. The largest section focused on understanding, and presented 15 multiplication problems solved by the computer. Each solution was contained in an animation lasting between 11 and 28 seconds, which played on each student's computer. Students could control the playback, which stopped after completion of each major step in the solution, and could watch the entire animation as many times as they wanted. Then they answered three or four multiple choice and one constructed response questions for each problem. The questions in the understanding section focused a construct multiplication consisting of eight strategies and five concepts (see Appendix A). Several strategies and concepts applied to each problem in order to assess them multiple times (see Figure A.1).

The other two sections of the test consisted of five performance questions in which the students had to supply only the answer and three questions about their experiences taking tests on computers. The understanding section preceded the performance section for half the students; the order was reversed for the other half. All

participants took the assessment within a one-week span, shortly after they had completed a unit on multiplication.

Data were gathered using Qualtrics survey software, a commercial product with secure data storage and access methods. Responses were anonymous: the researcher was not present during data collection and no identifiable information was collected. Data were prepared for analysis in Microsoft Excel by coding responses dichotomously (correct/incorrect) for the multiple-choice and performance questions. Qualitative data codes were developed and applied to the constructed responses.

Summary of Results

Analysis of the responses gathered during the four days of administration of the assessment instrument revealed a considerable amount about how students think about and understand the mathematical concepts related to the multiplication of one and two digit positive integers. Patterns become discernable, while inconsistencies revealed other dimensions of students' thinking. The data were viewed from the perspectives of consistency, group-wise patterns, reactions to mistakes and varied strategies, and their strategy suggestions.

In broad terms, it showed different levels among students of recognition or acceptance of various strategies and concepts. The students as a whole demonstrated they were better at determining whether an answer was correct than at determining the appropriateness of a particular strategy. The instrument also showed that of the five concepts rated, place value, arguably the most important, was the weakest. The group

scored highest on the concept of flexibility, operationalized in this study by the use of many different methods and unusual variations in standard methods.

Analysis began with extensive conversion and coding of data. The quantitative data showed the assessment instrument had poor reliability at the levels of individual students' understanding of specific concepts and strategies. When viewed as a group, the data demonstrated better reliability. While still below the levels expected of a standardized achievement test (Gay & Airasian, 2003), reliability coefficients for the group are at levels that make plausible the examination of the validity and results of this new assessment.

At the group level, the assessment was able to reveal strengths and weaknesses in various strategies and concepts. Students performed best on problems that used repeated addition and halving and doubling to solve the problems, while their lowest performance was on problems that used the FOIL method and the shortcut for multiplying by 10 by adding a zero to the right-hand side of a number. The most notable finding regarding concepts was weakness in problems for which place value was an issue.

Students tended to respond consistently across the questions for a given problem, but when they did not, they tended to get more questions right as they went through the five questions for the problem. Again, it is difficult to know whether this is a result of students' increasing understanding of the problem as they progress through the five questions for each problem or differences in the questions and the skills and knowledge upon which they draw.

Students' reactions to the problems in their constructed responses showed a great deal about how they think and how they understand multiplication. Students generally accepted the demonstrated solution, although less frequently for problems that contained a mistake. However, whether the solution contained a mistake or not, they often suggested another way to solve the problem. While only three strategies dominated suggestions at the group level and most individual students relied on only one or two strategies, a number of suggestions demonstrated understanding by being appropriate for the problem.

The constructed responses also demonstrated limits to students' understanding or possibly limits of the assessment to reveal their thinking. In problems that contained a mistake, few students directly fixed or explained the error. Most commonly, they would simply offer another solution or strategy. Students suggested general strategies more often that specific ones.

Overall, constructed responses could be judged to a degree not just by what they said but by what they did not say. Students who suggested and worked out a specific strategy demonstrated more understanding than those who give a general strategy or simply gave the answer, a result corroborated by responses on the first four questions for that problem. Students who took the time to write a response that contained less or no mathematical content probably had lower levels of understanding. While more information in a response obviously demonstrated greater understanding, analysis also

suggested less information was not just an omission but demonstrative of lower levels of understanding.

Comparing the scores between the animation/understanding and the performance sections revealed a moderate positive correlation. Understanding multiplication and performing multiplication are related but distinct capabilities. Comparisons of the scores from the computer familiarity scale with the other two sections revealed very weak positive correlations, demonstrating that students' experience or, in a few cases, lack of experience with computers did not noticeably affect their performance on the assessment.

Discussion of the Results

Understanding mathematical concepts and why strategies work or do not work are some of the goals of cognitively guided instruction (CGI). The purpose of the instrument designed for this study was to provide teachers and students with a tool to reveal and assess that understanding and use it for formative assessment. Before answering the research questions directly, the results are viewed through the two lenses of understanding and use of strategies.

Interpretation of Responses

Determining the meaning and intent of responses to the final question in each group was challenging. Merely figuring out whether they agreed or disagreed with the given strategy involved shades of meaning that were difficult to categorize. "I would use a different strategy" could represent disagreement, whereas "I would use different

strategy to check my answer" could mean the student agreed with the strategy used but wanted to double check her answer, which is something students are taught to do (Carpenter et al., 1999).

In some cases, constructed responses had to be interpreted by reading previous responses by the same student, even though earlier responses would have been related to different problems and possibly different mistakes. For example, "couleter" is not understandable without comparison to "couckulater," the same student's response to a previous question. This type of interpretation would be difficult if not impossible when attempting to use a computer to categorize responses, and it tests even the limits of human memory and associative skills.

