The essential role of curricular analyses in comparative studies of mathematics achievement: Developing "fair" tests

Óscar Chávez, *University of Missouri — Columbia*
Ira Papick, *University of Nebraska — Lincoln*
Dan J. Ross, *University of Missouri — Columbia*
Douglas A. Grouws, *University of Missouri — Columbia*

Please address all correspondence to:
Óscar Chávez
University of Missouri
College of Education
121E Townsend Hall
Columbia, MO 65211-2400
578-882-4521
chavezo@missouri.edu

Paper presented at the
Annual Meeting of the American Educational Research Association

Denver, April 30–May 4, 2010

The essential role of curricular analyses in comparative studies of mathematics achievement: Developing "fair" tests[†]

Óscar Chávez, *University of Missouri — Columbia*
Ira Papick, *University of Nebraska — Lincoln*
Dan J. Ross, *University of Missouri — Columbia*
Douglas A. Grouws, *University of Missouri — Columbia*

In this paper we describe the process of development of assessment instruments for the Comparing Options in Secondary Mathematics: Investigating Curriculum (COSMIC) project. The COSMIC project was a three-year longitudinal comparative study focusing on evaluating high school students' mathematics learning from two distinct approaches to content organization (McNaught, Tarr, & Grouws, 2008). More than 90% of U.S. secondary schools follow a curriculum built around a sequence of three full-year courses, Algebra 1, Geometry, and Algebra 2 or Algebra 1, Algebra 2, and Geometry (Dossey, Halvorsen, McCrone, 2008). Integrated curriculum materials developed since 1990 have been adopted in some high schools. These materials integrate algebra and geometry content, together with functions, data analysis, and discrete mathematics each year of the secondary mathematics curriculum (Hirsch, 2007). We collected data in schools that were using both curricular approaches but with different groups of students. The study was conducted in six school districts in five states involving over 4,000 students.

Measures of student learning were central in the research design of our study. Developing and using instruments that focus on the specific aims of the programs being compared is essential in order to establish a causal link between the curriculum program and student outcomes. Nevertheless, developing assessment instruments responsive to the needs of mathematics curriculum comparison studies is challenging and poses unique difficulties. The research literature on test development usually focuses on large-scale assessments that are not directly linked to the content of specific curricula. For example, for an evaluation of the second edition of The University of Chicago School Mathematics Project curriculum (UCSMP), the evaluators used tests developed specifically for students using UCSMP, in addition to a standardized test, because the authors recognized the need for balanced test with respect to skills, properties, and real-world applications (Thompson et al, 2001, 2003, as cited in National Research Council [NRC], 2004). This example suggests that researchers have made use of *ad hoc* solutions in order to find adequate outcome measures. Moreover, there is a dearth of research to inform this kind of work. We found the process described in this paper successful and offer it as a framework on which to build better tests that can help establish fair comparisons between curricular programs.

[†] This paper is based on research conducted as part of the Comparing Options in Secondary Mathematics: Investigating Curriculum (COSMIC) project, a research study supported by the National Science Foundation under grant number REC-0532214. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

More specifically, this work presents a model sequence for developing high quality assessment instruments that accurately and validly measure students' mathematical learning in comparative research contexts. An ultimate goal of research in mathematics education is directly concerned with what students have learned and are able to do mathematically, and assessments provide one means of gathering this information.

The use of large-scale standardized tests in evaluations of curricular effectiveness is frequently questioned. Results from these tests may arguably represent an incomplete depiction of student learning. In this sense, standardized tests are "exceedingly blunt instruments for measuring what students might learn in a given year *from a given curriculum*" (Kilpatrick, 2003, p. 479, emphasis ours). The NRC (2004) concluded that "comparative studies need to attend more closely to… more precise measures of content-strand outcomes, especially in relation to curricular validity of measures" (p. 199). Nonetheless, standardized tests can be useful in comparison studies because of their psychometric properties. Webb (2003) suggests that standardized tests measure traditional concepts that are important for students to learn. It is also important to note that school districts are interested in the information such tests provide, for they offer outcome measures that are widely accepted by the public. For these reasons, we chose to use a standardized measure of achievement along with specially designed assessment instruments. In this paper, however, we focus on the process of designing our own assessments.

Developing assessments for COSMIC

Measures of student outcomes should include, according to the NRC (2004), "a variety of measures of the highest quality" (p. 165). It is further recommended that assessment instruments should include a variety of question types and should ensure curricular validity. We therefore sought to develop assessments that offer *curricular validity* (sensitivity to a curriculum's stated goals and objectives). By doing so, valid inferences could be drawn about what mathematics students learn from different approaches to curriculum organization.

