Instructional Validity, Opportunity to Learn and Equity: New Standards Examinations for the California Mathematics Renaissance

CSE Technical Report 484

Bokhee Yoon
New Standards
Office of the President, University of California

Lauren B. Resnick
CRESST/University of Pittsburgh

July 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences. Lauren Resnick and Michael Young, Co-Project Directors, CRESST/University of Pittsburgh

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B6002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

INSTRUCTIONAL VALIDITY, OPPORTUNITY TO LEARN AND EQUITY: NEW STANDARDS EXAMINATIONS FOR THE CALIFORNIA MATHEMATICS RENAISSANCE

Bokhee Yoon
New Standards
Office of the President, University of California

Lauren B. Resnick
New Standards
Learning Research and Development Center,
University of Pittsburgh

Abstract

The study used data from the Spring 1996 administration of the New Standards Mathematics Reference Examination for Middle Grades to examine the instructional validity of the Reference Examination, opportunity to learn and equity in the context of the California Mathematics Renaissance program. Student and teacher responses to the opportunity to learn questionnaires and student performance were compared between the Renaissance and the non-Renaissance groups (called the *multi-state* group). Our results showed that Renaissance teachers had more opportunities to participate in reform-oriented professional development activities than teachers in a multi-state comparison group and that Renaissance classroom practice reflected these teacher learning opportunities. Renaissance students also showed significantly higher levels of performance on Skills and Problem Solving clusters of the New Standards examination compared to students in the multi-state group, after adjusting for student background characteristics. Furthermore, the achievement results on Skills showed a smaller performance gap between white and minority students in the Renaissance group than the multi-state group regardless of levels of SES (i.e., parent education). The results also showed a positive effect of reform-oriented instructional strategies on the outcomes in all three clusters. These results suggest the instructional validity of the New Standards Reference Examinations and a generally positive effect on equity of educational opportunity in the Renaissance program.

Introduction

Standards and accompanying assessments are being proposed today as instruments for raising academic achievement (e.g., Resnick & Resnick, 1991;

Shepard, 1995). The argument, broadly stated, is that if teachers and students know clearly what kinds of learning are expected they can direct their teaching and learning energies in a targeted way to meeting standards that will matter in their lives. Some (e.g., Howard, 1995; Resnick, 1995) argue that standards, with public examinations that students can study for, are key tools for making American education more equitable, for they will make it difficult to continue to exclude poor and minority students from opportunities to learn challenging academic material.

The recent growth in development and use of performance assessment is related in part to this vision. For assessments to be effective promoters of instructional change they must examine students on material of the complexity and depth specified in the standards. This calls for the inclusion in exams of extended tasks and constructed responses along with a much clearer relation to what is expected to be taught than is the case in most American testing. This new use of assessment as a legitimate target of instruction and learning efforts by teachers and students, demands attention both to the validity of the assessment in relation to the instructional program and to the ways in which the assessment is used to create opportunities for teacher and student learning.

We use the term *instructional validity* to refer to the extent to which an assessment is systematically sensitive to the nature of instruction offered. An instructionally valid test is one that registers differences in the amount and kind of instruction to which students have been exposed. Because of this sensitivity, an instructionally valid assessment might be expected to show somewhat lower effects of student background characteristics on achievement than tests that are not explicitly related to the curriculum. Instructionally valid assessments are thus important tools demonstrating improvements in the equitable distribution of teaching and learning resources.

The general concept of instructional validity, although not necessarily the term, is a familiar one in the assessment research community. Researchers have long recognized that the degree of overlap between the content tested and the content taught can have a strong impact on test scores (Airasian & Madaus, 1983; Anderson, 1990; Haertel & Calfee, 1983; Leinhardt & Seewald, 1981; Mehrens & Phillips, 1986). Degree of overlap is sometimes described as *opportunity to learn* what is tested. But in recent policy discussions opportunity to learn (OTL) has come to refer not only to the overlap between what has been taught and what is

tested, but to a more proactive concern with providing appropriate learning opportunities for all groups of students (Darling-Hammond, 1994; Guiton & Oakes, 1995; Porter, 1995; Stevens, 1993). An assessment—or, preferably, the standards an assessment is built to examine—might well become the focus for professional development programs aimed at changing teachers' capacity to teach the kinds of new, usually more challenging, concepts and skills that are the goal of much current education reform (e.g., Loucks-Horsley, Stiles, & Hewson, 1996; Resnick & Nolan, 1995; Wiley & Yoon, 1995). In such uses, the assessment is expected to lead instruction, not just reflect what is actually taught.

The instructional validity of a test is an aspect of its larger *consequential validity*. Since Messick (1989) first introduced the term, issues of consequential validity have received increasing attention in the educational measurement field (e.g., Messick, 1994, 1995; Popham, 1997; Shepard, 1997). Consequences are a logical part of the evaluation of test use; therefore, examination of effects following from test use is essential in evaluating test validity (Shepard, 1997). A test that is instructionally valid, in the sense of being systematically sensitive to differences in opportunity to learn can be further evaluated in terms of its consequential validity—that is, its effectiveness in leading teachers to spend time on classroom activities helpful to learning goals and responsive to individual student learning styles and needs (Darling-Hammond, 1994; Glaser, 1990).