Operationalized Construct of Mathematical Understanding

In the context of the construct of understanding multiplication presented in this study, the responses that indicate understanding must be described. Students with a high level of understanding would have correctly identified the presence and location of mistakes in problems, explained the mistakes, suggested a variety of strategies appropriate for the nature of individual problems, and seen that unusual strategies obeyed basic concepts and were therefore acceptable. They would have had an idea of whether an answer was correct (before they solve the problem themselves) by following the strategies and steps in the presented solution. Students would understand why a strategy did or did not work. At the other end of the scale, students with a poor understanding would not have spotted mistakes at all, would suggest the same strategies regardless of

the nature of the problem, and be uncomfortable with strategies that were unusual despite their adherence to mathematical principles. They would accept flawed strategies that contain familiar elements but did not respect mathematical principles. They would rely heavily on processes, such as the standard algorithm.

Interpretation of Results

Students demonstrated all the above characteristics of understanding and lack of understanding at various points and at varying levels in the assessment. This is not surprising and is indicative of variation in mathematical abilities and of concepts and procedures not yet fully internalized (Anderson, 1983). At the group level, the results might suggest some broad instructional strategies. Students need work in recognizing mistakes, and might be encouraged to follow through on suggested strategies by solving the problem, as described in Carpenter et al. (1999). Teachers should not be content with explanations of just a few words. The response patterns indicate that students should be given ample time to think about problems requiring knowledge that has not been fully mastered.

Students' poor understanding of place value is particularly worrisome because it is an important concept in many areas of mathematics they will be learning for years to come (Carpenter, Franke, & Levi, 2003). It is possible that its low ranking among concepts used in this assessment is indicative of the relatively large number of solutions and errors in which it was a factor, but poor understanding of place value is still a concern. The traditional algorithm simultaneously disregards and compensates for place

value, but other methods and strategies require more attention be paid to the true value a digit represents. Teachers might be well advised to pay close attention to their students' understanding of this concept.

Decomposition fared relatively poorly as a strategy, contrasting with the high frequency with which students suggested it in their constructed responses. The difference may be due to the fact that the provided animated solutions primarily broke numbers down by place value, whereas students' responses usually broke numbers into several small, friendly numbers. For example, several students broke the number 25 into two 10s and a 5, and broke 7 into three 2s and a 1. This difference highlights the difficulty in defining the construct of multiplication.

As a group, it seems obvious that these students are being taught multiplication in ways that downplay the standard algorithmic process; ways that attempt to involve more mathematical thinking. Out of 292 total suggestions for how to solve the 15 problems, students suggested using the traditional algorithm only six times. If the same students were given the same problems to solve without further instructions, it is easy to imagine that more might use the traditional algorithm. However, that imagined discrepancy should not be viewed as contradictory or a failure of teaching for understanding. When simply given a problem to solve, it is reasonable to expect students to solve it in the most expedient way, which for many might be the mathematical shorthand of the traditional algorithm.

At the individual level, such an assessment would ideally pinpoint specific areas of high and low understanding. The reliability of items at this level is not sufficient to make such judgments. Even if such judgments were possible at the individual level, they would only be of use if teachers could individualize instruction to meet the revealed needs. This also could be completed with small groups for instruction, and computers could, once again, individualized practice.

Hints of understanding surface in unusual statements, such as "if it is 12 x 27 it cant [*sic*] be 27!" How much more this student understood is not known, but she at least understood that multiplying 27 by a number other than one could not result in 27. Another student wrote "stop making a 10 a 1!" after seeing several deliberate place value errors. Finally, one student could not understand how the solution could have gone wrong after starting correctly: "youn [*sic*] can fix it by you need to add 4 more lines of seven and when you put the numbres [*sic*] down on your strategi [*sic*] you earased [*sic*] it why." Although these three responses were strong indicators of understanding, they, too, would be difficult for a computer to interpret.

The finding that shorter constructed responses were indicative of lower levels of understanding could have implications in other assessments, whether formal or informal; formative or summative. Students encounter open-ended questions in many classroom situations: being called upon in class, descriptive writing, and even in purposeful drawing. The amount of information in any of these responses could be a significant indicator of knowledge and understanding. Students may be encouraged to include any

information in open-ended questions that might be relevant in any way. Although this may seem to give students the best opportunity to demonstrate what they know, this finding shows there is a potential downside to the "everything but the kitchen sink" answer strategy.

Of course, the test can only reveal understanding to the extent that students possess it. They all completed a unit on multiplication about a month previous to administration of the test, but their understanding of multiplication is neither fully mature nor complete. Multiplication might be viewed as a unitary concept for many purposes, but of all the possible ways to solve multiplication problems, only the traditional algorithm treats it that way.

Although a majority of students answered all questions for a problem either all correctly or all incorrectly, some seemed to experience an "ah ha!" or "Eureka!" moment when they went from not understanding the solution or the problem to understanding. This was evidenced by response patterns that changed from zeros (incorrect answers) to ones (correct answers) and by the fact that students who missed both of the first two questions were more likely to identify the mistake than students who answered one of the first questions correctly. The former group might be those who experienced that moment of realization; the latter was unsure and remained so. This pattern could also be because assessing the appropriateness of strategies, knowing whether an answer was correct, and identifying procedural errors are related but not fully overlapping abilities. However, the

numbers are not large and the percentages not that different, so this anomaly could be due to chance.

The fact students got to spend time on and answer multiple questions about a single problem seems to have given some of them a chance to increase their understanding over the course of the problem. If true, the assessment would help define students' zone of proximal development (Vygotsky, 1978). Young learners whose mastery of the material, multiplication in this case, is still forming benefit from having time to spend on a problem, even on a test. *Benefit* in this case is defined as students having the fullest opportunity to demonstrate their knowledge. This is surely desirable in formative assessment, but formative might describe any assessment of learners who cannot be expected to have fully mastered the material. This could extend the description of formative to a great many tests of students this age.