In their analysis of extended constructed-response (ECR) tasks in the National Assessment of Educational Progress (NAEP), Silver, Alacaci, & Stylianou (2000) argued that large-scale tests, by relying almost exclusively on multiple-choice items, provide insufficient evidence of students' capacity for mathematical problem-solving and reasoning. On the other hand, they acknowledged that the design of some of the NAEP ECR tasks "seemed to constrain students' responses in ways that were not in line with the expectations evident in the scoring rubrics" (p. 303), hence their call for test designers to design better tasks and more appropriate scoring rubrics. In the following section, we provide a detailed description of the process we followed to develop tests that were largely based on ECR tasks, and rubrics that would enable us to analyze and understand what students learn.

Fair and conservative tests

In its examination of comparative curricular studies, the NRC (2004) explicitly recommended the use of multiple outcome measures. The committee found interesting the

approach used in the studies conducted by the developers of the UCSMP textbook series, in which *fair* and *conservative* test scores were reported. UCSMP developers made a distinction between these two types of scores, using *fair* to refer to the scores on topics that were taught, taking into account opportunity to learn; while *conservative* scores were scores on common items, that is, items taught in both programs under consideration. In this sense, the assessment instruments developed for the COSMIC project reflected a conservative approach because we chose to use items written around topics common to both curriculum programs. Nevertheless, we refer to our tests as *fair* tests because the tests were developed with the deliberate goal of not being biased towards either of the two curriculum programs studied. For each of the first two years of the study, two tests were developed: a test composed of items based on common topics (the fair test) and a test that assessed students' reasoning and problem solving skills. The items in the *reasoning* tests were based on topics that were appropriate to the grade level, according to the content in the textbooks, and as identified during our internal and external reviews. Furthermore, reasoning and sense making should be integral parts of all secondary mathematics (NCTM, 2009), so we developed tests that would allow us to collect data to document differences in student strategies and provide us insight into students' reasoning and sense-making.

Content analysis

Measurement experts agree that the content of an assessment instrument must reflect the content and processes of the subject being tested (e.g., Kline, 1986, 2000; Murphy and Davidshofer, 2005). Thus, the primary goal of our content analysis was to inform the development of fair assessment instruments, that is, instruments utilizing significant content common to the curricular materials involved in the study. Our content analysis involved understanding both quantitative and qualitative aspects of the textbooks used at each grade level in the study. In this paper we describe the process of test development in detail, referring to specific examples from our first year tests.

Characterizing the textbooks.

The analyses used to support the development of the assessment instruments for the study was conducted with the *Core-Plus (Course 1–Course 3)* textbook series (Coxford et al., 1998) and the *Glencoe Algebra 1, Geometry, and Algebra 2* textbook series (Holliday et al., 2005). *Core-Plus* was the textbook series used by all teachers of integrated mathematics classes in the COSMIC study. The teachers in subject-specific classes predominately used the *Glencoe* textbook series (which also owns the largest national market share of Algebra 1, geometry, and Algebra 2 textbooks in the U.S.), however, other subject-specific textbooks were also in use in classrooms included in our study: *Glencoe*, *Holt*, and *McDougal Littell*. We examined these books for differences among them in content and pedagogical approaches. Based on the number of classrooms using *Glencoe* in our study, the similarity among these subject-specific textbooks, and *Glencoe*'s larger market share in the US, we decided to conduct the content analyses on the Glencoe textbook series.

Lessons in the integrated mathematics textbooks are structured in a progression of learning stages beginning with a brief introduction of the concepts to be studied, moving on to in-depth explorations of those ideas, and then culminating in activities meant to give students opportunities to summarize and share the important ideas that emerged from the

lesson. Typical lessons span several days, and are based on the assumption that teachers will utilize collaborative learning strategies. Topics in discrete mathematics, data analysis, and modeling are central to the mathematical focus of the integrated mathematics textbooks examined, and technology is integrated throughout the course. Emphasis is placed on "real world" data driven problems, with formal definitions and theory being initially delayed and revisited later in the same course or in subsequent courses.

In contrast to the extended (multiple day) lessons common to the integrated mathematics textbooks, typical lessons in the subject-specific textbook series are designed to reach closure in a standard class period. Mathematical topics frequently begin with definitions, worked-out examples, and culminate with multiple opportunities to practice problems designed on variations of the given examples. Technology is not an integral part of the materials, but it is considered at various junctures within the materials.

Identifying common topics.

Year 1 Common Content

Our first step was to compare the table of contents of each textbook (*Core-Plus Course 1* and *Glencoe Algebra 1*) used in the first year of our study in order to determine significant content topics common to both textbooks. The substantive common topics that emerged from this comparison were: linear equations (slope, rate of change, graphing, etc.), data and statistics, linear inequalities, exponential functions, Pythagorean theorem, and introduction to matrices. Each of these textbooks also covered many other topics but were not common to both. For example, some topics on symbolic algebra appeared in *Glencoe Algebra 1* and not extensively in *Core-Plus Course 1*, such as: sequences, polynomials, factoring, quadratic functions, radical expressions, and rational expressions. Some topics covered in *Core-Plus Course 1* and minimally or not at all in *Glencoe Algebra 1* were: recursive formulas, simulation models and probability, elementary graph theory, and a variety of geometry concepts, such as some properties of polygons, three-dimensional shapes, symmetry, among others.

