# MEASURING STUDENT LEARNING GAINS IN CONCEPTUAL MATHEMATICS WHEN SCALING A TECHNOLOGICAL INTERVENTION FOR MIDDLE SCHOOL MATHEMATICS

Nicole Shechtman, Jeremy Roschelle, Geneva Haertel, Jennifer Knudsen

SRI International

Deborah Tatar

Virginia Tech

Presented at the American Educational Research Association (AERA)

Friday April 15, 2005

Montréal, Canada



Address all correspondences to
Nicole Shechtman
nicole.shechtman@sri.com

Abstract

U.S. students are less likely than their international peers to have opportunities to learn complex and conceptually difficult mathematics (CCDM), and they lag in achievement. The Scaling Up SimCalc project is a multiyear longitudinal randomized experiment to test the efficacy of an intervention that integrates technology (our software, SimCalc MathWorlds) and curriculum to create opportunities for students to learn CCDM. One of the major challenges of this research has been to create a measurement instrument that would meet rigorous evidentiary standards for validity and have the capacity to demonstrate growth in students' knowledge of CCDM. In this paper, we present our assessment development process, its rationale, and the data we have collected (and continue to collect) to support our argument for the instrument's validity.

# MEASURING STUDENT LEARNING GAINS IN CONCEPTUAL MATHEMATICS WHEN SCALING A TECHNOLOGICAL INTERVENTION FOR MIDDLE SCHOOL MATHEMATICS

Analyses from data from the Trends in International Mathematics and Science Study (TIMSS) reveal that American students lag behind their international peers in mathematics achievement beginning in middle school because they are less likely to be taught or to master complex and conceptually difficult mathematics[1] (CCDM; Schmidt et al., 2001; Suter, 2002). National Assessment of Educational Progress (NAEP) scores likewise show that we have made steady progress in improving the number of students mastering mathematics at a basic level but considerably less progress reaching the more conceptual and complex "proficient" level (Sowder, Wearne, Martin, & Strutchens, in press).

Technology is often cited for its potential to create new opportunities for more students to develop deeper knowledge (National Research Council, 1999) and for "democratizing access to advanced mathematics" (Kaput, 1994). Although there are many compelling case studies in the literature that show students can learn deeper mathematics with representationally rich technology, the field has just begun to measure systematically the effectiveness of various interventions in creating opportunities to learn CCDM. More work in this area is urgently needed, as in today's climate of accountability, an innovative approach deserves attention only if its effect can be systematically documented.

The Scaling Up SimCalc project is a multiyear longitudinal randomized experiment to test the efficacy of an intervention that integrates technology (our software, SimCalc MathWorlds), curriculum, and professional development to create opportunities for students to learn CCDM.

[1] Note that since the text of this paper was written, we have decided that the term "complex and conceptually difficult mathematics" is controversial and ambiguous. We are considering other terms such as "foundational," "generative," and "complex building block concept." This is still open to debate.

Our aim in this research is to use randomized experimentation to gather systematic data to understand what works and what does not work with a wide variety of teachers.

One of the major challenges of this research has been to create a measurement instrument that would meet rigorous evidentiary standards for validity, have the capacity to demonstrate growth in students' knowledge of CCDM, and be developed with only a fraction of the budget for our experimental research. One possible approach to instrumentation could have been to rely on previously validated standardized tests. Because our study takes place in the state of Texas, we examined the Texas Assessment of Academic Skills (TAAS). However, we found that the items on this exam did not cover the depth of CCDM that our intervention addresses. Using such a measure would not capture the conceptual depth students could reach using our technology and curriculum. Therefore, we designed our own student test, extending beyond the content covered in TAAS to assess more CCDM. We also discovered that although most people will agree that it is important for student to learn complex and conceptually difficult mathematics, the mathematics education community does not have a single vision of what constitutes CCDM. Thus, our task is not only to measure the construct of CCDM, but also to begin to define it.

In this paper, we present our assessment development process, its rationales, and the data we have collected (and continue to collect) to support our argument for the instrument's validity.

## Our Research: Scaling Up SimCalc

### *Purpose*

The overarching purpose of the Scaling Up SimCalc research program is to test at scale the following hypothesis:

> *A wide variety of middle school teachers can use an innovative integration of technology and curriculum to create opportunities for their students to learn complex and conceptually difficult mathematics.*

We discuss briefly below each of the key components of this hypothesis: complex and conceptually difficult mathematics, the technology, and the innovative integration of technology and curriculum.

*Complex and Conceptually Difficult Mathematics*

In this project, we focus on rate and proportionality at the middle school level. Proportionality is a central, crosscutting theme that involves number and operations, geometry, and algebra, and connections among them (Hiebert & Behr, 1988) and is a large focus area in both the National Council of Teachers of Mathematics (NCTM) and state standards (NCTM, 2000).

We differentiate the conceptually simple aspects of proportionality (hereafter, "$M_1$") from the conceptually complex aspects of proportionality (hereafter, "$M_2$"). Generally, we consider $M_1$ to be the application of routine procedures. This would include, for example, solving simple problems for one value using the formulas a/b = c/d, y = kx, or d = rt. $M_2$, in contrast, includes making mathematical connections among mathematical representations, mathematical ideas, and realistic situations. This would include, for example, reasoning about a representation (e.g., graph, table, or y = kx formula) in which a multiplicative constant "k" represents a constant rate, slope, speed, or scaling factor across many pairs of values.

Operationalizing the precise constructs of $M_1$ and $M_2$ was critical to our assessment development process. We have iterated through two operational definitions.

*The Technology: SimCalc MathWorlds*

SimCalc MathWorlds was designed to help middle and high school students make the connections required to understand $M_2$ concepts. The software makes dynamic links between simulations of moving objects (linear motion), the associated position and velocity graphs, and

tables and formulas. Through exploring challenging problems, students make rich connections among these representations, resulting in their understanding of rate of change.