The moderate positive correlation between performance and understanding may be viewed as surprising low or surprising in its strength. It would be easy to imagine a student who had been taught nothing of multiplication but algorithms and procedures being unaware of its underlying concepts. Conversely, a student who only answered multiple-choice, non-performance based questions about multiplication might have a difficult time performing multiplication. Ideally, as students' mastery of multiplication improves, the strength of the correlation between performance and understanding would increase. However, if understanding is neglected or unused, students may be left with

only procedures. It is easy to imagine adults who can perform multiplication but have forgotten why their procedures work.

Recommendations for Educators

CGI recognizes and builds upon informal mathematics knowledge students have before their formal training in mathematics begins, but most students in this study preferred a limited number of strategies. Young students, like most everyone else, have a comfort zone. Expanding that zone will require ongoing attention of teachers.

This assessment, like any other, would suffer greatly from "teaching to the test." While mathematics teachers may hope their students are familiar with and use a variety of appropriate problem-solving strategies, teaching them in ways that cause them to repeat names of various strategies because they think that is what the teacher or, in this case, the researcher wants to hear would skew the results.

As noted above, students at this exact age but also throughout elementary school are constantly increasing their knowledge in all the basic skill areas. Even if a test is the last time a teacher will assess a given skill set, it would be helpful for students to treat the assessment as formative. This would mean allowing students ample time to work on the test and giving feedback beyond a simple grade.

Suggestions for Additional Research

Motivation may be been a factor in performance because students received no grade or credit of any sort for taking this test. The test generally took longer than

expected, causing increasing numbers of unanswered questions as the test progressed. In future iterations, incorporating even experimental administrations into an assignment or assigning a grade even for completion might motivate students to do their best.

When asking students to render their thoughts on a computer, language skills can also become a factor. If students are to give some responses in written prose, their ability in that medium must be controlled or known to separate it from their mathematical ability. The level of grammar and spelling skills in the participants made interpretation of the constructed responses challenging. The ability for some students to express concepts clearly may not be sufficiently developed to make constructed responses reliable and valid. Poor language skills may reduce the validity of the assessment if they prevent students from expressing themselves clearly. English may not be the primary or home language of some of the Hispanic students, which could hamper their ability to fully express their knowledge.

Some of the process descriptions were difficult to categorize by strategy – the constructed response questions gave students the opportunity to display their individual natures and constructions of knowledge. While educators recognize and often celebrate individual characteristics of students, those same qualities make interpretation of responses such as those gathered in this study difficult.

Strategies chosen for the animated problems may have influenced students. Besides the influence noted above of what was presented with number facts, what was seen may have also played a role in other responses. For example, the given solutions for

two problems used halving and doubling. Neither of these contained a mistake. Although no student suggested using halving and doubling to solve any of the other problems, the number who agreed and disagreed with the strategy was similar and in a similar ratio compared with problems using the strategies most often suggested by students (decomposition and arrays). With a correct answer displayed and no mistakes on which to focus, separating students who truly understand the strategy and might ever use it from those who, lacking an obvious reason to object, simply went along with it. The use of multiple strategies, a component of CGI, is supposed to develop understanding of mathematical thinking by allowing students to see how mathematical principles apply across various strategies. However, if not taught well, multiple strategies could lose their effect as a conceptual scaffold and become the new procedural algorithm (Hannafin, Hannafin, & Gabbitas, 2009; Oliver & Hannafin, 2001), ultimately failing to impart improved understanding.

To increase the validity of future versions of this assessment, individual items may need to have fewer factors loading on them. A typical problem in the current version related to several strategies and concepts, strategy selection, and possibly error recognition. The number of items was not sufficient to determine the effects of so many factors, especially concepts and strategies at the individual level. Tests might need to focus on a smaller number of factors to have acceptable reliability and be of a reasonable length. This might prevent a single such test from being comprehensive.

Constructed responses are always subject to interpretation. If this type of assessment is to produce consistently useful data with efficiency, responses might need to be scripted into a selected response format. With fewer factors to measure, such questions could still capture more information than standard multiple choice questions. Given the high Hispanic populations of the schools from which the sample was drawn, offering students additional response methods for the open-ended questions might increase the validity of the test by removing the language factor.

Capturing and assessing students' mathematical thinking remains a necessary goal if the stated purpose of instruction is to improve their mathematical thinking and understanding. In higher levels of mathematics, understanding is more important than proficiency in arithmetic. Narrowing the focus of such assessments may provide the reliability and specificity teachers need to effect change in their instruction.

REFERENCES

Airasian, P. W. (2005). *Classroom assessment: Concepts and applications*. Boston: McGraw Hill.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228*(4698), 456-462.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199-209.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*, 269-282.

Baker, E. L., & O'Neil Jr., H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18*, 609-622.

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the American Educational Research Association, San Francisco.

Boud, D., & Feletti, G. (1997). *The challenge of problem-based learning*. New York: Routledge.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington, D.C.: National Academy Press.

Brendefur, J. L., Strother, S., & Bunning, K. (2009). *Developing Mathematical Thinking: Year 2 Technical Report*. Boise, ID: Boise State University, Center for School Improvement and Policy Studies.

Caldwell, Idaho. (2009a). Retrieved September 29, 2009, from http://www.city-data.com/city/Caldwell-Idaho.html

Caldwell, Idaho. (2009b). Retrieved September 19, 2009, from http://www.city-data.com/city/Caldwell-Idaho.html

Caliandro, C. K. (2000). Children's inventions for multidigit multiplication and division. *Teaching Children Mathematics, 6*(6), 420-426.

Carpenter, T. P. (1986). Conceptual knowledge as a foundation for procedural knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 113-132). Hillsdale, NJ: Lawrence Earlbaum Associates.

Carpenter, T. P., Blanton, M. L., Cobb, P., Franke, M. L., Kaput, J., & McClain, K. (2004). *Scaling up innovative practices in mathematics and science*. Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal, 97*(1), 3-20.