Year 2 Common Content

From our analysis of *Core-Plus Course 2* and *Glencoe Geometry*, we established that the main topics common to these two textbooks were: the geometry of the Cartesian coordinate system with concepts related to lines, distance, midpoints, and slope; transformations in the coordinate plane; perimeter, area, volume, and surface area; proportionality and similarity; and trigonometric ratios. We also examined the topics that these two textbooks had in common with the textbooks used in year 1. For example, among the topics we found in both *Core-Plus Course 2* and *Glencoe Algebra 1* were systems of linear equations, graphs of linear equations, quadratic equations, radicals and rational exponents, and the study of some basic polynomial functions. Some of the topics found in both *Core-Plus Course 1* and *Glencoe Geometry* were three-dimensional figures and nets, symmetry and transformations, Pythagorean theorem, and quadrilaterals.

Year 3 Common Content

Core-Plus Course 3 includes topics in algebra, geometry and trigonometry, and statistics and probability. It is particularly important to note that the concept of function,

used informally throughout *Course 1* and *Course 2*, is formalized in *Course 3.* An emphasis on modeling situations is prevalent throughout the different investigations within each unit of this course. In *Course 3* there are opportunities for students to use symbolic algebra and formal proofs in geometry, mostly around topics of congruence and similarity of triangles, topics that were found in the *Algebra 1* and *Geometry* textbooks.

In *Glencoe Algebra 2*, major topics deal with equations: linear equations, systems of equations, zeros of polynomial functions, rational equations. These topics are also common to *Core-Plus Course 3*. Other common topics to these two curricula are sequences and series, trigonometric functions, and exponential functions. Although both textbooks include probability and statistics, the emphasis given is quite distinct: in *Glencoe Algebra 2* only 1 out of 14 chapters is dedicated to probability and statistics, while 2 out of 7 units in *Core-Plus Course* focus on these topics. On the other hand, *Glencoe Algebra 2* includes chapters on matrices and conic sections, which are not included in *Core-Plus Course 3*, although *Course 1* includes some investigations where matrices are used. Both textbooks give particular weight to functions as a common thread throughout the different sections of each textbook.

There are some topics that appear in only one of the textbooks analyzed. For example, the integrated mathematics textbooks included lessons where graphs are examined in-depth. Problems from graph theory, graph-coloring, modeling problems with graph models, and applications to optimization problems are some examples of the tasks found in these lessons.

Among the topics that are included in the subject-specific textbooks, but not in the integrated mathematics materials we could find radical equations, division of polynomials, Cramer's rule, logarithms, remainder theorem, conic sections, binomial theorem, mathematical induction, geometry of circles (tangents and secants), and three dimensional figures (surface area and volume of cylinders, cones, and spheres, congruent and similar solids, coordinates in space). We should point out that several of these topics are included in *Core-Plus Course 4*, which was not part of our study.

While we can identify topics that appear in both sets of curriculum materials, there are important differences in extension, depth, emphases, and placement in the curriculum that create different experiences for students and must therefore be taken into account.

This process of purposeful comparison and identification of common topics allowed us to make informed decisions on what topics to include in our tests so that they would not be biased, deliberately nor inadvertently, towards either one of the textbooks.

Writing and adapting items

After the initial content analysis, we began the process of garnering and writing items. As we did this, our decisions were influenced by several factors, which are discussed below. We carefully examined test items that were included in both curricula. We examined items from large-scale assessments, including the Trends in International Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP), the Program for International Student Assessment (PISA), and the Balanced Assessment of Mathematics (BAM). We also gathered items from the NCTM Assessment Sampler (Dahl, Johnson, Morton, & Whalen, 2005), the Singapore curriculum (Yoong, 1999), Algebra Connections (Papick, 2006), and several websites that included released items for different state tests. We examined many items in order to identify items that addressed the common

specific objectives we had identified in our content analysis. Along with examining items, we also wrote items aimed at identified objectives. Kline (1986) recommended developing twice as many items as will be used; accordingly, we gathered or wrote at least one alternate for each item that we planned to use.

Linn, Baker, and Dunbar (1991) recommended involving subject matter experts in item writing. Consistent with this recommendation, we involved mathematics educators and mathematicians in the writing and reviewing of test items. In addition, we consulted an expert in Item Response Theory at various stages of the test development.

Selecting and designing items.