This study examines issues of instructional validity of an assessment, opportunity to learn, and equity in the context of the California Mathematics Renaissance program. The Middle Grades Math Renaissance is a component of the California Alliance for Mathematics and Science (CAMS). The Middle Grade Mathematics Renaissance works with middle schools to help them transform mathematics programs so that all students are engaged in a thinking-centered mathematics curriculum. Employing a professional development strategy, the Renaissance works directly with teachers to help improve instruction. The Renaissance is committed to ensuring that all students, especially those traditionally placed at risk, have access to high-quality math education. Earlier evaluation of CAMS can be found in various reports (Derghazarian, 1996; Shields, Marder, & Wilson, 1996; WestEd, 1996; Yoon, 1996, 1997). During 1996 the Renaissance administered the New Standards (NS) Mathematics Reference Examination to students in participating middle schools as a basis for evaluating its effects on student achievement.

A number of factors such as the school curriculum, time devoted to mathematics instruction, types of typical classroom activities, and inservice education for teachers can be modified by the educational system. Other variables such as socioeconomic conditions and the level of parents' education are largely beyond the power of the schools to modify. This study examined the factors that are modifiable by the educational system and policy and evaluated student performance after controlling for the factors that are less modifiable.

Method

The New Standards Mathematics Reference Examination

The New Standards Mathematics Reference Examination is a performance on-demand assessment that is systematically referenced to the New Standards *Performance Standards* for mathematics (*New Standards*, 1997).[1] The *Performance Standards* identify eight standards:

Standard 1: Number and Operation

Standard 2: Geometry and Measurement

Standard 3: Functions and Algebra

Standard 4: Statistics and Probability

Standard 5: Problem Solving and Mathematical Reasoning

Standard 6: Mathematical Skills and Tools

Standard 7: Mathematical Communication

Standard 8: Putting Mathematics to Work

Of these, Standards 1 through 7 can be assessed in an on-demand setting, and tasks on the Mathematics Reference Examination are designed explicitly to assess them. For purposes of score reporting, Standards 1 through 7 are grouped into three standards clusters: *Conceptual Understanding*, covering Standards 1-4;

[1] The New Standards *Performance Standards* are built directly upon the consensus content standards developed by the relevant professional organizations. The Mathematics performance standards are based on the content standards produced by the National Council of Teachers of Mathematics [Commission on Standards for School Mathematics (1989), *Curriculum and Evaluation: Standards for School Mathematics*. USA: National Council of Teachers of Mathematics]. The *Performance Standards* consist of two parts: (1) *Performance descriptions* describe what students should know and the ways they should demonstrate the knowledge and skills they have acquired; (2) *Work samples and commentaries* are samples of student work selected for their capacity to illustrate the meaning of the performance descriptions together with commentary that shows how the performance descriptions are reflected in the work sample.

Mathematical Skills, covering Standard 6; *Mathematical Problem Solving*, covering Standards 5 and 7.

The Mathematics Examination calls for three class periods of testing, usually administered on three successive days. The 1996 NS mathematics exam consisted of open-ended constructed response tasks that were 2, 5, 15, and 45 minutes in duration. There were no multiple-choice items. Table 1 presents the configuration of the Middle Grades Mathematics Reference Examination. Descriptions of the nature of tasks that assess competence in the three standards clusters are provided in technical reports on the New Standards Examinations (*New Standards*, 1994, 1996).

Since no single task can adequately represent a standard, reporting a student's performance against a standard requires summarizing information from several tasks. The classification rules for determining standards levels are set through a process that involves qualified content and assessment experts in consultation with New Standards staff. There were three steps involved in assigning standards levels: The first step was the weighting of the tasks that make up a cluster. Judges assigned weights to each task on the basis of the task's centrality to the standard (the task's difficulty or score distribution was not considered). Each judge assigned weights independently and the weights were discussed until a consensus was reached. The second step was using the weights to create a weighted average score for each cluster of tasks. The weighted average scores ranged from 0.0 to 4.0 in 0.1 increments. The third step was setting cutpoints, or score levels that determine the absolute classification of the performance.

Table 1

1996 New Standards Mathematics Reference Examination (Middle Grades)

	Skills	Concepts	Problem solving	Total
Number of tasks	8	12	4	24
Time allotted for each task	2 minutes	2 - 5 minutes	5 - 45 minutes	2-45 minutes
Maximum possible score	32	48	18	98

Students receive an examination grade (called a *standards level*) on each standards cluster. Grades are reported in five categories: Achieved the Standard with Honors; Achieved the Standard; Nearly Achieved the Standard; Below the Standard; and Little Evidence of Achievement. The patterns of task scores resulting in these categories were established in a standard-setting exercise in which judges compared patterns of performance with the specifications in the *Performance Standards*. The standard-setting exercise was carried out prior to exam administration; judges had no access to comparative performance data in setting the grading levels.

The reliability of the examination was examined in both the reliability of the weighted means that were used to produce standards levels for the clusters and the accuracy and consistency of decisions based on the standards. Estimates of accuracy and consistency were made for the decision of whether a student had met the standard (called "Meeting the Standard") for each cluster reported in Mathematics. A student was said to have met the standard if s/he had achieved one of the top two standards level categories (i.e., Achieved the Standard with Honors and Achieved the Standard). A student was said not to have met the standard if s/he was in one of the bottom three categories (i.e., Near the Standard, Below the Standard, and Limited Evidence of Achievement). These definitions were applied to students' standards levels grades for the Skills, Concepts, and Problem Solving clusters.

The *accuracy* of the decisions is the extent to which they would agree with the decisions that would be made if each student could somehow be tested with all possible forms of the examination. The *consistency* of the decisions is the extent to which they would agree with the decisions that would have been made if the students had taken a different form of the New Standards Examination, equal in difficulty and covering the same content as the form they actually took. Additional analyses were performed to examine the overall reliability of the Reference Examination as well as the decision accuracy and consistency of the composite score. The "Meeting the Standard" cutpoint for this composite was defined as the sum of the "Meets the Standard" cutpoints of the clusters within the composite. For further details of the procedures employed, see the 1996 New Standards Reference Examination Technical Summary (*New Standards*, 1997) and Young and Yoon (1997). Table 2 presents the reliability of the 1996 Mathematics Reference Examination.