Figure 1 shows two windows in MathWorlds representing the same phenomenon. The simulation window shows two runners moving across a field. The position graph window shows a plot of distance (from start) versus time, with lines representing each runner's movement created in synchronization with the simulation. Students can change the speed of the runners by changing the slope of their associated lines, and they can also create more complex motions comprised of linear segments. Additionally, students can run the simulation in discrete steps and drop marks in order to apply numerical reasoning to the situation. Tables and formulas are also available. Velocity graphs, showing plots of speed versus time, are not included in the intervention in this study, but they are an important part of building an intuitive understanding of calculus. Fuller descriptions of the software are available elsewhere (Kaput & Roschelle, 1998; Roschelle & Kaput, 1996; Roschelle, Kaput, & Stroup, 2000).

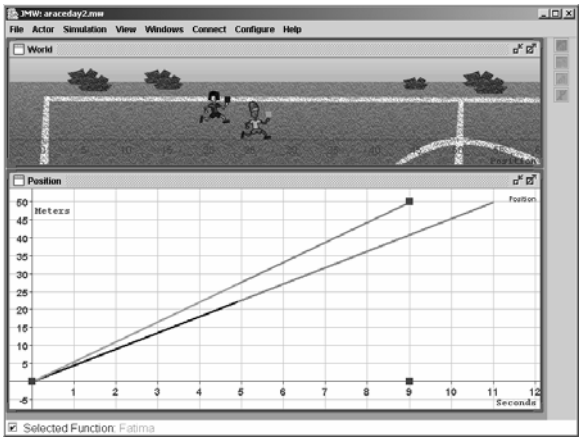


**Figure 1. MathWorlds simulation and line graph windows**

*Innovative Integration of Technology and Curriculum*

For the SimCalc intervention, we use a replacement unit strategy, which has been shown to be somewhat successful at a large scale in mathematics reform in California (e.g., see Cohen &

Hill, 2001). Our replacement unit, *Managing the Soccer Team*, takes a linear function (y=kx) approach to rate and proportionality (Greenes & Findell, 1999), exploiting SimCalc curriculum and MathWorlds software. Through examination of a number of situations involving motion or accumulation, students learn about how y=kx can be used to express change, how k describes the rate of change, the connections among representations of that change, and how to solve problems integrating all of these.

### *Experimental Design*

Scaling Up SimCalc uses a delayed treatment design (Campbell, Shadish, & Cook, 2001; Slavin, 2002), in which teachers are randomly assigned to one of two conditions: Immediate Treatment (IT) and Delayed Treatment (DT). This design affords the implementation of a control group, as well as equity among all research participants. Both the IT and DT groups receive treatment interventions; however, the treatment for the DT group is always delayed by one year. The DT group therefore serves as the IT's control group in any given year. To contrast with the impacts of the treatment intervention, the experimental control is designed to have teachers teach comparable material as they would under typical circumstances.

The basic experimental design is as follows (Figure 2). Teachers attend a series of professional development workshops (see Table 1 for an overview of their purposes) and then implement the replacement unit in their classrooms. In the summer, both groups attend a 2-day preparatory workshop. After this workshop, DT teachers have completed their summer participation, while IT teachers attend a 3-day SimCalc workshop. In the fall, IT teachers come together once again for a weekend planning workshop. During the school year, IT teachers are asked to teach the replacement unit, *Managing the Soccer Team*, in the place of their usual rate and proportionality unit, while DT teachers are asked to teach rate and proportionality as usual.