Our item selection and test construction was guided by the aspiration to assess students' understanding and knowledge of important mathematical topics. The goal of the assessments was to identify and measure student learning outcomes of each type of curriculum, and thus we needed detailed information about student understanding. The NRC (2001) advocates designing items that allow for collection of evidence of students' underlying cognition. In addition, the NCTM *Assessment Standards* (1995) endorse tests that assess "students' growth in mathematical power, [rather than] students' knowledge of specific facts and isolated skills" (p. 56). Royer, Cisero, and Carlo (1993) also comment that tests which only assess factual knowledge do not provide information about where students are along a continuum of skill development.

In connection to the preceding principle, rather than relying on selected response items, we chose to write constructed response items that require students to show their work and explain their reasoning. Although the NRC (2001) warns that poor quality items of any format cannot offer valuable information, constructed response items have the potential to provide greater opportunities for more direct insights into student thinking (Webb, 1992). Linn, Baker, and Dunbar (1991) argue that "many writers have criticized the limitations of multiple-choice tests" (p. 15) and that multiple-choice items "place too much emphasis on factual knowledge" (p. 19). Even Kline (1986), who reveals a preference for multiple-choice items, comments that open response items "actually involve some further element of reasoning... The free item is therefore richer and more difficult" (p. 52). Nevertheless, the use of open response items is not undisputed. Pearson and Garavaglia (2003) argue that multiple-choice items on large-scale assessments may generate equivalent information more efficiently than open response items. Nevertheless, an advantage of using open response items and asking for students to show their work and explain their reasoning in the tests developed for the COSMIC study is that it allowed us to closely analyze selections of responses for more direct insight into student thinking. As Gorin (2007) argued, "Good diagnostic items provide opportunities to observe the process of student responses and increase the amount of information available from student answers" (p. 175). Finally, because the students in the study also took the Iowa Test of Educational Development (ITED), which solely consists of multiple-choice items, selecting a different format allowed us to view student learning from multiple points of view.

Silver and Lane (1993) note the importance of avoiding items with construct-irrelevant easiness or difficulty. That is, if items or sets of items give clues or contain flaws that allow students to arrive at the correct answer without knowledge of the intended construct, then it is impossible to draw valid conclusions about the knowledge of the

students. Aspects of this position include reading and writing ability required, unfamiliarity of contexts, and space in which to work (Baker & Herman, 1983; Kline, 1986; Linn, et al., 1991; Webb, 1992). Thus, we provided ample space with each item for students to show their work and strove to produce items with clear and simple language set in contexts with which most students in the U.S. should be familiar.

We set many of our problems in realistic contexts. Students studying from the integrated mathematics curriculum materials have limited exposure to problems without any realistic context, thus problems without context might be considered biased against them. In addition, students studying from subject-specific texts have exposure to items both in contexts and not in contexts. Moreover, items in a contextual setting have the potential to provide a better measure of the transfer value of the concept being tested than do items devoid of context (NRC, 2000).

Linn, Baker, and Dunbar (1991) argue that if students have prior experience with an item or if items are similar to problems previously seen by students, then the item may be ostensibly easier. Consequently, we avoided items that came directly from either curriculum in order to prevent inadvertently measuring students' recall rather than their learning. However, we did select several items to use longitudinally in order to assist with studying students' progress over three years. These items retained their original phrasing each time they appeared. We assumed that most students would not be able to recall their answers to items from at least one year earlier.

With respect to technology, we decided to collect and write items which would allow calculator use, but which should not be decisively more difficult without access to a calculator. Linn, Baker, and Dunbar (1991) state that "… equitable access would be an important consideration in a calculator-active assessment" (p.17). We intended calculators to be used for our assessments as they were used in each classroom. Calculators are heavily integrated into students' learning through the integrated mathematics textbooks, thus the test would be biased against them if calculators were not allowed. However, although we could encourage that every student have equal access to calculators, we could not guarantee this. Thus, we sought problems in which the calculator would not have an influence on item response. We made similar considerations regarding other tools, such as compasses or rulers.

Our initial selection of items to pilot test followed discussions centered on the matters listed above, along with several other concerns. For the *fair* tests in the three years of our study (to which we will refer as tests A, C, and E), we selected items that matched the objectives we identified as common to both curricula. For these tests we selected items with an emphasis on linear function objectives, geometry, and functions, respectively. For the reasoning tests in years 1 and 2 of our study (hereafter, tests B and D), we selected items that required nontrivial mathematical reasoning, but were appropriate for the level of students in the study. We also selected reasoning items that allowed students to use mathematical concepts and methods that we expected them to have acquired through study of their curriculum and we made deliberate efforts to select items that did not display an obvious bias toward one type of curriculum. The number of items included in each assessment was commensurate with the amount of testing time that would be available. Within these bounds—appropriate level, unbiased content, and reasonable time demands—we selected items that would assess the associated constructs as broadly as

possible and provide subtle insights into students' differing levels of understanding relating to the underlying constructs, which we next describe more fully.