Table 2

Reliability Coefficients, the Decision Accuracies and Consistencies

		Meeting the standard	
	Reliability coefficients[a]	Accuracy(%)[b]	Consistency(%)[c]
Skills	0.85	90	84
Concepts	0.74	88	83
Problem solving	0.66	91	86
Mathematics composite	0.90	92	89

[a] Cronbach Alpha for clusters and stratified Cronbach Alpha for Math Composite were used in estimating the reliabilities.

[b] Accuracy is the percent of agreement between the classifications based on the NS Reference Examination actually taken and the classifications that would be made on the basis of the test takers' true scores.

[c] Consistency is the percent of agreement between the classifications based on NS Reference Examination actually taken and the classifications that would be made on the basis of an alternate form of the NS Reference Examination.

Opportunity-to-Learn Questionnaire

In conjunction with the administration of the Reference Examination, a set of questions was asked of teachers and students. These questions were adapted from the 1994 CLAS questionnaires (Wiley & Yoon, 1995). Students were asked about their background information (e.g., socioeconomic status [SES], ethnicity) and classroom activities. Teachers were asked about their classroom activities, instructional strategies and participation in staff development. The questionnaire items and the response patterns appear in Tables 5 through 8. Most of the questions about classroom activities and instructional strategies were aimed at identifying reform-oriented activities of the kind that Mathematics Renaissance aims to promote. The response categories for classroom activities were "Never," "Once or twice per semester," "Monthly," and "At least weekly." The response categories for instructional strategies were "Never," "Several times a semester," "Several times a week," and "Daily." The response categories for staff development were "Not at all," "Once," "2-5 times," and "5+ times/ Ongoing."

Study Design

The data used in this study were taken from the Spring 1996 administration of the New Standards Mathematics Reference Examination for Middle Grades. The exam was administered to middle grade students (mostly Grade 8) in a number of jurisdictions throughout the nation. The sample used for the present study includes 1,936 students and 105 teachers. Forty-three of the teachers were participants in California Math Renaissance. Their students numbered 673. The remaining teachers (62) and students (1,263) came from eight districts in several states. The students and teachers in the multi-state group form a comparison population for examining the instructional sensitivity of the New Standards Reference Examination along with the effects of the Math Renaissance professional development. The comparison group was formed by selecting, from among those students who responded to all questions on background characteristics and whose teachers could be identified, a sample whose background characteristics matched those of the Renaissance group. Only white, Hispanic, and African American students from both the Renaissance and multi-state groups were included in the analyses**.** Table 3 presents the description of the variables used in the study.

Our study design included three sets of data analysis. First, we compared the Renaissance and multi-state samples on the opportunity to learn (OTL) variables described earlier. Separate comparisons of student responses and teacher responses to the OTL Questionnaire items were made. Second, we compared overall achievement levels of the Renaissance and multi-state student populations on each of the three standards clusters. Third, we examined the issue of equity and instructional validity. We applied a Hierarchical Linear Modeling (HLM) technique in evaluating the effects of factors at both student and teacher (classroom) levels so that all estimated effects are adjusted both for individual students and classroom level influences on the outcome.

In education, data structures are often hierarchical. Students are grouped in classes and classes are grouped in schools. We have variables describing students and variables describing classes. When the nested structure of education is considered, it is important to think of ways in which statistical techniques should take this hierarchical structure into account. Traditionally, studies of the

Table 3

Description of the Variables Used in the Study

Variable	Description
Student-level variables	
SES (parent education)	1 = Not high school graduate; 2 = High school graduate; 3 = Some college; 4 = College graduate; 5 = Advanced degree
Minority	African American or Hispanic = 1; White = 0
Skills	Weighted average of task scores in Skills, ranges from 0.0 to 4.0.
Concepts	Weighted average of task scores in Concepts, ranges from 0.0 to 4.0.
Problem solving	Weighted average of task scores in Problem Solving, ranges from 0.0 to 4.0.
Teacher-level variables	
Group	Math Renaissance group = 1; Multi-state group = 0
STAFFDEV	An average of teacher responses on staff development: NCTM Standards, Group work, Problem Solving, Mathematics Concepts, Use of Technology in Teaching, Performance Assessment, Portfolio Assessment. Cronbach alpha = 0.88
TRADIT	An average of teacher responses on traditional teaching strategies: Lectures, Seat work, and Worksheets or Workbooks. Cronbach alpha = 0.74.
INSTACT	An average of teacher responses on instructional strategies and activities: Homework, Group assignments, Lab work, Field work, Performance assessment, Portfolio assignments, Draft and revision, Oral presentations, Portfolios, Student or teacher designed rubrics, Journals and logs, Open-ended questioning, and Student self-evaluations. Cronbach alpha = 0.82

correlates of achievement used standard regression techniques to summarize relationships between variables. Generally, standard regression techniques give standard errors that are misleadingly small when we fail to take into account the similarities or dependencies among observations within groups. Furthermore, we may overlook how relationships between variables of interest vary across organizational contexts (e.g., schools and classrooms) when we ignore the multilevel structure of educational data.

Hierarchical linear models help resolve this confounding by facilitating a decomposition of any observed relationship between variables, such as

achievement and social class, into separate student-level and school/teacher-level components (Bryk & Raudenbush, 1992). Hierarchical linear models also can explore "cross-level interactions" (Bryk & Raudenbush, 1992), in which a variable measured at a higher level interacts with a variable measured at a lower level (Raudenbush & Willms, 1991).