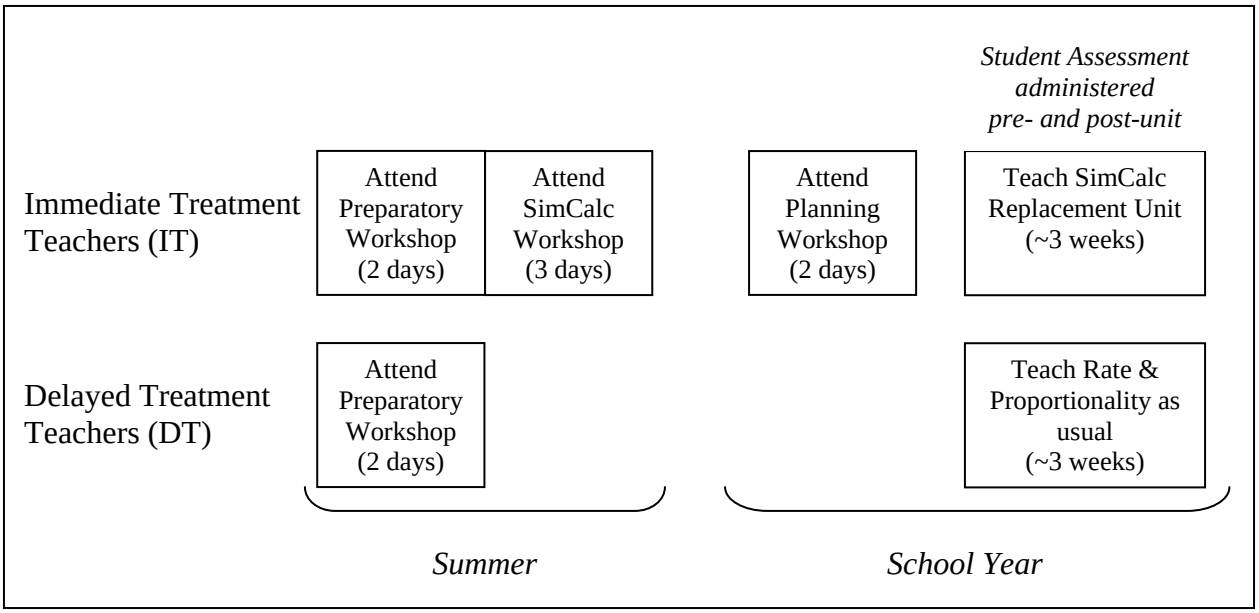


**Figure 2. Scaling Up SimCalc first-year research design**

**Table 1. Goals of the three types of teacher professional development workshops**

| Workshop | Purpose |
|---|---|
| Preparatory (2 days) | • To introduce the reframing of the traditional approach to proportionality<br>• To prepare teachers to use technology in their instruction of proportionality |
| SimCalc (3 days – IT only) | • To deepen teachers' knowledge of CCDM for proportionality<br>• To familiarize teachers with the SimCalc MathWorlds software<br>• To familiarize teachers with the scope and sequence of the SimCalc replacement unit |
| Planning (2 days – IT only) | • To plan in detail how the teachers will implement the SimCalc unit during the school year |

In subsequent years, IT teachers are presented with further opportunities for training, and DT teachers always receive identical opportunities with a 1-year lag. In this paper, we address the first year only.

The student assessment is administered in both conditions pre- and postunit. It is designed to measure learning gains over the 3-week units taught in both the IT and DT conditions. Other key measures include teacher math knowledge, classroom observations, teacher logs during the unit, surveys about teachers' attitudes toward teaching and student capabilities, and in-depth yearly

interviews with teachers about their experiences in the project.

Our research has progressed through two phases, both using this research design. Phase I, implemented during the 2003-2004 school year, was a small-scale pilot ($n$ = 21 teachers total) in which we established our curriculum, research design, recruitment strategies, and instruments. Phase II implementation will begin in summer 2005 and will be at full scale ($n$ = 140 teachers total).

## Approach to Assessment Development

The overarching goal of our assessment design process was to create a valid and reliable instrument that could be used to assess conceptually simple and complex knowledge of rate and proportionality. Although we could draw items from several standardized tests and instruments in the literature that capture various aspects of this content, we had to build our own instrument to fully embody the target construct. We describe our general approach to validation below. In the next sections, we describe our assessment development methodology and findings.

Drawing from the Standards for Educational and Psychological Testing (AERA/APA/ NCME, 2002), we approach validation as the establishment of a set of evidentiary arguments. Table 2 outlines the set of arguments, primary concerns, and sources of evidence we decided were important for this assessment.

The first argument is content validity: evidence obtained from an analysis of the relationship between the test's content and the construct it is intended to measure. We began by mapping out the conceptual framework of the assessment – what, operationally, the assessment would cover. We decided that the validity of the conceptual framework of CCDM would be best established by using the judgments of an expert panel of mathematicians and math educators. Given this conceptual framework, we would then need evidence that our specific assessment items were

**Table 2. Evidentiary arguments for validity of the assessment instrument**

| Type of Argument | Primary Concerns | Sources of Evidence |
|---|---|---|
| Content validity | Knowledge and skills assessed in items address CCDM for proportionality | Operational definition of CCDM for proportionality by expert panel of mathematicians and math educators |
| | Alignment of items with SimCalc conceptual framework | Formative expert panel review |
| | Alignment with Texas state standards | Formative expert panel review |
| | Grade-level appropriateness | Formative expert panel review |
| Construct validity | Appropriateness of response processes to intended construct definition | Think alouds |
| | Internal consistency of relationships of total and subscale scores to individual items within a test | Cronbach's alpha of complete test and $M_1$ and $M_2$ subscales |
| | Internal structure | Confirmatory factor analysis to validate $M_1$ and $M_2$ constructs |
| Instructional sensitivity | Capacity to detect change in knowledge relevant to instruction | Gains on items and the test as a whole pretest to posttest in the IT condition |
| Discriminant validity | Capacity to discriminate performance of students in the IT versus DT groups | Differences in gains between the groups |

aligned with the framework's content, the Texas state standards, and the type of content that is considered appropriate for assessments for seventh grade students. We decided that expert panel evaluation of the items, with respect to our conceptual framework, would be the most compelling evidence.

The second argument is construct validity: evidence that the test scores are to be interpreted as indicating the test taker's standing on CCDM. We decided to evaluate construct validity with three sources of evidence. First, to determine the appropriateness of response processes to the

intended construct definition, we would conduct think alouds to make sure that the items evoke the appropriate proportional reasoning. Next, we would empirically derive internal structure (factor analysis), as well as investigate the internal consistency of the test and subscales (Cronbach's alpha).

The third and fourth arguments establish the assessment's capacity to detect learning in the IT group as compared with the DT group. Instructional sensitivity, or the capacity of the item to detect change in knowledge relevant to the replacement unit, would be evaluated through gains on items and the test as a whole, pretest to posttest in the IT condition. Discriminant validity, the capacity to discriminate performance of students in the IT versus DT groups, would be evaluated by examining differences in gains between the groups in the experiment. We would expect that the DT group would demonstrate less gain in knowledge.

To create an instrument with these characteristics, we followed an iterative assessment design process through Phase I and Phase II of the experiment (Figure 3). For Phase I, we created a prototype assessment to be used in the pilot. We began by establishing our conceptual framework, the operational definition of $M_1$ and $M_2$. Next, we formulated an assessment blueprint, a set of specifications for the types of items (content and format) needed to construct the test. We then collected a pool of items from various sources that were reviewed by the members of the formative expert panel for various characteristics essential to our validation process. These items were subjected to cognitive think-alouds and field testing. The revised items were included in a prototype instrument in the Phase I pilot study.