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., 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. (2004). Cognitively guided instruction: Challenging the core of educational practice. In T. K. Glennan, S. J. Bodilly, J. R. Galegher & K. A. Kerr (Eds.), *Expanding the reach of educational reforms: Perspectives from leaders in the scale-up of educational interventions* (pp. 41-80). Santa Monica, CA: RAND.

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

Chappuis, J., & Chappuis, S. (2002). *Understanding school assessment: A parent and community guide to helping students learn*. Portland, OR: Assessment Training Institute.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2).

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior, 18*, 669-684.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Technology, 33*(5), 593-602.

Clements, D. H., & McMillen, S. (1996). Rethinking "concrete" manipulatives. *Teaching Children Mathematics, 2*(5), 270-279.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Dewey, J. (1960). *The child and the curriculum*. Chicago: University of Chicago Press.

Dick, W., Carey, L., & Carey, J. O. (2005). *The systematic design of instruction* (6th ed.). Boston: Allyn & Bacon.

Driscoll, M. P. (2005). *Psychology of learning for instruction* (3rd Ed.). Boston: Pearson Education.

English, F. W. (2000). *Deciding what to teach and test: Developing, aligning, and auditing the curriculum - Millennium Ed.* . Thousand Oaks, CA: Corwin Press.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 403-434.

Fosnot, C. T., & Dolk, M. (2001). *Young mathematicians at work: Constructing multiplication and division*. Portsmouth, NH: Heinemann.

Fuson, K. C. (2003). Toward computational fluency in multidigit multiplication and division. *Teaching Children Mathematics, 9*(6), 300-305.

Gagne, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning* (2nd Ed.). New York: Longman.

Gay, L. R., & Airasian, P. (2003). *Educational research: Competencies for analysis and applications*. Upper Saddle River, NJ: Merrill Prentice Hall.

Gierl, M. J. (2004, February 11). *Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences.* Paper presented at the America Educational Research Association, San Diego, CA.

Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In I. E. Sigel & K. A. Renninger (Eds.), *Handbook of child psychology* (5th ed., Vol. 4, pp. 401-477). New York: Wiley & Sons.

Giordani, A., & Soller, A. (2004, August 22-27). *Strategic collaboration support in a web-based scientific inquiry environment.* Paper presented at the 16th European Conference on Artificial Intelligence, Valencia, Spain.

Glennan, T. K., Bodilly, S. J., Galegher, J. R., & Kerr, K. A. (Eds.). (2004). *Expanding the reach or educational reforms: Perspectives from leaders in the scale-up of educational inverventions*. Santa Monica, CA: Rand Corporation.

Glesne, C. (1999). *Becoming qualitative researchers: An introduction* (2nd ed.). New York: Addison Wesley Longman.

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal, 5*(4). Retrieved from http://escholarship.bc.edu/jtla/vol5/4/

Hannafin, M., Hannafin, K., & Gabbitas, B. (2009). Re-examining cognition during student-centered, Web-based learning. *Educational Technology Research & Development, 57*(6), 767-785.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research, 46*(1), 29-42.

Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., & Murray, H., et al. (1997). *Making sense: Teaching and learning mathematics with understanding*. Portsmouth, NH: Heinemann.

Hoeft, R. M., Jentsch, F. G., Harper, M. E., Evans III, A. W., Bowers, C. A., & Salas, E. (2003). TPL-KATS concept map: A computerized knowledge assessment tool. *Computers in Human Behavior, 19*, 653-657.

IMMEX. (2007). Retrieved October 7, 2007, from http://www.immex.ucla.edu/iWeb/Agencies/142115/default.aspx

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal, 4*(5). Retrieved from http://escholarship.bc.edu/jtla/vol4/5/

Joyce, B., Weil, M., & Calhoun, E. (2000). *Models of teaching* (6th Ed.). Boston: Allyn and Bacon.

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1993). *Extending the rule space model to a semantically-rich domain: Diagnostic assessment in architecture*. Princeton, NJ: Educational Testing Service.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal, 4*(2). Retrieved from http://escholarship.bc.edu/jtla/vol4/2/

Kim, J. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the Mid-Western American Educational Research Association, Chicago.

Kulm, G. (1990). Assessing higher order mathematical thinking: What we need to know and be able to do. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 1-6). Washington, D.C.: American Association for the Advancement of Science.

Lesh, R. (1990). Computer-based assessment of higher order understandings and processes in elementary mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 81-110). Washington, D.C.: American Association for the Advancement of Science.

Lesh, R., Hoover, M., Hole, B., Kelly, A. E., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. In A. E. Kelly & R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 591-646). Mahway, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Kelly, A. E. (2000). Multitiered teaching experiments. In A. E. Kelly & R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 197-230). Mahway, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Lamon, S. J. (1992). Assessing authentic mathematical performance. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 17-62). Washington, D. C.: American Association for the Advancement of Science.

Lesh, R., & Lehrer, R. (2000). Iterative refinements cycles for videotape analyses of conceptual change. In A. E. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 665-708). Mahwah, NJ: Lawrence Erlbaum Associates.

Lipson, J. I., Faletti, J., & Martinez, M. E. (1990). Advances in computer-based mathematics assessment. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics*. Washington, D.C.: American Association for the Advancement of Science.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied Measurement in Education, 5*(2), 151-169.

Maslow, A. H. (1966). *The psychology of science: A reconnaissance*. New York: Harper & Row.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology*. New York: Simon & Schuster Macmillan.

McClain, K., Cobb, P., Gravemeijer, K., & Estes, B. (1999). Developing mathematical reasoning with the context of measurement. In V. L. Stiff & R. F. Curcio (Eds.), *Developing mathematical reasoning in grades K-12, 1999 yearbook* (pp. 93-106). Reston, VA: National Council of Teachers of Mathematics.