In the HLM analyses, student performance were compared using weighted means, rather than Standards Levels, for each cluster. The variables used in controlling the differences in student background characteristics between Math Renaissance and multi-state groups were *SES* (Parent Education; ranged from 1 to 5) and *Minority* (ethnic group membership, 1 = African American or Hispanic; 0 = White). The variables used to adjust for the differences among classes (or teachers) were Group (Renaissance or multi-state), STAFFDEV (teacher staff development), TRADIT (traditional teaching methods), and INSTACT (reform-oriented instructional strategies).

Results

Opportunity to Learn: Student Responses

Classroom activities. Table 4 summarizes student responses to the OTL questions. Substantially more Renaissance students than multi-state students reported at least weekly activities of *writing an explanation of how I solved a problem* (62% in the Renaissance group; 46% in the multi-state), *using a calculator to work on problems* (93% in the Renaissance group; 64% in the multi-state), and *working in small groups* (76% in the Renaissance group; 51% in the multi-state). Neither group reported using a computer frequently.

Most students in both groups reported activities of *making an oral presentation* and *working on problems or investigations that took from one to two weeks* at least once or twice per semester; however, one fifth of the multi-state students (23% and 20%, respectively) reported having never done these activities in their math class compared with a small percentage of the Renaissance students (4% and 6% respectively). Student responses for activities of *work on problems that can be solved in more than one way*, *solve problems using tables and graphs*, *working on projects that took from two to six weeks*, and *judging my own work* were similar in both groups. For the activity of *adding their work to portfolios*, more multi-state students responded

Table 4

Percentage of Students Reporting on Classroom Activities

	Math Renaissance				Multi-state			
Activities in math class	Never	Once or twice per semester	Monthly	At least weekly	Never	Once or twice per semester	Monthly	At least weekly
Work on problems that can be solved in more than one way	1	5	20	74	2	4	17	77
Solve problems using tables and graphs	0	11	55	34	4	15	43	38
Write an explanation of how I solved a problem	1	9	28	62	9	17	28	46
Use a calculator to work on problems	1	2	5	93	5	11	20	64
Use a computer to work on problems	58	27	13	2	77	12	6	5
Making an oral presentation	4	30	64	2	23	42	26	9
Working on problems or investigations that took from one to two weeks	6	29	58	6	20	37	33	10
Working on projects that took from two to six weeks	22	58	16	4	29	47	20	4
Adding my work to a portfolio	23	26	32	19	17	21	33	29
Working in small groups	0	3	21	76	4	13	32	51
Judging my own work	10	21	26	43	20	14	23	43

Note. The question in the survey: During this school year, how often have you done or participated in the following activities?

that they had had the activity in their math class than did the Renaissance students. In fact, this is the only activity for which more Renaissance students (23%) than multi state students (17%) reported no experience. Overall—with the exception of portfolios—the Renaissance students seem to have had more exposure to reform-oriented classroom activities compared with the multi-state students.

Opportunity to Learn: Teacher Responses

Instructional strategies and classroom activities. Table 5 reports percentages of teacher responses on questions about their instructional strategies. According to their self-reports, Renaissance teachers used lecture and seat work methods much less frequently (daily use only 2% and 7%, respectively) than did multi-state teachers (daily use 18% for lectures, 35% for seatwork). Renaissance teachers, on the other hand, reported using several strategies associated with constructivist mathematics learning more frequently than their multi-state counterparts did. These strategies included group assignments, lab work, and field work. Neither group of teachers reported much use of the draft and revision strategy, but Renaissance teachers used it slightly more frequently. Thirty-nine percent of the multi-state teachers (vs. 2% in the Renaissance group) reported that they had never used draft and revision in their math instruction.

Teachers in both groups reported giving homework to their students at least several times a week. All of the Renaissance teachers reported using performance assessment at least several times a semester, whereas 15% of the multi-state teachers reported that they never used performance assessment in

Table 5

Percentage of Teachers Reporting on Instructional Strategies

	Math Renaissance				Multi-state			
	Never	Several times a semester	Several times a week	Daily	Never	Several times a semester	Several times a week	Daily
Lecture	19	49	30	2	2	15	66	18
Seat work	26	12	56	7	0	16	48	35
Homework	0	7	51	42	0	11	44	45
Group assignments	0	16	49	35	3	60	23	15
Lab work	26	28	47	0	73	24	3	0
Field work (outside of classroom)	26	74	0	0	73	27	0	0
Performance assessment	0	81	5	14	15	61	23	2
Portfolio assignments	19	70	7	5	39	41	16	3
Draft and revision	2	84	9	5	39	46	10	5

Note. The question in the survey: How often do you use the following instructional strategies?

multi-state teachers reported that they never used performance assessment in their math classes. On the other hand, more teachers in the multi-state group (19%) reported giving portfolio assignments at least several times per week than did Renaissance teachers (12%); however, 39% of the multi-state teachers (vs. 19% in the Renaissance group) reported that they had never given portfolio assignments.

Table 6 reports percentages of teachers reporting on classroom activities used during their math classes. Renaissance teachers reported higher rates in all of the activities associated with active, problem-solving classrooms except for portfolios. All of the Renaissance teachers reported using oral presentations at least once or twice per semester during their math classes, whereas 26% of the multi-state teachers had never used them. Portfolio use was a more frequent classroom activity in the multi-state group than in the Renaissance. Sixteen percent of the multi-state teachers reported regular portfolio use (at least weekly) in their math classes compared with 0% in the Renaissance group.