Currently, in Phase II, we are refining our prototype to create a final instrument. We evaluated our evidence and process and made some essential refinements. Specific refinements were strengthening of our conceptual framework and a more rigorous approach to content

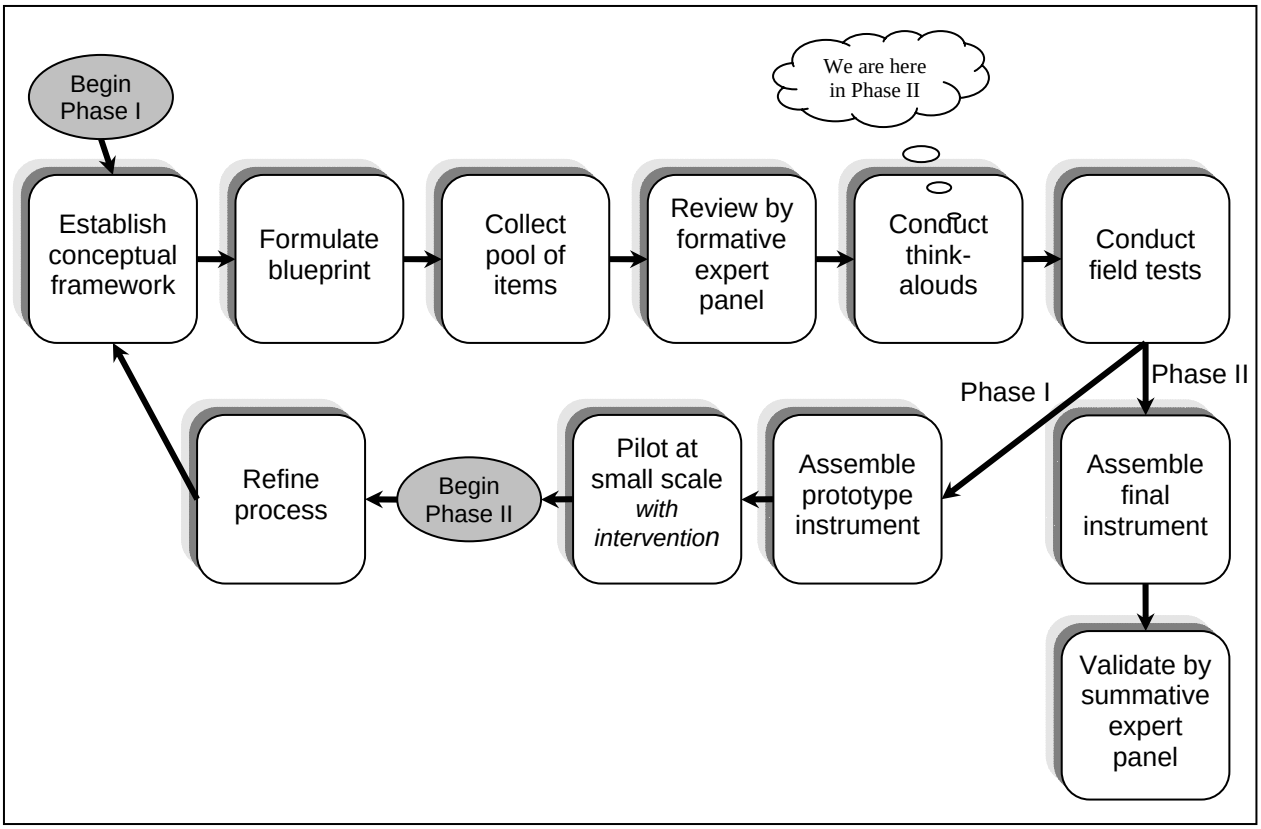


**Figure 3. SimCalc's iterative design process for the student assessment**

validity. We are doing each of the first six steps of the process again. We are currently in the process of doing think alouds and field testing our instrument. We will convene a final expert panel in the summer to provide summative evidence of the revised items' content validity. The refined instrument will be used in the Phase II experiment beginning in fall 2005.

## Prototype Assessment Development in the Phase I Pilot Study

In this section we describe the evolution of the prototype assessment in Phase I and evidence for its validity. We followed the first six steps of the assessment development process, assembled the prototype instrument, and then piloted a prototype assessment in the Phase I randomized experiment pilot.

## *Method for Phase I Assessment Development*

### *Establish Conceptual Framework*

In Phase I, we began our assessment development process with the initial working definition of CCDM for proportionality indicated in Table 3.

**Table 3. Initial working definition of CCDM for proportionality**

| |
|---|
| **$M_1$ – Conceptually Simple Proportionality**<br>Mathematical ability to apply routine procedures as assessed through calculation and multiple choice. These would include all the nonredundant rate and proportionality items released on the TAAS Web site. |
| **$M_2$ – Complex and Conceptually Difficult**<br>The more advanced mathematical ability to form mathematical connections among ideas as assessed through students' expression of mathematical ideas in symbols, graphs, diagrams, and words in addition to multiple choice and calculation. By "connections," we mean to include connections between mathematical ideas, across representations, and to realistic situations. |

### *Formulate Blueprint*

The Phase I blueprint had four dimensions. The cells in this complex four-dimensional matrix were used to guide us toward coverage of important math. The dimensions are described in Table 4. In this early stage of our assessment development process, we considered $M_1$ items to be any in the blueprint matrix that originated from the TAAS exam. $M_2$ items were therefore those that dealt with the more complex mathematics beyond what was tested on the state exam.

**Table 4. Phase I dimensions of the assessment blueprint**

| |
|---|
| 1. Conceptual knowledge and skills, including problems that involved unit rate, speed, the formula y = kx, graph reading, rate as slope, multiple representations, piecewise graphs, and the definition of proportionality |
| 2. Mathematical context, including money and motion |
| 3. Representation type, including graph, table, narrative description, and formulas |
| 4. Task types, including identify, calculate describe/explain, and construct |

### *Collect Pool of Items*

Guided by our blueprint, we searched for candidate items in relevant sources, including released TAAS items from previous years, items used in early SimCalc design research, items

from the rate and proportionality literature, items from the math of change and variation literature (e.g., Carlson, Jacobs, Coe, Larsen, & Hsu, 2002), and items adapted from the SimCalc unit. The resulting pool had 35 items that were dichotomously scored as correct or incorrect.