McKnight, C. C. (1990). Critical evaluation of quantitative arguments. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 169-186). Washington, D.C.: American Association for the Advancement of Science.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 9*, 287-304.

Mislevy, R. J. (2004). *The case for an integrated design framework for assessing science inquiry* (No. 638). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2003). Improving educational assessment. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: effective research designs for improving learning* (pp. 149-180). New York: Teachers College Press.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). *A cognitive task analysis, with implications for designing a simulation-based performance assessment*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April 1999.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363-389.

Niemi, D. (1996). *Instructional influences on content area explanations and representational knowledge: Evidence for the construct validity of measures of principled understanding (CSE Technical Report 403)*: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Niemi, D., Vallone, J., & Vendlinski, T. (2006). *The power of big ideas in mathematics education: Development and pilot testing of POWERSOURCE assessments (CSE Report 697)*: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil Jr., H. F., & Klein, D. C. D. (1997). *Feasibility of machine scoring of concept maps*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California.

Oliver, K., & Hannafin, M. (2001). Developing and refining mental models in open-ended learning environments: a case study. *Educational Technology Research & Development, 49*(4), 5-32.

Oosterhof, A., Conrad, R. M., & Ely, D. P. (2008). *Assessing learners online*. Upper Saddle River, NJ: Pearson.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

Piaget, J. (1969). *Science of education and the psychology of the child*. New York: Viking.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6).

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal, 2*(6). Retrieved from http://escholarship.bc.edu/jtla/vol2/6/

Richardson, V., & Placier, P. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 905-950). Washington DC: American Educational Research Association.

Roschelle, J., & Jackiw, N. (2000). Technology design as educational research: Interweaving imaginatin, inquiry, and impact. In A. E. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 777-798). Mahway, NJ: Lawrence Erlbaum Associates.

Saettler, P. (1990). *The evolution of American educational technology*. Englewood, CO: Libraries Unlimited, Inc.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., & Kaplan, B., et al. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (No. NCES 2005-457): U.S. Department of Education, Institute of Education Sciences.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal, 4*(6). Retrieved from http://escholarship.bc.edu/jtla/vol4/6/

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil Jr., H. F. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 15*, 403-418.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-15.

Siegler, R. S., & Alibali, M. W. (2005). *Children's thinking*. Upper Saddle River, NJ: Pearson.

Snow, R. E. (1987). Aptitude complexes. In R. E. Snow & M. J. Farr (Eds.), *Aptitude learning and instruction* (Vol. 3). Hillside, NJ: Lawrence Erlbaum Associates.

Statistics. (2009). Retrieved September 19, 2009, from http://www.sde.idaho.gov/site/statistics/

Stroup, W. M., & Wilensky, U. (2000). Assessing learning as emergent phenomena: Moving constructivist statistics beyond the bell curve. In A. E. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 877-912). Mahwah, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, D. C.: American Council on Education.

Thorsen, C. (2009). *TechTactics: Technology for teachers* (3rd ed.). New York: Pearson Education.

Vygotsky, L. S. (1978). *Mind and Society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Walvoord, B. E. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education*. San Francisco: Jossey-Bass.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219-238.

Wilhelmi, M. R., Godino, J. D., & Lacasta, E. (2007). Didactic effectiveness of mathematical definitions: The case of absolute value. *International Electronic Journal of Mathematics Education, 2*(2), 72-90.

Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. *Educational Researcher, 30*(9), 12-17.

APPENDIX A

Assessment for Understanding: Problem Matrix

Table A.1

Description of Problems

Item	Problem	Context	Error	Reason / Justification
1	23 x 7	N	Place value	Simple decomposition, only 1 two digit number to keep simple
2	5 x 28	N	-	Numbers that could be halved and doubled easily, with one of them producing friendly number
3	3 x 4	N	-	Small numbers that would produce easy numbers for repeated multiplication
4	5 x 6	N	Concept of area model	Numbers large enough that students may not know as number fact, small enough to easily count, possible confusion w/ 6x2, 6+6, or 6+7
5	12 x 27	Y	Place value	Area model not as digit dependent, but wanted some partial sums to be 2 digit, and easy partial products. Also, these numbers unique in problem set
6	25 x 23	Y	Place value	Easy partial products, with possibility of confusion about how zeros to add for zero trick
7	12 x 14	N	-	Numbers than would produce an unusual look when partial products were reversed.

table continues

Table A1 (continued)

Item	Problem	Context	Error	Reason / Justification
8	13 x 16	N	Place value	Easy computation of partial products, and easy addition of same
9	18 x 25	Y	-	One small enough that most would put on bottom in algorithm, both partial products end in zero to create confusion.
10	16 x 12	N	-	Numbers that would be easy to halve and double, even numbers easier
11	20 x 30	N	Place value	Multiplication of significant digit would be easily known number fact, both multiples of 10 to create confusion
12	8 x 23	N	-	One single-digit number to make algorithm look unusual, avoid other obvious methods with odd & prime 23
13	35 x 9	Y	-	Access to friendly number, correctional also contains friendly numbers.
14	19 x 4	N	Concept of multiplication	Access to friendly number, decomposition gives easy, known number facts
15	15 x 29	Y	-	Access to friendly number, decomposition easy, but requires 3rd technique

		Strategies								Concepts				
No.	Problem	Repeated addition	Foil	Zero trick	Friendly numbers	Halving & Doubling	Algorithm	Area/Grid	Decomposition	Place value	Dist. Property	Assoc. Prop.	Flexibility	What is Mult?
1	23 x 7								X	X*	X			
2	5 x 28			X	X	X				X	X		X	
3	3 x 4	X											X	X
4	5 x 6							X						X*
5	24 x 27						X	X	X	X*	X		X	
6	25 x 23		X							X*			X	
7	12 x 14				X		X			X			X	
8	13 x 16		X	X					X	X	X*			
9	18 x 25						X			X		X		
10	16 x 12					X			X				X	
11	20 x 30			X	X					X*			X	X
12	8 x 23						X			X		X	X	
13	35 x 9			X	X							X	X	
14	19 x 4	X		X	X				X	X				X*
15	15 x 29			X	X				X		X		X	
	totals	2	2	6	6	2	4	2	6	10	5	3	10	3