The tests were intended to allow inferences about student knowledge on constructs underlying the content of the tasks on the tests, rather than merely being able to make inferences about student ability only on the tasks themselves. Messick (1994) discusses the difference between a construct-centered approach and a task-centered approach. He writes that a construct-centered approach begins with consideration of "what complex of knowledge, skills, or other attributes… are tied to explicit or implicit objectives of instruction" (p. 16). Thus, the nature of the construct guides the writing, selection, and scoring of the tasks. These factors, along with how well the items represent all possible items, then influence the generalizability of results (Webb, 1992). We began with our focus on the constructs underlying tasks by conducting the content analysis. Our goal was not only to include items that we felt were valuable for students to be able to demonstrate their ability to complete, but also select items that represented aspects of the construct as well. We developed items to reveal students' conceptual understandings of the content. Our goals were to understand students' underlying knowledge and skills that allow them to successfully complete items.

Writing scoring rubrics

Several goals guided our development of scoring rubrics. We wanted scoring rubrics that would effectively assist scorers to interpret student responses in order to reflect what students knew about the underlying construct. For example, we wanted the scoring rubric to assign different scores to students who provided evidence that they had better understanding of a particular aspect of linear function compared to students who did not. We also wanted scoring rubrics to provide different student scores for subtle differences in understanding in order to make our tests sensitive to differences in student performance. As we constructed the scoring rubrics, we recognized the importance of achieving high levels of inter-rater reliability (Domino & Domino, 2006; Kline, 1986, 2000; Reynolds, et al., 2006). Linn, Baker, and Dunbar (1991) also emphasize that it is critical that scoring procedures reflect students' true capabilities and not perceptions or biases of the scorer. For portions of items where raters need to interpret student work or explanations, we examined samples of student responses to model anticipated responses in the scoring rubric, included wording in the scoring rubric to allow for any possible portrayal of correct responses, and held group discussions about interpretation of student responses to all items.

Reviewing process

The process of developing these tests had an iterative nature. Reviewing the test items at different points in the cycle was a central component, as described in the following sections.

Piloting items

Once items were selected for the assessments, in order to better understand how the assessments would work with students, we arranged small-scale pilot studies of the tests with fewer than 20 students each. These pilot studies helped us understand possible

weaknesses in item surface features, judge whether items addressed constructs that they were designed to assess, and provided rough estimates of timing issues and difficulty levels. Face validity, whether students believe the test appears to be a valid measure of what it claims to measure, was also valuable to ascertain because higher face validity suggests higher motivation for students to complete the test (Domino & Domino, 2006; Kline, 1986, 2000; Linn, et al., 1991; Murphy & Davidshofer, 2005; Reynolds, Livingston, & Willson, 2006). We asked students in the small-scale pilots to write comments and reactions to the items, which allowed us to understand better the assessments' face validity.

External review process

Concurrent with the small-scale pilots, external reviewers evaluated the items individually and the tests as entities. Linn, Baker, and Dunbar (1991) suggest that "systematic judgments of the quality of the tasks… are needed from subject matter experts" (p. 19), and other measurement experts also commend review of items by experts to support content validity (Domino & Domino, 2006; Kline, 1986, 2000; Reynolds, et al., 2006).

Our reviewers included notable mathematics educators and educational researchers, research mathematicians, large-scale mathematics assessment developers, and experts in high school mathematics instruction. Among the many test features, we explicitly asked the reviewers to provide feedback about clarity, quality, and weaknesses of individual items on the assessments and the alternate items. We also asked them to provide judgments about whether each item was a reasonable measure of its underlying construct, if the content was appropriate for the intended students, whether the test was potentially biased toward either curriculum, and whether students would have enough time to complete all items. We received valuable and frequently concurring feedback on items and overall assessments and the general consensus supported the content validity of the assessments and indicated bias in only a few select items, which were subsequently modified or deleted.

Rewriting items.

Item revision took place based on our continued discussion, analysis of items, continuing analysis of the curricula, and feedback from the small-scale pilots, COSMIC team members, and external reviewers. Revisions were made to surface features of items as well as content of items where there was evidence that an item was redundant, biased, or did not provide sufficient information about what was intended. Throughout our ongoing revision of items, we focused on ways to reduce biases toward any of the curriculum programs and ways to enhance quality and volume of information from students.