Table 6

Percentage of Teachers Reporting on Classroom Activities

	Math Renaissance				Multi-state			
	Never	Once or twice per semester	Monthly	At least weekly	Never	Once or twice per semester	Monthly	At least weekly
Oral presentations	0	70	19	12	26	47	5	22
Portfolios	19	51	30	0	38	28	18	16
Student or teacher designed rubrics	0	19	60	21	34	33	26	7
Journals and logs	9	19	37	35	60	15	18	8
Open-ended questioning	0	0	19	81	2	21	31	47
Worksheets or workbooks	9	19	44	28	3	8	31	57
Student self-evaluations	0	37	58	5	18	46	18	18

Note. The question in the survey: How often have you used the following instructional strategies or activities with these students?

Although they used portfolios more frequently, multi-state teachers also tended toward heavy use of worksheets and workbooks. Thus, overall, the results showed that the Renaissance teachers used "reform-oriented" instructional strategies and classroom activities more frequently during their math instruction, whereas multi-state teachers were more likely to use "traditional" instructional strategies such as lecture, seat work, worksheets.

Comparison of Student Responses With Teacher Responses

There were a few overlapping questions between the sets of student (Table 4) and teacher questions (Tables 5 and 6), such as oral presentation, portfolio, group work (group assignments), student self-evaluations, and homework. Some discrepancies between student and teacher responses can be expected, because student responses would be based mostly on their personal experiences, whereas teacher responses would be based on their teaching plan for the whole class. Indeed, in both the Renaissance and multi-state groups, students were more likely to report working in small groups at least weekly than were their teachers. But a higher percentage of teachers than students reported using oral presentations at least weekly. A mixed pattern appeared for portfolio use, with a higher percentage of students than teachers reporting at least weekly use in both the Renaissance and the multi-state groups; however, a higher percentage of students than teachers also reported never using portfolio in the Renaissance. Despite these discrepancies in responses between students and teachers, the overall pattern is the same for both students and teachers: both students and teachers gave responses indicating that the Renaissance students were exposed more frequently to reform-oriented activities.

Opportunity for Teacher Learning

The need for appropriate learning opportunities is not limited to students. Teachers also need opportunities to learn both the content and the teaching strategies appropriate to the new, more demanding achievement standards embodied in the Reference Examinations. A comparison of the learning opportunities of the Renaissance teachers with those of the multi-state comparison group reveals some interesting results. Teachers were asked a series of questions concerning their participation in staff development. Table 7 summarizes their responses.

Table 7

Percentage of Teachers Reporting Various Categories of Staff Development Activities

	Math Renaissance				Multi-state			
	Not at all	Once	2-5 times	5+ times/ ongoing	Not at all	Once	2-5 times	5+ times/ ongoing
NCTM standards	0	0	31	69	10	17	43	30
Group work	0	0	21	79	4	4	46	47
Problem solving	5	0	14	81	7	14	29	50
Mathematics concepts	5	0	19	77	9	16	27	48
Use of technology in teaching	2	5	44	49	12	14	61	12
Performance assessment	5	2	49	44	3	7	53	37
Portfolio assessment	5	5	67	23	17	32	32	19

Note. The question in the survey: How often have you led or participated in the following staff development during the last 5 years?

The Renaissance teachers reported participating in various professional development activities focused on the NCTM Standards, group work, problem solving, mathematics concepts, use of technology in teaching, performance assessment, and even portfolio, much more frequently than the multi-state teachers. Over 90% of the Renaissance teachers responded that they had participated in these activities two to five times or more during the last five years, whereas many teachers in the multi-state group had participated only once or never during that time period. Most of these activities appear to be ongoing and regular for the Renaissance teachers. However, staff development on portfolio assessment was the least frequent ongoing activity for Renaissance teachers. In sum, the Renaissance teachers seemed to have had more opportunities to participate in various reform-oriented professional development activities.

Student Achievement

Standards-level comparison. Table 8 presents the distribution of overall student performance for the Renaissance and the multi-state groups in Skills,

Table 8

Standards Level Comparison

	Standards level	Skills	Concepts	Problem solving
Math Renaissance	Achieved the standard with honors	19	2	0
	Achieved the standard	33	18	12
	Nearly achieved the standard	25	26	32
	Below the standard	16	26	43
	Little evidence of achievement	6	28	13
Multi-state	Achieved the standard with honors	17	5	2
	Achieved the standard	25	15	8
	Nearly achieved the standard	22	19	25
	Below the standard	22	22	40
	Little evidence of achievement	14	39	25

Concepts, and Problem Solving. In both groups, more students received "Achieved the Standard with Honors" in Skills (19% in the Renaissance group; 17% in the multi-state group) than in Concepts (2% in-Renaissance; 5% in multi-state) and in Problem Solving (0% in Renaissance; 2% in multi-state group). At the other end of the scale, fewer received the lowest score, "Little Evidence of Achievement," in Skills (6% in Math Renaissance; 14% in multi-state) than in Concepts (28% in Renaissance; 39% in multi-state) and Problem Solving (13% in Renaissance; 25% in multi-state).

The differences in the percentages of students "Meeting the Standard" (i.e., the top two categories) between the two groups were 10% in Skills, 0% in Concepts, and 2% in Problem Solving. The differences in the percentages were examined using z-statistic. In this comparison, significantly higher percentages of the Renaissance students met or exceeded the Standard (i.e., the top two categories) in Skills ($p < .05$), but not in Problem Solving and in Concepts. In Skills, which is presumably taught in most classrooms, the Renaissance teachers not only pushed many more students to "Meeting the Standard" level (cumulative frequency of 52% in the Renaissance group; 42% in the multi-state group) but also more successfully pulled students out of the two bottom categories ("Below the Standard" and "Little Evidence of Achievement";

cumulative frequency of 22% in the Renaissance students vs. 36% in the multi-state students, significant at $p < 0.05$).