* indicates source of the error in problems with an error

Figure A.1 Matrix of strategies and concepts

Table A.2

Numbers Used in Problems

Group	Used as digits	Used as numbers
1-9	All	3, 4, 5, 6, 7, 8, 9
10-19		12, 13, 14, 15, 16, 18
20-29		20, 23, 24, 25, 27, 29
30-39		30, 35

APPENDIX B

Reference List for Multiplication Strategies

Caliandro, C. K. (2000). Children's inventions for multi-digit multiplication and division. *Teaching Children Mathematics, 6*(6), 420-426.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal, 97*(1), 3-20.

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., 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.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Fosnot, C. T., & Dolk, M. (2001). *Young mathematicians at work: Constructing multiplication and division*. Portsmouth, NH: Heinemann.

Fuson, K. C. (2003). Toward computational fluency in multi-digit multiplication and division. *Teaching Children Mathematics, 9*(6), 300-305.

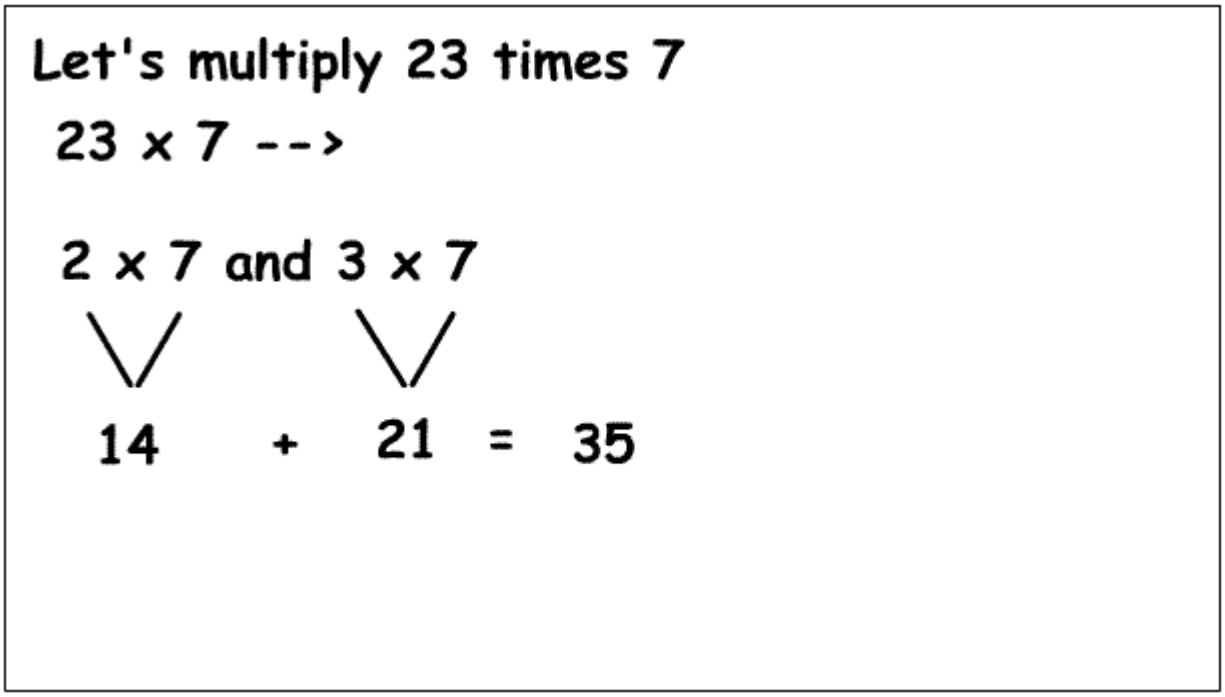
Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In I. E. Sigel & K. A. Renninger (Eds.), *Handbook of child psychology* (5th ed., Vol. 4, pp. 401-477). New York: Wiley & Sons.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., & Murray, H., et al. (1997). *Making sense: Teaching and learning mathematics with understanding*. Portsmouth, NH: Heinemann.