*Review by Formative Expert Panel*

These items were then given to an expert panel for review. A panel of three math education experts considered each item on three dimensions. They rated the difficulty level, clarity, and degree to which the item was aligned with knowledge in the SimCalc unit and rate and proportionality. All ratings were made by consensus in group discussion; thus, there was no check for reliability. We then chose from the pool of items those rated as having a range of difficulty, high clarity, and high relevance.

In Phase II, we decided to place much greater emphasis on expert panel review as a source of evidence for validity. We created a much more rigorous and quantitative process for this review. This is described in detail below.

*Conduct Think Alouds*

To yield information about how individual students would solve the problems, we conducted think-alouds with six students from a local (Massachusetts) class. These six were composed of two high, two middle, and two low math achievers, as rated by their math teacher. We used these data to eliminate items that were too easy or difficult, eliminate items that could be solved successfully by strategies other than proportional thinking, and clarify items when necessary.

*Administer Field Testing*

These items were then administered to two classes (55 students total), each with a wide distribution of student achievement levels (based on teacher report). We used item data from this administration to select items that could measure a range of student performance. We eliminated

items that the majority of students got correct. Because the instrument is intended to measure growth in knowledge, items that had a low percentage correct were kept.

*Assemble Prototype Instrument*

Using data from the formative expert panel review, think alouds, and field testing, we created a prototype instrument with coverage of the dimensions specified in the original assessment blueprint. The resulting instrument had 27 items total, 9 $M_1$ and 18 $M_2$.

*Pilot at Small Scale*

The Phase I pilot of the full randomized experiment took place during the 2003-2004 school year. There were 21 middle school math teachers, 11 in the IT group and 10 in the DT group (assignment was random except for two minority and rural teachers who were explicitly assigned to the IT group to ensure representation). As shown in the research design in Figure 2, the assessment was administered to students by their teachers pre- and post-unit. We describe below the pilot with respect to validation of our assessment instrument.

Table 5 describes the attributes of the students in the sample, as reported by their teachers. Students represented are those who turned in both pretest and posttest. Other students in the class may have been absent on the days these tests were administered. These data were provided by teachers at the beginning of the school year. Data are missing for 18% of the students, primarily because of student turnover after the beginning of the school year.

The 714 pretests and posttests were scored by two undergraduate math education majors. Since all items were scored dichotomously, there was little training necessary. Scorers were blind to participant condition. To check for reliability, 89 assessments (12.5%) were scored by both scorers. Their agreement was 97.5%. Most discrepancies were due to one particular item. Instructions for scoring this item were modified midcourse for greater clarity.

**Table 5. Attributes of student sample**

| Variable | Delayed Treatment (DT) | Immediate Treatment (IT) |
|---|---|---|
| N | 176 | 181 |
| Mean number of students per class | 17.6 | 16.5 |
| Gender (%) | | |
| Female | 47.2 | 55.8 |
| Male | 52.8 | 44.2 |
| Student achievement, rated by teacher (%) | | |
| Low | 18.5 | 41.8 |
| Medium | 50.3 | 41.1 |
| High | 31.1 | 17.0 |
| Ethnicity (%) | | |
| White | 50.6 | 44.0 |
| African American | 14.3 | 7.3 |
| Hispanic/Latino/Latina | 30.5 | 47.3 |
| Asian/Pacific Islander | 1.9 | 1.3 |
| Native American | 2.6 | 0 |

### *Evidence for Validity of the Prototype Assessment in Phase I*

Table 2 outlined the evidentiary arguments we use to support the validity of the instrument. We now consider each type of evidence collected in Phase I.

#### *Content Validity*

In Phase I, we began with a working definition of CCDM for proportionality and got informal feedback from a panel of experts. We did not, however, follow principled processes for operationalizing the conceptual framework, validating whether the content was appropriate, or evaluating the alignment of items with this framework, Texas state standards, or grade-level expectations. These weaknesses are addressed in Phase II.

#### *Construct Validity*

To address construct validity, first, we used cognitive task analyses to determine whether students used proportional reasoning in their strategies for solving the items. We used only items that evoked this type of reasoning and could not be solved with an alternate algorithm.

Second, we examined internal consistency of the whole test and subscales. When administered at baseline, test scores from the 357 students showed that the 27-item assessment had a Cronbach's alpha of .82 when administered at baseline. The 9-item $M_1$ subscale had a Cronbach's alpha of .71, and the 18-item $M_2$ subscale had a Cronbach's alpha of .76. These alphas are acceptable for unidimensional scales of achievement.

Given these findings, we did a confirmatory factor analysis to examine the empirical internal structure of the assessment. Using principal components analysis, we found a solution with one factor that accounted for 20% of the variance and several other factors accounting for less than 7% each. Factor loading did not correspond to our $M_1/M_2$ distinction. We find these data difficult to interpret and plan to do additional exploration of internal structure of our refined instrument in Phase II.

*Instructional Sensitivity*

For instructional sensitivity, we consider whether there are pre-post gains in the IT group. Table 6 shows the scores. We found that students who had the SimCalc replacement unit showed considerable gains on the assessment. This difference was most notable for the $M_2$ subscale. This indicates that the assessment, particularly the $M_2$ items, is sensitive to the SimCalc instruction.