In Concepts, the percentages of students "Meeting the Standard" was the same for the two groups; however, the percentage of multi-state students who received "Little Evidence of Achievement" was 7% higher than that of the Renaissance group ($p < 0.05$). This indicates that the Renaissance teachers were able to pull students from the bottom two categories ("Below the Standards," and "Little Evidence of Achievement") to "Nearly Achieved the Standard," but were not able to help many of them to actually reach the standard.

For Problem Solving, the difference in the percentages of students "Meeting the Standard" was small and not significant. However, as for Concepts, the greatest success of the Renaissance teachers was in pulling students out of the bottom achievement categories. The percentage of multi-state students who received "Little Evidence of Achievement" in Problem Solving was almost double the percentage in the Renaissance group ($p < 0.05$).

Equity and Instructional Validity: Hierarchical Linear Modeling

The data reported above suggest that, overall, Renaissance teachers are generally more successful than their multi-state counterparts in producing good levels of examination performance among their students in Skills and in pulling students from the bottom two categories to "Nearly Achieved the Standard" in Concepts and Problem Solving. We also wanted to know whether Renaissance professional development experience helped teachers reduce the achievement gap between white and minority students and whether the differences between Renaissance and multi-state achievement were indeed due to the differences in modes of instruction of the two teacher groups. For this purpose we conducted a set of hierarchical linear modeling analyses**.**

Table 9 presents the means and standard deviations of the variables used in these analyses, for both the Renaissance and multi-stage samples. Table 10 shows the intercorrelations among these variables. As can be seen, Group was positively correlated with STAFFDEV ($r = 0.38$) and INSTACT (0.48), but was negatively correlated with TRADIT ($r = -0.48$). In other words, the Renaissance teachers participated in various staff development activities and used "reform-oriented" activities in their instruction more frequently than the multi-state

Table 9

Means and Standard Deviations of the Variables Used in the Study

Variable	Renaissance		Multi-state	
	Mean	*SD*	Mean	*SD*
Student-level variables				
Mathematics performance				
Skills	2.74	0.75	2.50	0.89
Concepts	2.03	0.78	1.93	0.91
Problem solving	2.00	0.62	1.84	0.68
SES	3.43	1.25	3.32	1.19
Minority (Percentage of minority students)	0.37	0.48	0.45	0.50
Teacher-level variables				
Group				
STAFFDEV	3.53	0.51	3.02	0.62
TRADIT	2.50	0.71	3.20	0.53
INSTACT	2.61	0.28	2.15	0.46

Table 10

Correlations of Teacher-level Variables and Outcome Variables

Variables	Group	STAFFDEV	TRADIT	INSTACT	Minority	Skills	Concepts	Problem solving
Group						0.15***	0.05*	0.12***
STAFFDEV	0.38***					0.21***	0.21***	0.19***
TRADIT	-0.48***	-0.23***				-0.10***	0.01	-0.06*
INSTACT	0.48***	0.54***	-0.50***			0.17***	0.15***	0.21***
Minority	-0.09***	-0.16***	-0.04	-0.16***		-0.34***	-0.39***	-0.32***
SES	0.08**	0.10***	0.08**	0.11***	-0.21***	0.19***	0.22***	0.21***

***significant at $p < 0.001$; **significant at $p < 0.01$; *significant at $p < 0.05$.

teachers. STAFFDEV and INSTACT were positively correlated (r = 0.54) while TRADIT was negatively correlated with STAFFDEV (r = -0.23) and INSTACT (r = -0.51). The relationship between STAFFDEV and INSTACT implies that the more teachers were aware of the reform-oriented goals and practices through

various staff development activities, the more likely they would implement the practices in their classrooms. Reflecting what we have reported earlier, participation in staff development and the use of "reform" methods of instruction were positively associated with achievement (ranged from 0.15 to 0.21). The use of traditional modes of instruction was negatively associated with achievement in Skills (r = -0.10) and Problem Solving (-0.06), but no association was found with achievement in Concepts. SES was positively associated with achievement while Minority status was negatively associated with it. There were also significant associations of student background with staff development and the two modes of teaching: SES was positively related to STAFFDEV, TRADIT, and INSTACT while Minority was negatively correlated with STAFFDEV and INSTACT, but not with TRADIT. This indicates that minority students received "reform" methods of instruction less frequently then white students and their teachers participated in staff development less frequently than white students' teachers.

Equity. Table 11 shows results of the HLM analysis that examined the effects of Group (whether the class was taught by a Renaissance or a multi-stage teacher), minority status of students and SES on student achievement. In this analysis, the adjusted mean achievement in each class was allowed to vary across classrooms (random effect) while the slopes of Minority and SES were not (fixed effect). The variance of mean achievement across classes (τ_{00}) was modeled with Group (γ_{01}) and an interaction effect (γ_{11}) between Minority and Group was also examined.

Class mean achievement (γ_{00}) is an adjusted mean achievement in the multi-state group. The Group effect (γ_{01}) indicates the estimated mean difference between the Renaissance and the multi-state adjusted for student background (0.23 in Skills; 0.08 in Concepts; 0.16 in Problem Solving). This difference was significant for Skills and Problem Solving, but not for Concepts. For Skills and Problem-Solving, in other words, class means were, on average, higher in Renaissance than in multi-state classrooms.