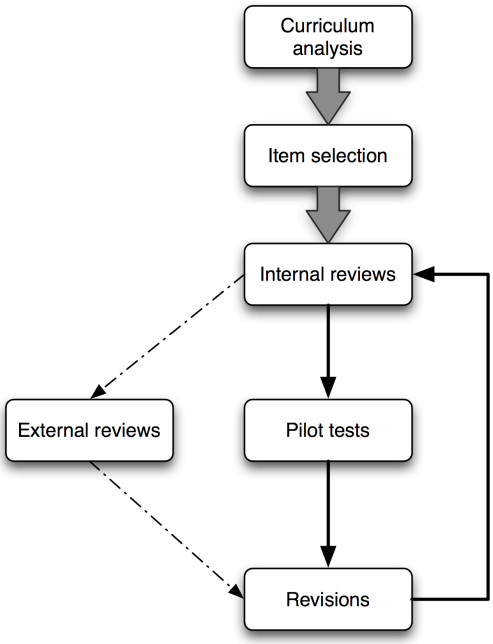
Following the external review and adjustment of items, the tests were piloted with students in numerous schools. Students in each pilot had experienced an appropriate level of curriculum for each test. Each test was piloted with more than 70 students. The tests were scored and results for the items were used to calculate the homogeneity and discriminatory power of the items. Responses to the items were used to calculate discrimination indices and difficulty levels of items (Domino & Domino, 2006; Kline, 1986;

Murphy & Davidshofer, 2005; Reynolds, et al., 2006). Figure 1 illustrates the cycle of review, piloting, and revision of the test items.

The scoring rubrics were refined in an iterative matter, following a process parallel to the development of the tests. The external reviewers examined the scoring rubrics and made some recommendations. Rubric developers first used the rubrics to score tests obtained from the pilot studies and then made revisions to the rubric.

Linn, Baker, and Dunbar (1991) argue that careful training is essential to reduce bias in scoring. As we drafted the scoring rubrics, a few scorers were trained in the use of the scoring rubrics. This served the purpose of informing the training process as well as an opportunity to examine the reliability of the scoring rubrics. These scorers scored copies of the same tests from students in the pilot studies. Their scores were compared with each other and with rubric developers' scores in order to identify any problems and ascertain a preliminary idea of the scoring reliability that the rubric would provide. Further revisions were based on this information. Once the scoring rubrics were finalized, no changes were made prior to their use in scoring the tests. This process of training, scoring, and refining informed our strategies for training the scorers of the tests.

Figure 1. Iterative process of writing and revising items for tests.



Developing an Item: An Example of Refining a Test Item

To illustrate the item refinement process, in this section we follow the development of one item. After the curriculum analysis, we selected the following item (Yoong, 1999, p. 167):

> *John bought a certain number of apples at 30¢ each and he had $3 left. If he bought the same number of pears at 40¢ each instead of the apples, he would be short $1. How many apples did he buy?*

In the original source, this item appeared in a section on writing equations. Although it was clear that students could use a system of equations to solve it, they could also use other strategies such as guessing and checking or creating tables. The external reviewers found this item appropriate, and they classified it as medium to difficult. Some of their comments:

> *While this item is a bit traditional, it is a good example of a systems problem that can be worked in a variety of ways. Students have to reason in a bit non-traditional way here as they have to think about the amount of money originally held by the person.* [Reviewer 1]

> *This is just a pair of simultaneous equations—is that in the curriculum or do they have to do it by trial and error—or some other reasoning?* [Reviewer 2]

> *A beauty. I really think it fits as an item to compare students from the two courses. Traditional algebra teacher would easily identify it as worthy. This one would be fun to compare student attacks to solution.* [Reviewer 3]

> *Also good, and can be solved in various ways.* [Reviewer 4]

After the first pilot test, we realized that the fact that the problem included apples and pears created some confusion among students. A revised version of the item changed the wording slightly:

> *John bought a certain number of apples at 30 cents each and he had $3 left. If instead, apples were 40 cents each, he would have been short $1. How many apples did he buy? Show your work.*

Based on results from a group of students with whom we pilot tested this item, we considered adding a simpler problem, making this a two-part item, so that students would have an opportunity to engage with the problem:

a) *Mary bought a certain number of pears at 25 cents each and she had $2 left. If she originally had $6, how many pears did she buy? Show your work.*
b) *John bought a certain number of apples at 30 cents each and he had $3 left. If instead, apples were 40 cents each, he would have been short $1. How many apples did he buy? Show your work.*

After a second pilot test we noticed that part *a* was unnecessarily easy and did not seem to help students to solve part *b*. So we left only part *b* as the final version of this item. Preliminary results from the actual tests suggest that this problem provides the kind of rich information that we were looking for. In a separate paper, we examine in detail the strategies used by students as they worked on this item (Ross, Reys, Chávez, McNaught, Grouws, in press).

The final rubric for this item can be seen in Figure 2. The rubric was developed through an iterative process along with the item in order to match changes in the problem and to best meet our goals for rubric development. As Silver et al. (2000) noted, scoring

rubrics may limit what is accepted as a correct response. Attending to the issues they raised, we made an effort to anticipate all strategies that students could use to solve the problem. Specifically, we sought to create a scoring rubric that allowed for any possible portrayal of correct responses. In the rubric, we explicitly included three different types of solutions: one algebraic, one based on using a table or guessing and checking, and one based on an arithmetic approach. We also included the option of "Other valid method" in order to allow the possibility of legitimate strategies beyond the three we anticipated that explicitly appear in the rubrics.