**Table 6. Test scores in the IT group (*n* = 181).**

| | **Test Scores** | | |
|---|---|---|---|
| **Subscale** | **Pretest mean (*sd*)** | **Posttest mean (*sd*)** | **Difference mean (*sd*)** |
| All items (27 total) | 8.9 (4.1) | 15.0 (5.2) | 6.1* (4.4) |
| $M_1$ items (9 total) | 5.2 (2.2) | 5.8 (2.2) | 0.6* (1.8) |
| $M_2$ items (18 total) | 3.7 (2.5) | 9.1 (3.6) | 5.5* (3.6) |
| *All difference scores were significantly greater than 0 ($p < .0001$). | | | |

We also found that individual items on the $M_1$ scale were not necessarily strong enough to detect differences in performance. We found that our nine $M_1$ items ranged in percentage correct from 44% to 81% at baseline, indicating that their level of difficulty was too low for a test

intended to measure growth in knowledge. On the other hand, with the exception of three of the 18 items, $M_2$ items ranged in percentage correct from 3% to 48% at baseline, indicating a capacity to measure a range of student performance.

*Discriminant Validity*

To check for discriminant validity, we consider whether the test showed discrimination of learning gains between the IT and DT students (i.e., students taught the SimCalc replacement and students taught the usual rate and proportionality). As Figure 4 shows, students of teachers in the IT group showed greater growth in knowledge than students of teachers in the DT group. An ANOVA of student difference scores with factors condition (IT versus DT) and teacher nested within condition showed a significant difference in gains between the groups [$F(1,282) = 126.7$, $p < .0001$]. The gain was mostly on $M_2$. An ANOVA of difference scores (again teacher nested within condition) of $M_2$ was significant [$F(1,282) = 178.0$, $p < .0001$]; for $M_1$ this statistic showed no significant difference [$F(1,181) = 1.2$, $p = .28$, n.s.]. The effect size for the combined $M_1$ and $M_2$ gain in the group that used SimCalc was 1.01. This evidence supports the discriminant validity of the assessment.

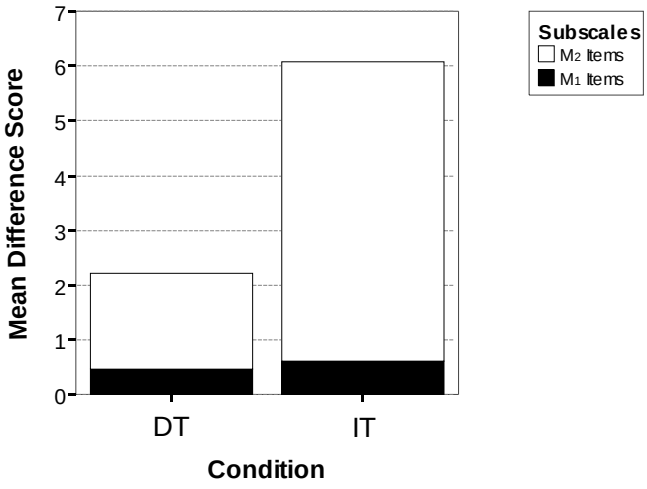


**Figure 4. Student gains (posttest minus pretest) across treatment groups**

## Assessment Refinement in Phase II

In Phase I, we created a prototype assessment instrument with several sources of evidence for validity. We collected evidence of the instructional sensitivity and discriminant validity of many of our items. Phase I data for content and construct validity was not as strong as we would like, and in Phase II we improve upon these aspects of validity. As Figure 3 illustrates, transitioning to Phase II, we build on our data and processes to refine and strengthen our instrumentation. The Phase II instrument development process is currently in progress. Implementation of classroom instruction will begin in the 2005-2006 school year. We describe below our method in progress.

### *Method in Progress*

#### *Establish More Operational Conceptual Framework*

To operationalize CCDM for proportionality, we assembled an expert panel of five mathematicians and math education experts. The first goal was to distinguish simple and complex proportional reasoning. The second goal was to extract the key facets of knowledge necessary to both a simple understanding of the relationship $a/b = c/d$ and the understanding of the formula $y = kx$ as a continuous function. Table 7 shows the conceptual categories.

We also add emphasis of alignment with the Texas Essential Knowledge and Skills (TEKS). A stated goal of the introduction to the seventh-grade mathematics TEKS is that "students use algebraic thinking to describe how a change in one quantity in a relationship results in a change in the other; and they connect verbal, numeric, graphic, and symbolic representations of relationships." There are several specific knowledge and skills that are aligned to simple proportional reasoning and CCDM for proportionality. Through expert panel review (see below), we selected only items that address these relevant TEKS.

**Table 7. Operational definition of CCDM for proportionality in Phase II**

| | |
|---|---|
| **$M_1$ – Conceptually Simple Proportionality** | |
| *Solving for a specific value* | |
| A. Solving problems using the formula a/b = c/d | Simple a/b = c/d problem in which three of the values are provided and the fourth must be calculated or the proportion must be recognized. |
| B. Solving unit rate problems with y=kx or d = rt | Simple y = kx or d = rt problem in which two values are provided and a third must be calculated (even if it is based on own prior work). |
| *Reading a specific value* | |
| C. Basic graph reading of linear relationships | • Reading values at specific points, without interpreting their meaning as a rate.<br>• Using the labels of axes to determine the meaning of a given pair of x, y coordinates.<br>• Sketching or plotting a given pair of x, y coordinates. |
| D. Basic table reading of linear relationships | Given a particular value, find the corresponding value in a table of a relationship. |
| **$M_2$ – Complex and Conceptually Difficult** | |
| E. Solving problems that invoke the function y = kx | Reasoning about a representation (e.g. graph, table or y = kx formula) in which a multiplicative constant "k" represents a constant rate, slope, speed, or scaling factor across many pairs of values (three or more pairs) that are given or implied. |
| *Within representations* | |
| F. Algebraic expression | Interpreting the behavior of a proportional function represented by an algebraic expression; or constructing an algebraic representation of a proportional function |
| G. Table | Filling in table cells of a table with many (3 or more pairs) values that are related by the same constant of proportionality |
| H. Graph | Interpreting or constructing the graph of a proportional or linear function. |
| I. Graph with a piecewise linear function | Interpreting or constructing a piecewise linear graph (e.g., with respect to narrative description of change over time) |
| *Making connection(s) or comparison(s)* | |
| J. Across two or more functions | Interpreting, comparing, or constructing two or more linear or piecewise linear functions |
| K. Across multiple representations | Reasoning about the same proportional relationship across at least two of the following representations: graph, table, formula |
| L. Additive versus multiplicative | Distinguishing problems that require additive versus multiplicative strategies for solution. |