Minority (γ_{10}) indicates the gap between minority and white students' performance. On all three standards, minorities performed significantly less well than white students (-0.38 in Skills; -0.45 in Concepts; -0.29 in Problem Solving), with the largest gap for Concepts and the lowest for Problem Solving. Group (γ_{11}) is a cross-level interaction that estimates the difference in the minority gap

Table 11

Results of HLM Analyses: Group Effect

Variables	Skills		Concepts		Problem solving	
	Effect	(SE)	Effect	(SE)	Effect	(SE)
Class mean achievement, γ_{00}	2.57***	(0.04)	1.95***	(0.05)	1.88***	(0.03)
Group, γ_{01}	0.23**	(0.08)	0.08	(0.10)	0.16*	(0.07)
Minority, γ_{10}	-0.38***	(0.04)	-0.45***	(0.04)	-0.29***	(0.03)
Group, γ_{11}	0.21*	(0.08)	0.07	(0.07)	0.03	(0.06)
SES, γ_{20}	0.05**	(0.01)	0.04**	(0.01)	0.05***	(0.01)
Estimates of variances:						
Variance of class mean, τ_{00} (between-class variance)	0.126		0.230		0.098	
Within-class variance, σ^2	0.510		0.432		0.294	

Note. All variables were Grand-mean centered; ***significant at $p < 0.001$; **significant at $p < 0.01$; *significant at $p < 0.05$.

within Renaissance and multi-state groups. This interaction effect was significant for Skills, but not for Concepts or Problem Solving. The interaction effect in Skills indicated that the difference in student performance between white and minority students depended on group membership (i.e., whether the student was in the Renaissance or the multi-state sample). Minority students in the Renaissance group were somewhat less disfavored than those in the multi-state group. Overall, the greater minority gap in Skills in the multi-state group and the Group effect in Problem Solving suggest that the Renaissance program had an impact on student performance, particularly beneficial effect for classically underserved students (low SES and minorities).

Instructional validity. In the next analysis, teacher-level variables (STAFFDEV, TRADIT, and INSTACT) were added to the model in order to examine the effects of staff development and instructional strategies on class means (student performance). Among the teacher-level variables, only TRADIT and INSTACT were included in the model in Table 12 because STAFFDEV was not significant and the model was improved after deleting it in all three clusters. The results are reported in Table 12.

Table 12

Results of HLM Analyses: Effect of Classroom Instruction

Variables	Skills		Concepts		Problem solving	
	Effect	(SE)	Effect	(SE)	Effect	(SE)
Class mean achievement, γ_{00}	2.58***	(0.04)	1.95***	(0.05)	1.88***	(0.03)
Group, γ_{01}	0.15	(0.10)	0.01	(0.12)	0.07	(0.08)
TRADIT, γ_{02}	0.02	(0.07)	0.15	(0.08)	0.01	(0.06)
INSTACT, γ_{03}	0.21*	(0.10)	0.37**	(0.13)	0.22*	(0.09)
Minority, γ_{10}	-0.37***	(0.04)	-0.45***	(0.04)	-0.29***	(0.03)
Group, γ_{11}	0.21**	(0.08)	0.08	(0.07)	0.03	(0.06)
SES, γ_{20}	0.05**	(0.01)	0.04**	(0.01)	0.05***	(0.01)
Estimates of variances:						
Variance of class mean, τ_{00} (between-class variance)	0.123		0.214		0.093	
Within-class variance, σ^2	0.510		0.432		0.294	

Note. All variables were Grand-mean centered; ***significant at $p < 0.001$; **significant at $p < 0.01$; *significant at $p < 0.05$.

When teacher-level variables were added in the model, Group (γ_{01}) effect disappeared in Skills and Problem Solving; however, there was no change in the effect of Minority (γ_{10}). The lack of a positive Group (γ_{01}) effect means that once teacher-level variables are taken into account, the difference between Renaissance and multi-state achievement levels disappeared. In other words, the overall higher achievement of Renaissance students on Skills and Problem Solving seen in Table 11 is due to the specific kinds of teaching they experienced. More specifically, there was a positive effect of reform-oriented instructional strategies and activities (INSTACT, γ_{03}) on class means for all three standards after controlling for student background characteristics. The reported use of traditional teaching strategies (TRADIT, γ_{02}), on the other hand, did not affect achievement.

Discussion

This study examined the relationship between professional development opportunities for teachers, the kinds of instruction offered to students, and student performance on a mathematics examination the New Standards Mathematics Reference Examination designed explicitly to function as a tool for reforming instruction and making high level instruction more equitably available to different groups of students. By comparing teachers (and their students) who had participated in the California Mathematics Renaissance professional development program with teachers and students elsewhere we were able to evaluate both the effectiveness of the Renaissance program and the instructional validity of the Reference Examination.

Our results showed that Renaissance teachers had more opportunities to participate in reform-oriented professional development activities than teachers in a multi-state comparison group and that Renaissance classroom practice reflected these teacher learning opportunities. As reported by both teachers and their students, Renaissance teachers were more likely to engage their students in problem-solving, explanation of problem solutions, small group work and other activities associated with the kinds of challenging content advocated by the NCTM *Standards* and embodied the New Standards *Performance Standards*. Thus, the enhanced teacher learning opportunities offered by the Renaissance appeared to have the intended effect on student opportunity to learn.

The differential opportunity to learn for students was, in turn, reflected in student performance on the New Standards exam. Renaissance students showed significantly higher levels of performance on Skills and Problem Solving clusters of the New Standards examination compared to students in the multi-state group, after adjusting for student background characteristics. The "reform-oriented" instruction also showed a positive effect on student performance after controlling for student background characteristics. The New Standards exam composed of constructed response tasks, some of which were quite extended in length and required complex problem solutions and explanations was well matched to the kind of instructional goals and methods espoused by the Renaissance. The generally higher performance of Renaissance than comparison group students and the effect of instruction thus confirms the instructional validity of the New Standards Reference Examination. The examination is

sensitive to broad differences in instruction and is able to register the effects of particular kinds of instructional strategy and content.