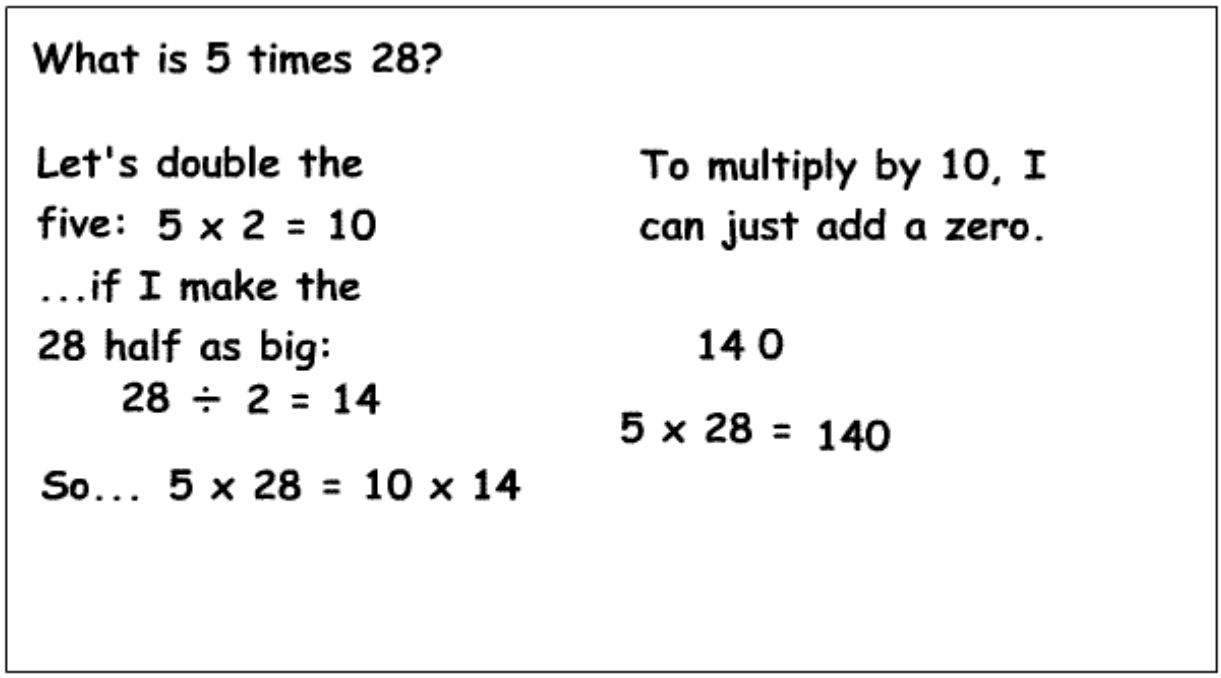
APPENDIX C

Graphics of Animated Problem Solutions

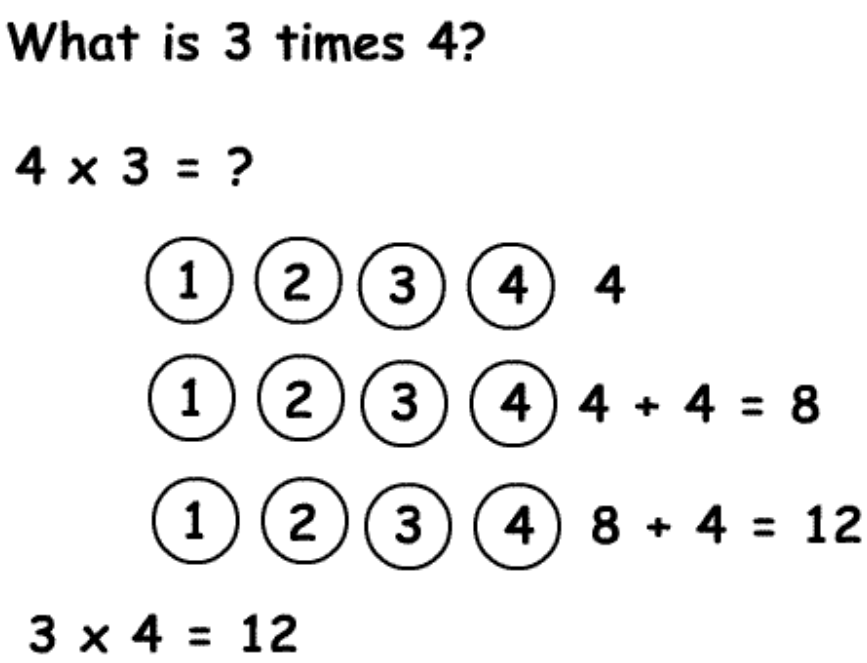
Problem 1



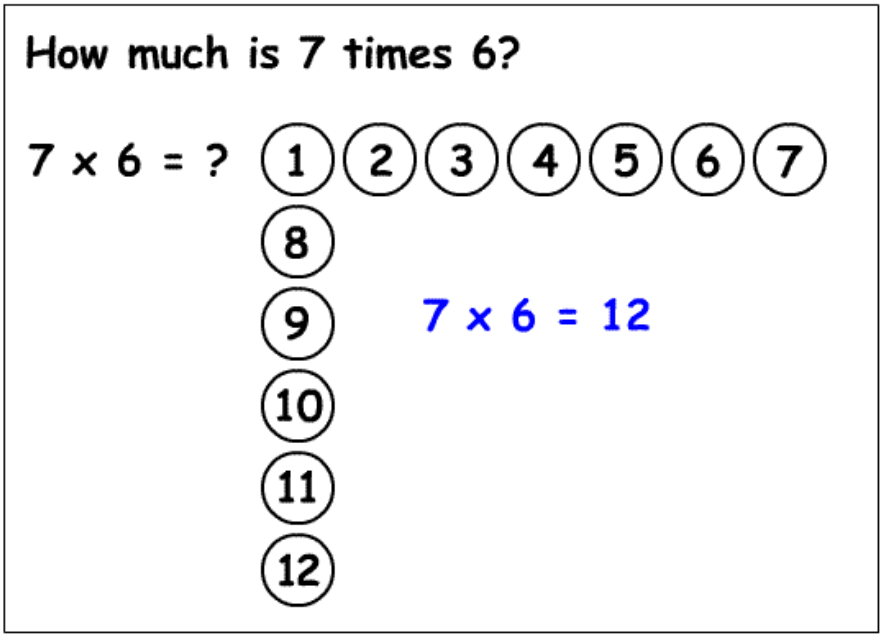
Problem 2



Problem 3



Problem 4



Problem 5

How much is 12 times 27?