*Reformulate Blueprint*

We revised our blueprint to scaffold an assessment with stronger content and construct validity. The Phase II blueprint has four dimensions. Again, we do not aim to fill every cell in this complex four-dimensional matrix; rather, we use these dimensions to help guide coverage. The dimensions are described in Table 8.

**Table 8. Phase II dimensions of the refined assessment blueprint**

| |
|---|
| 1. Reliable classification as addressing the concepts in $M_1$ or $M_2$ according to our operational definition |
| 2. Alignment to relevant TEKS |
| 3. Mathematical context, including money and motion |
| 4. Task types, including identify, calculate, describe/explain, and construct |

*Collect New Pool of Items*

Guided by our findings in Phase I and the reformulation of the structure of our assessment, we created a new pool of items from several sources. The first source was our prototype assessment. On the basis of percentage correct and magnitude of pre-to-post gains, we selected items that showed strong instructional sensitivity and capacity to discriminate between the groups. From the 27-item prototype assessment, there were 13 items that met these criteria. A second source of items was proportionality items from released standardized tests. We evaluated released items from the 8th-grade NAEP, the 8th-grade TIMSS, the 7th-grade TAKS, California High School Exit Examination (CHSEE), and $8^{th}$- and $10^{th}$-grade Massachusetts Comprehensive Assessment System (MCAS). A third source of items was the rate and proportionality literature (e.g., Kaput & West, 1994; Lamon, 1994; Lobato & Thanheiser, 2002). Finally, guided by our CCDM definition, we generated some new items to fill empty cells in our assessment blueprint. The final pool contained 59 items.

*Review by Formal Formative Expert Panel*

In Phase II, we conducted a formal formative expert panel review to evaluate our 59 items

for alignment with our conceptual framework, alignment with the TEKS, and grade-level appropriateness. The panel met for one day and took place in Austin, TX.

*Ratings.* The panelists were divided into two panels. The Content Validity and Alignment (CVA) panel made two ratings for each item: (1) they selected which of the 12 CCDM categories (A through L in Table 7), if any, were addressed by the item; and (2) they selected which of five relevant TEKS, if any, were addressed by the item. The Grade-Level Appropriateness (GLA) panel made three ratings for each item: (1) appropriateness of the reading load, (2) appropriateness of the computation load, and (3) appropriateness of the graphics. As a formative panel, all panelists were encouraged to suggest modification to the items that would improve their clarity, validity, or grade-level appropriateness.

Because interrater reliability was critical in this process, the CVA panel was trained to rate with high interrater reliability. Training consisted of rating together 15 practice items. In order to guard against rater "slide," panelists also discussed their ratings to every fifth item. They were instructed not to change any previous ratings based on this discussion. Any rating of grade-level inappropriateness was considered a signal to reevaluate the item.

*Panelists*. There were six panelists, four on the CVA panel and two on the GLA panel. On the CVA panel was: (1) a professional development program coordinator with experience as a teacher and administrator; (2) a math education researcher with experience teaching high school and college mathematics, focusing on mathematical modeling in teacher and student learning; (3) a math education researcher with experience teaching college mathematics, focusing on district benchmarking and evaluation of math teacher professional development; and (4) an expert in standard and assessment alignment. On the GLA panel was: (1) a math teacher professional development developer and leader, and author of a mathematics textbook; and (2) a district

math/science instructional strategist with experience leading curriculum development and facilitating staff development.

*Procedure.* All six panelists met together in the morning. They were given an overview of the project, the assessment, the validation process, and the rating scales. The two panels then went to separate rooms and were trained separately on their respective rating scales. Both groups took approximately 7 hours to rate the 59 items.

*Interrater reliability for CVA.* We evaluated the interrater reliability of categorizing items into: (1) $M_1$, $M_2$, both $M_1$ and $M_2$, or neither $M_1$ nor $M_2$; and (2) aligned with TEKS or not aligned with TEKS. An item was considered "unanimous" if all raters agreed on the classification. One rater omitted ratings on seven items. In the case of a missing rater, agreement by the remaining three was also considered unanimous. An item was considered "majority rated" if there was exactly one minority dissenter. All other cases were considered "indeterminate." For $M_1/M_2$, 86% of the items were unanimous (64%) or majority (22%); only 14% were indeterminate. For TEKS alignment, 92% were unanimous (80%) or majority (12%); only 8% were indeterminate. Overall, this represents high interrater agreement.

*Results.* Using the CVA panel ratings, we selected items that met both of the following criteria: (1) unanimous or majority rated $M_1$, $M_2$, or both; and (2) unanimous or majority rated as aligned with TEKS. This rendered a total of 43 out of 59 items (17 $M_1$, 22 $M_2$, and 4 both). Of these 43 items, all but 10 were rated as grade-level appropriate by both GLA panelists with respect to reading, computation, and graphics. Of those rated inappropriate, all the ones rated as too easy were dropped, and all the ones rated as too difficult were modified to grade level according to the panelists' recommendations. We also made other minor modifications of the items based on the recommendations of both panels.