An instructionally valid test is particularly useful in examining the equity effects of educational programs. A fundamental strategy for improving minority and low SES academic performance is to end the practice of providing *de facto* different, lower demand curricula to poor and minority students than to those who are more privileged. This strategy, while it will not alone eliminate achievement differences between social groups, should at least narrow the gap. Assessments used to evaluate such equity-oriented education efforts need to be able to register the effects of the different instruction students receive. The New Standards Reference Examination was able to do this for the Renaissance program, at least in Skills. The achievement results in Skills cluster showed a smaller performance gap between white and minority students in the Renaissance group than the multi-state group. The Renaissance program was apparently effective in bringing more demanding instruction and learning opportunities to minority students and this, in turn, was helping to reduce the minority-white performance gap. A reduction in the performance gap, while controlling for level of SES, suggests that the Renaissance program was having a generally positive effect on equity of educational opportunity.

Instructional validity is a necessary, but not sufficient, attribute of the consequential validity of an assessment aimed at influencing the character of instruction and its equitable distribution. Establishing the consequential validity of an assessment would require also showing what *causal* effects the assessment has on educational practice of its users or on student opportunities. Can the assessment stimulate new forms of instruction? Can it improve the quality of professional development? Can it be used as a positive part of an equity program actively encouraging teachers to provide the same demanding curriculum for all students? On the negative side, does it unduly narrow the focus of instruction? Does its use create barriers for some students that the educational program as a whole is not able to overcome?

The positive effects of the Renaissance program on overall achievement levels and its reduction of minority and SES achievement gaps cannot, of course, be causally attributed to the introduction of the New Standards exams. Renaissance had been working toward the NCTM *Standards* for several years before the New Standards exams became available. Instead, the directors of the

Renaissance program selected the New Standards Reference Examination because they needed an assessment that would be sensitive to the instructional goals they were already promoting and at the same time provide comparative information for a larger, national sample of students. They were, in other words, "betting" on the instructional validity of the New Standards exams with respect to the learning opportunities for teachers and students that they were already providing.

In other jurisdictions, the New Standards exams are often chosen as a tool for initiating instructional change. In these cases, jurisdictions provide professional development aimed at informing teachers about the nature of the performance standards to which the exams are referenced and about curriculum and instruction that will prepare students to meet the standards. These districts, in other words, encourage active "teaching to the standards." Since examinations are systematically referenced to the standards, such teaching can be expected to yield measured achievement gains without the narrowing effect of rehearsing students on specific test items. Evaluation of this claim would require examining both New Standards exam results and results on other tests within a jurisdiction that has adopted the New Standards assessments in order to implement a "teaching to the standards" strategy. We will be reporting on studies of such districts in later papers in this series.

References

Airasian, P., & Madaus, G. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103-117.

Anderson, L. W. (1990). *Opportunity to learn and the National Assessment of Educational Progress: An analysis with recommendation*. Unpublished manuscript.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioral research: applications and data analysis procedures*. Beverly Hills, CA: Sage.

Derghazarian, E. (1996). *Classroom mathematics study: A preliminary comparative analysis of TIMSS and mathematics renaissance teachers at the eighth-grade level*. San Francisco, CA: WestEd.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*, 5-30.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*, 119-132.

Howard, J. (1995). You can't get there from here: The need for a new logic in education reform. *Daedalus*, *124*(4), 85-92.

Glaser, R. (1990). *Testing and assessment: O tempora! O mores!* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*, 323-336.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught. *Journal of Educational Measurement, 18*, 85-96.

Loucks-Horsley, S., Stiles, K., & Hewson, P. (1996). *Principles of effective professional development for mathematics and science education: A synthesis of standards*. NISE Brief, University of Wisconsin-Madison.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*, 185-196.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

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

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

New Standards. (1994). *The 1994 mathematics reference examination: Design, administration, results, and accuracy assessment.* Oakland, CA: New Standards.

New Standards. (1996). *1996 New Standards reference examination technical summary.* Oakland, CA: New Standards.

New Standards. (1997). *Performance standards (Vols. 1, 2, 3).* Rochester, NY: National Center on Education and the Economy.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13, 24.

Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21-27.

Raudenbush, S. W., & Willms, J. D. (1991). The organization of schooling and its methodological implications. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schools from a multilevel perspective.* London: Academic Press.

Resnick, R. B., & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Resnick, L. B. (1995). From aptitude to effort: A new foundation for our schools. *Daedalus*, *124*(4), 55-62.

Resnick, L. B., & Nolan, K. J. (1995). Standards for education. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* Washington, DC: Brookings Institution.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership, 52*(5), 38-43.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

Shields, P. M., Marder, C., & Wilson, C. L. (1996). *A case study of the California alliance for mathematics and science (CAMS), 1992-1993*. Menlo Park, CA: SRI International.

Stevens, F. I. (1993). *Opportunity to learn: Issues of equity for poor and minority students.* Washington, DC: National Center for Education Statistics.

WestEd. (1996). *Examination institutionalization of mathematics renaissance and CSIN in California schools: CAMS 1995-1996 Evaluation report case study component.* Oakland, CA: Author.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*, 355-370.

Yoon, B. (1996). *1994 California Renaissance study*. Mathematics Renaissance, Ventura, CA.

Yoon, B. (1997). *1996 Middle grades Mathematics Renaissance study*. Oakland, CA: New Standards.

Young, M. J., & Yoon, B. (1997). *Estimating consistency and accuracy of classifications in* standards-*referenced assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.