Figure 2: Rubric for the Apples problem

Student shows $0.3n+3=0.4n-1$ or equivalent OR Student shows $\begin{cases}0.3x+3=y\\0.4x-1=y\end{cases}$ or equivalent OR Student shows a correct table (or other representation) tracking changes or guesses and checks for number of apples and costs (or one correct guess of 40 **AND** a correct check that it works) OR Student shows $\frac{4}{.1}=40$ (or equivalent) **AND** justification OR *Other valid method*	3 pts
Partial Credit *(2 points)* *Student shows equation(s) above **WITH** one error (not computational error)* *OR* *Student shows a table (or other representation) tracking changes or guesses and checks for number of apples and costs with minor errors (e.g. an arithmetic error, entries off by one row, gives same sign for $3 left and $1 short)* (or one correct guess of 40 **AND** an incorrect check that it works) *OR* *Student shows* $\frac{4}{.1}=40$ *(or equivalent) **WITHOUT** justification*	
Partial Credit *(1 point each)* *Student shows one incorrect guess **AND** check* *OR* *Student indicates 3 - -1 = 4 or -1 – 3 = -4* *OR* *Student indicates .3 - .4 = -.1 or .4 - .3 = .1*	
Student provides correct answer (40)	1 pt
Total points	**4 pts**
Strategy Employed: A – equation(s) B – table/guess and check C – arithmetic D – other	

For the anticipated strategies, our goal was for the rubric to capture any differences in student understanding of the key concepts in the problem. Thus, the rubric included a separate score for the answer and three scoring categories for work. Each higher score indicates a higher level of conceptual understanding. Finally, we sought to minimize biases of the scorers and support a high level of inter-rater reliability between scorers by increasing the level of detail included in the rubric.

Scoring reliability

Prior to the scoring process, to measure inter-rater scoring reliability for each of the tests developed, we selected randomly 100 tests from the tests completed by the students who participated in the study. Copies of these tests were randomly placed back in the complete set of tests, so that each one of these tests was scored twice.

Having scored these 100 tests twice, we calculated the inter-rater reliability as a percent of agreement for each item for each test. We also calculated Cohen's κ (kappa), a statistic that takes into account the agreement that could be attributed to chance, and may therefore provide a more robust measure of agreement than simple percentages of agreement (Landis & Koch, 1977). The mean percent of agreement for each of the five tests was 96.5%, 96%, 97.3%, 94%, and 96%. That is, in average, in more than 94% of paired codings the scores were identical. All values of the kappa statistic showed at least a substantial agreement, and, for most of the items, almost perfect agreement (Landis & Koch, 1977). The results, after the tests were administered to thousands of students (around 2600 or more, depending on the year), showed that the scoring process using the rubrics developed was highly reliable.

Summary

Developing fair assessments

We have described how we developed the five project-developed tests. Each test was developed as described above, that is, as a result of curriculum analyses and several rounds of reviews and revisions. In addition, due to the longitudinal nature of the study, we decided that some items would be administered longitudinally, so that analyses of students' scores on these items could give us insight into growth over the three years of the study. In particular, the *apples problem* previously discussed in detail was one of those longitudinal items.

For year 1, when we tested primarily 9th-grade students, the test with common content, Test A, included nine items, all of them constructed response. The goal for Test A was to assess students' knowledge of mathematical content that was significant and common to the subject specific textbooks and the integrated textbook. The majority of Test A deals with linear relationships, a topic holding a central position in both Algebra 1 and integrated textbooks. The remaining items were based on the common topics of measures of central tendency, exponential relations, and the Pythagorean theorem. The problem on exponential relationships was used again in year 3. The *reasoning* test for year 1, Test B, included five problems. The content included data analysis, algebra, and geometry.

In year 2, the test with common content, Test C, included some items on algebraic topics, although it was focused primarily on geometric concepts (coordinate geometry,

perimeter and area, and trigonometry), and it had 10 items. The *reasoning* test, Test D, included five items, four geometric items and the longitudinal item, an algebra problem that was used in Test B.

In year 3, we developed only one test, Test E. Since, as discussed earlier, both textbooks give particular weight to functions as a common thread, the central topic in Test E is the concept of function and its different representations, besides items in symbolic algebra and one proof related to congruent triangles. It included a multiple-choice item and an item in which students had to match graphs of functions and their equations. The eight remaining items were constructed response. Three of these items were used in previous years' tests: the algebra problem used in tests B and D, and two items used in Test A.