*Conduct Think Alouds*

We will conduct think-alouds and cognitive task analyses to yield information about the ways individual students solve each of the items. We will interview 12 students representing a range of achievement levels. These data will be used to help characterize the cognitive work required by each item, clarify appropriate items, and eliminate inappropriate items.

*Conduct Field Testing*

We will field test appropriate items with about 200 students. We will consider both classical test theory and item response (IRT) measurement models. We consider an IRT approach because that is the approach being used in our assessment of math knowledge for teachers (see Ball & Hill, n.d.).

*Assemble Final Assessment*

Using all sources of evidence collected in Phase I and Phase II, we will create our final student assessment instrument scaffolded by the requirements outlined in our Phase II assessment blueprint.

*Validate by Summative Expert Panel*

We will assemble as a panel the research advisors for our project, including experts in math education and research methodology. This panel will finalize the content and construct validity of our items by rating them in their final form on alignment with the SimCalc conceptual framework, alignment with TEKS, and grade-level appropriateness.

## Discussion

Although we are still in the process of creating our assessment instrument, in Phases I and II we have built much of our evidentiary base for its validity as an instrument that meets rigorous scientific standards and has the capacity to demonstrate growth in students' knowledge of

CCDM. In Phase I, we developed a set of items that were demonstrated in the pilot as having instructional sensitivity and discriminant validity. On reevaluation of our assessment development process, we decided to place greater emphasis on the content and construct validity of the instrument. In Phase II, we have strong evidence for content validity for the conceptual framework of our assessment (its creation by expert panel) and alignment of items with the SimCalc conceptual framework and Texas state standards (formative expert panel review). Formative expert panel review has also provided strong evidence for the grade-level appropriateness of our items. Data collected through the spring and summer will allow us to evaluate the appropriateness of response processes to intended construct definition, internal consistency of the refined instrument, and internal structure. A final expert panel review will allow us to evaluate all of the evidence for the validity of the final instrument as a whole.

As we discussed in the introduction, current benchmark assessments are not necessarily sensitive to efforts to increase access to deeper, more important, and more conceptually difficult mathematics. Merely adopting existing high stakes tests as outcome measures could dampen innovation that has real potential to democratize access to important mathematical concepts. However, establishing credibility for new measures is difficult and can be very expensive.

In this paper, we have shared our work to forge a sensible middle ground. We have undertaken a serious effort to build an appropriate new measure but at a cost that is still only a fraction of the cost of our overall research project. Our process draws on published test items but seeks deliberate alignment to the $M_1/M_2$ distinction we target. We invite discussion with the field to further refine this process and make it useful for other projects with similar goals. Simultaneously, we acknowledge the underlying distinctions between simple and complex, or between basic and proficient, or between routine and profound as not fully settled in the field.

We encourage a healthy discourse about the nature of richer mathematics that is appropriate at the middle grades, how we can measure it, and how we can create opportunities for all students to learn it.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME) (2002). *Standards for educational and psychological testing.* Washington D.C.: American Educational Research Association.

Ball, D., & Hill, H. (n.d.). *Developing measures of mathematics knowledge for teaching*. Retrieved December 2, 2003 from http://soe.umich.edu/lmt/researchitems/index.html

Campbell, D. T., Shadish, W. R., & Cook, T. D. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin Co.

Carlson, M. P., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education*, *33*(5), 352-378.

Cohen, D. K., & Hill, H. (2001). *Learning policy: When state education reform works.* New Haven, CT: Yale University Press.

Greenes, C., & Findell, C. (1999). Developing students' algebraic reasoning abilities. In L. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12* (pp. 127-145). Reston, VA: National Council of Teachers of Mathematics.

Hiebert, J., & Behr, M. (1988). Introduction: Capturing the major themes. In J. Hiebert & M. Behr (Eds*.), Number concepts and operations in the middle grades* (pp. 1-18). Hillsdale, NJ: Lawrence Erlbaum.

Kaput, J. (1994). Democratizing access to calculus: New routes using old roots. In A. Schoenfeld (Ed.), *Mathematical thinking and problem solving* (pp. 77-155). Hillsdale, NJ: Erlbaum.

Kaput, J., & Roschelle, J. (1998). The mathematics of change and variation from a millennial perspective: New content, new context. In C. Hoyles, C. Morgan, & G. Woodhouse (Eds.), *Rethinking the mathematics curriculum*. London, UK: Falmer Press.

Kaput, J., & West, M. M. (1994). Missing-value proportional reasoning problems: Factors affecting informal reasoning patterns. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics*. Albany, NY: State University of New York Press.

Lamon, S. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics.* Albany, NY: State University of New York Press.

Lobato, J. & Thanheiser, E. (2002). Developing understanding of ratio and measure as a foundation for slope. In B. Litwiller & G. Bright (Eds.), *Making sense of fractions, ratios, and proportions: 2002 yearbook*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Roschelle, J., & Kaput, J. (1996). SimCalc MathWorlds for the mathematics of change. *Communications of the ACM, 39*(8), 97-99.

Roschelle, J., Kaput, J., & Stroup, W. (2000). SimCalc: Accelerating students' engagement with the mathematics of change. In M. J. Jacobson & R. B. Kozma (Eds.), *Innovation in science*

*and math education: Advanced designs for technologies of learning.* (pp. 47-75). Hillsdale, NJ: Erlbaum.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Sowder, J. T., Wearne, D., Martin, G., & Strutchens, M. (in press). What grade 8 students know about mathematics: Changes over a decade. In P. Kloosterman & F. Lester (Eds.), *NAEP Mathematics assessments: Interpretive analyses and materials development*. Reston, VA: National Council of Teachers of Mathematics.

Suter, L. (2002). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research, 70*(4), 529-545.