12 x 27 = 27

	2	7
1	2	7
2	4	14

2 + 7 + 4 + 14 = 27

Problem 6

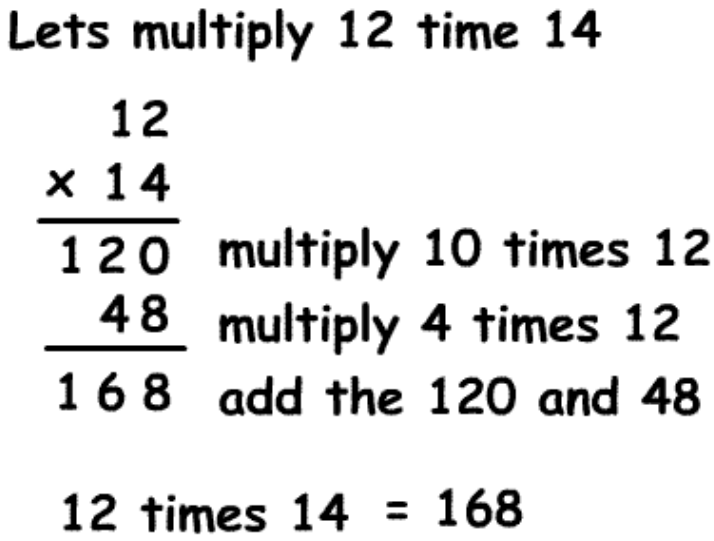
If each student in a class of 25 brought in 23 soda cans for recycling, how many cans would the class have collected?

total cans = 25 x 23

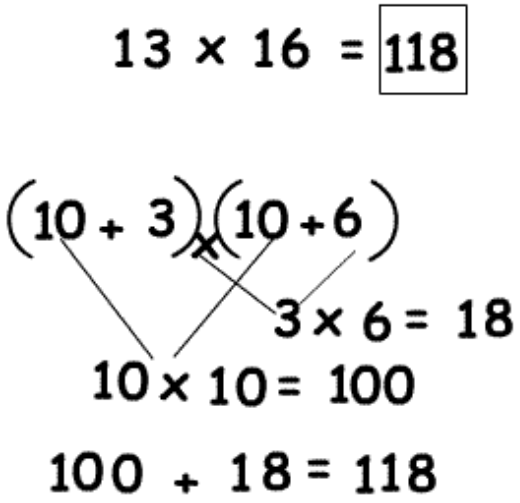
25 x 23 = 125

```
   25
 x 23
-----
2x20  40
2x5   10
3x20  60
3x5   15
-----
  125
```

Problem 7



Problem 8



Problem 9

In the library, there are 18 shelves that hold books. Each shelf has 25 books on it. How many books are there in the library?

$$\begin{array}{r} {}^{4} \\ 25 \\ \times\ 18 \\ \hline 200 \\ +250 \\ \hline 450 \end{array}$$

Problem 10

$16 \times 12 = \boxed{192}$

I know I can double the 12 if I divide the 16 in half:

$16 \times 12 = 8 \times 24$, then I keep halving and doubling:

$8 \times 24 = 4 \times 48$

$4 \times 48 = 2 \times 96$

$2 \times 96 = 1 \times 192$

$1 \times 192 = 192$

Problem 11

$20 \times 30 = \boxed{60}$

Well, $2 \times 3 = 6$

And: $20 = 2 \times 10$, and $30 = 3 \times 10$,

So I need to add a zero to the 6:

$20 \times 30 = 60$

Problem 12

$8 \times 23 = \boxed{184}$

$$\begin{array}{r} 8 \\ \times\ 23 \\ \hline 24 \\ +\ 16 \\ \hline 184 \end{array}$$

Problem 13

Ana, Manny, Eva, and Pete each order a slice of pizza. Each slice has 17 olives on it. How many olives do they have altogether?

$4 \times 17 =$ 68

17 = 10 + 5 + 2, so I can multiply each by 4

$4 \times 10 = 40$
$4 \times 5 = 20$
$4 \times 2 = 8$
68

And the add these up

Problem 14

$19 \times 4 =$ 61

It would be easier to multiply by 20 instead of 19, and then fix it at the end.

And 20 = 10 + 10,

So $20 \times 4 = (10 \times 4) + (10 \times 4)$

$40 + 40 = 80$

Now I need to take away the extra 19,

$80 - 19 = 61$

Problem 15

There are fifteen classrooms in a school. If each classroom has 29 desks in it, how may desks are in the school?

$15 \times 29 = \boxed{435}$

I know that $15 = 10 + 5$

And it would be easier to multiply by 30 instead of 29

So: $15 \times 30 = (10 \times 30) + (5 \times 30)$

$300 + 150 = 450$

Take away the extra 15: $450 - 15 = 435$

APPENDIX D

Qualitative Data Codes

Codes for qualitative data from constructed response questions

Original version

1. Demonstrated understanding
2. Did not demonstrate understanding

3. Agreed with strategy used
4. Disagreed with strategy used

5. Noted error correctly, but no explanation/correction
6. Error noted and correctly explained
7. Error noted by incorrectly explained

8. Suggested different general strategy (e.g. use 2 strategies to check answer, etc.)
9. Suggested different, correct specific strategy
10. Suggested different, incorrect specific strategy

11. Expressed undefined uncertainty

Final version

1. Agreed with strategy used
2. Disagreed with strategy used

3. Suggested different general strategy (e.g. use 2 strategies to check answer, etc.)
4. Suggested different, correct specific strategy
5. Suggested different, incorrect specific strategy

6. Noted error correctly, but no explanation/correction
7. Error noted and correctly explained
8. Error noted by incorrectly explained

9. Incomplete/undecipherable/meaningless response

Instructions for coding:

For each constructed response, write the number of each statement that applies. Some responses may have multiple statements apply, but others may only merit one. I anticipate that each response would use one statement from each group (3 or 4 or 5), but if you feel more than one statement from a group applies, then put both down.

Rules for coding specific types of responses:

1. The difference between a specific vs. general strategy is that a general strategy does not mention specific numbers, while a specific does.
2. If the response says there was no mistake, take that as an agreed with strategy (1)
3. If a response states there was a mistake, that's a disagreed with strategy(2)
4. Give responses consisting entirely of "yes" or "no" a code of 9.
5. Give responses consisting entirely or mostly of "different" a 2.
6. Give responses that merely restate the equation from the problem a 9.