The process of writing scoring rubrics and training scorers is of paramount importance when using extended constructed response items. Our scoring rubrics took into account a wide range of possible student responses and our scoring process was very reliable. In developing these tests, we addressed the NRC (2004) recommendations regarding comparative study design. Specifically, our project-developed tests have curricular validity and the student outcome data derived from these tests include measures that vary by question type.

Conclusions

We believe the data collected using the tests we developed will enable us to draw inferences about student learning because of the thorough process we followed. The development process ensured content validity of these instruments. Confirmatory factor analyses were conducted to examine construct validity. Our thorough process of scoring rubrics development, reviews, revisions, and pilot tests ensured that the scoring process was highly reliable. The assessment items have solid discrimination properties. These test data, used in conjunction with the curriculum implementation data, illuminate our understanding of what students learn from different ways of organizing mathematics content in secondary mathematics curriculum programs.

We believe that an iterative process of content analyses, identification of common topics, internal and external reviews, pilot tests, and revisions, such as the one we followed, has great potential for producing useful assessment instruments that can be used in curriculum comparison studies in ways that are fair to the curricula under study and useful in providing portraits of student learning from each curriculum. More rigorous research on mathematics curriculum should include multiple outcome measures, assessment instruments that are not biased towards any of the curriculum programs, as well as rich items that allow researchers to examine not only whether students are able to get correct answers but also show how they arrived at those answers. Learning what makes a curriculum effective, and under what circumstances it is effective, will require not one but a series of studies that build upon one another. We hope that the efforts described in this paper will contribute to this cumulative process and help to provide our field with the necessary evidence to answer important questions in mathematics curriculum research.

References

Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement, 20*(2), 149-164.

Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Burrill, G., Hart, E. W., et al. (1998). *Contemporary mathematics in context: A unified approach (course 1)*. New York: Glencoe/McGraw-Hill.

Dahl, T., Johnson, J., Morton, M., & Whalen, S. (2005). *Mathematics assessment sampler, grades 9-12: Items aligned with NCTM's principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction*. Cambridge: Cambridge University Press.

Dossey, J., Halvorsen, K, & McCrone, S. (2008). *Mathematics education in the United States 2008: A capsule summary book written for the eleventh International Congress on Mathematical Education (ICME-11).* Reston, VA: National Council of Teachers of Mathematics.

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173-201). New York: Cambridge University Press.

Hirsch, C. R. (Ed.) (2007). *Perspectives on the design and development of school mathematics curricula.* Reston, VA: National Council of Teachers of Mathematics.

Kline, P. (1986). *A handbook of test construction*. London: Methuen.

Kline, P. (2000). *The handbook of psychological testing* (Second ed.). New York: Routledge.

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

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

McNaught, M., Tarr, J., & Grouws. D. (2008). *Assessing Curriculum Implementation: Insights from the Comparing Options in Secondary Mathematics Project.* Paper presented at the Annual meeting the American Educational Research Association, New York.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (Sixth ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

National Council of Teachers of Mathematics (1995). *Assessment standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (2009). *Focus in High School Mathematics: Reasoning and Sense Making.* Reston, VA: Author.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

National Research Council (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 NSF-supported and commercially generated mathematics curriculum materials.* Washington, DC: National Academies Press.

National Research Council (2000). *How People Learn: Brain, Mind, Ex- perience, and School: Expanded Edition*. Washington, DC: National Aca- demies Press.

Papick, I. (2006). *Algebra connections*. Upper Saddle River, NJ: Prentice Hall.

Pearson, D. P., & Garavaglia, D. R. (2003). *Improving the information value of performance items in large scale assessments: NAEP validity studies*. Washington, D.C.: National Center for Education Statistics.

Reynolds, C. B., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education*. New York: Pearson.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*(1), 1-24.

Ross, D., Reys, R., Chávez, O., McNaught, M., & Grouws, D. (in press). Lessons learned from student strategies on an algebra problem. *School Science and Mathematics.*

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*(2), 201-243.

Silver, E. A., & Lane, S. (1993). Assessment in the context of mathematics instruction reform: The design of assessment in the QUASAR project. In M. Niss (Ed.), *Cases of assessment in mathematics education* (pp. 59-69). London: Kluwer Academic Publishers.

Silver, E. A., Alacaci, C., & Stylianou, D. A. (2000). Students' performance on extended constructed-response tasks. In E. A. Silver & P. A. Kenney (Eds.) *Results from the seventh mathematics assessment of the national Assessment of Educational Progress.* Reston, VA: NCTM.

Webb, N. L. (1992). Assessment of students' knowledge of mathematics: Steps toward a theory. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. Reston, VA: National Council of Teachers of Mathematics.

Yoong, W. H. (Ed.). (1999). *New elementary mathematics: Syllabus D, volume 1*. Singapore: Pan Pacific Publications.