The 1994 Mathematics Reference Examination: Design, Administration, Results, and Accuracy Assessment

Technical Report 438

New Standards

LRDC, University of Pittsburgh

December 1996

National Center for Research on Evaluation,

Standards, and Student Testing (CRESST)

Graduate School of Education & Information Studies

University of California, Los Angeles

Los Angeles, CA   90024-1522

(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA Catalog number 84.117G 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 policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The design. development, administration, scoring, and analysis of the 1994 New Standards Mathematics Reference Examination would not have been possible without the work of a number of individuals. Like the Reference Examination, this report represents the collaborative effort of the New Standards staff.

The individual sections of this report were written by David Wiley and Michael James Young, with additional material by Ann Borthwick, Phil Daro, and Lauren Resnick. Data analyses programming in support of the results reported here were provided by Winnie Young. Editing and coordination of the report was provided by Seth Beckerman.

Drafts of the report were reviewed by staff now or formally with New Standards. They include Jeff Adkins, Harold Asturias, Pam Beck, Mishaa Degraw, Sally Hampton, Kate Nolan, Christine Ross, Ann Shannon, Elizabeth Stage, John Tanner, Cheryl Tibbals, and Bokhee Yoon.

1. Introduction and Overview

This report presents results of the 1994 Reference Examinations in Mathematics. The examinations were taken by a nationwide sample of 30,000 fourth and eighth grade students.

The examinations were administered in a "matrixed" form, that is, each student took a selection of the examination tasks and no one student completed the full set of tasks included in the examination. This approach reduced the length of time each student spent taking the examination. The matrixed sample was designed to allow use of the collective results to estimate the proportion of students who would have performed at each level of achievement if these students taken the full examination. This form of examination administration means that the results must be interpreted as results for the partnership of states and school districts as a whole, rather than as results that can be interpreted for any specific state or school district, individual school, or individual student. In future years the examinations will be designed to provide such individual scores.

This report provides information about how students performed on examinations designed to find out what the students know and what they can do in relation to specified standards of performance at each grade level. This contrasts with traditional kinds of examinations that assess student performance in relation to their peers and previous test takers rather than reveal information about what they actually know and can do. This is the first report of a standards-based examination given at a national level.

Information about what students actually know and are able to do is essential to the efforts of schools that are setting clear, high standards of performance for their students and working hard to help their students achieve those standards. This report is an important milestone along the road to providing information that schools can use to monitor student achievement and develop strategies for improving performance. Analyses of the results referenced to specific clusters of standards that were assessed are included, as are analyses of the results broken out by gender, race, and ethnicity. In future years the examination will be designed to produce information that can be analyzed for individual schools and for individual students.

The remainder of this section introduces the New Standards Mathematics performance standards, the format of the 1994 Reference Examination, and its administration. Section Two describes the scoring of the Reference Examination tasks. This section includes a general description of the scoring rubric, the scoring procedures used and the accuracy of the scorers, and the task-level scoring results for the Reference Examination. Section Three describes the methodology used to combine these task-level scores and produce examination scores that are referenced to the Mathematics performance standards. Overall results, as well as results disaggregated by gender, ethnicity, and English language proficiency are presented in this section. Section Four of this report describes the methodology used to assess the accuracy of the Reference Examination results.

What was learned from the 1994 Reference Examination helped to revise the test design for subsequent years. Therefore, this report contains that pertains only to the 1994 Examination.

1.1 Mathematics Performance Standards

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 first four Mathematics performance standards delineate the important *conceptual content* that students need to learn:

Standard 1, Arithmetic and Number Concepts (elementary school)
Number and Operation Concepts (middle and high school)

Standard 2, Geometry and Measurement Concepts

Standard 3, Function and Algebra Concepts

Standard 4, Statistics and Probability Concepts

The next three standards delineate the important *procedures and strategies* that students must develop and use in the four conceptual areas listed above:

Standard 5, Problem Solving and Mathematical Reasoning

Standard 6, Mathematical Skills and Tools

Standard 7, Mathematical Communication.

The final standard calls for the application of conceptual understanding and procedures and strategies in extended mathematical projects:

Standard 8, Putting Mathematics to Work.

The tasks in the 1994 Mathematics Reference Examination are aligned explicitly with the Mathematics performance standards. This alignment ensures that the examination is appropriately referenced to the standards. The scoring process (Section 4) is also based directly upon these standards.

The tasks are aligned with Standards 1-7. Standard 8, Putting Mathematics to Work is not assessed in the Mathematics Reference Examination. It is assessed exclusively through the portfolio system.

For scoring purposes, Standards 1-7 are clustered into three categories:

A. Concepts (number and operations, geometry and measurement, functions and algebra, statistics and probability)

B. Mathematical Skills and Tools

C. Problem Solving and Mathematical Communication.

Specifying expectations for Concepts (category A) required identifying what must be understood and how well understanding must be reflected in performance. Tasks categorized as using these concepts were those which demanded a type of performance that demonstrated this quality of understanding. Categories of tasks were established similarly for Skills and Tools (category B) and for Problem Solving and Mathematical Communication (category C)

Scores for the tasks identified within category A were aggregated to produce a single “Concepts” score. Scores were also defined for Skills and Tools (category B) and Problem Solving and Mathematical Communication (category C).

Teachers and examiners familiar with the performance standards and with the reference examination at each grade level evaluated performance on the reference examination tasks. They also made judgments about the aggregate scores and patterns of performance across tasks required to meet or exceed the requirements of the performance standards. These judgments were used to calculate the proportions of students performing above standard, at standard, and below the standard for each of three categories of standards. The results are summarized in Section 2.

1.2 Examination Format and Administration

The mathematics examinations were constructed in two booklet formats — multi-task and single task. Both examination formats lasted 45 minutes.

Multi-Task Format. Multi-task booklets contained five to seven short tasks. Each booklet sampled important grade-level mathematics, but no single booklet sampled the entire range required to describe student performance across all the standards. These booklets contained both 15-minute and 5-minute tasks. One type of the multi-task booklet contained two 15-minute and three 5-minute tasks. The another type contained one 15-minute task and six 5-minute tasks.

The 15-minute tasks asked students to use concepts, skills and tools, and mathematical reasoning to solve a problem. Students were usually asked to explain and show how they got their answers. A multi-task booklet contained one or two 15-minute tasks. The 5-minute tasks assessed important concepts and had lighter problem-solving loads than the 15-minute tasks.

Single-Task Format. Each single task also took 45 minutes and appeared in its own student booklet. These tasks asked students to choose and implement a workable mathematical approach for which the purpose and product were clear but for which the strategies, mathematical concepts, and skills and tools were left for the student to decide. The names of the tasks and the forms on which they were administered are shown in Table 1.1.

Every examinee was given two booklets, one with multiple tasks and one with a single task, for a total of 90 minutes of testing. There were four multi-task booklets for each grade, which were distributed among the students in each class. About one-quarter of the students in each class were given the same booklet. Single-task booklets were distributed across schools so that all students in a single school received the same booklet. Table 1.2 presents, by partner and form, the number of elementary students who took the tests. Partners are either state education systems or large school districts. Table 1.3 presents similar data for middle school students. Note that these data are for students who completed both the multi-task and single-task booklets.

1.3 Examinee Demographics

Along with the examination booklets, each student was asked to complete a Student Information Form (SIF). categories The demographic breakdown of the students who were administered examination is shown in Table 1.4.

Approximately equal percentages of male and female students were administered the Reference Examination in both elementary and middle schools. White (not Hispanic) students were the majority of students administered the examinations (approximately 70% of students in elementary and middle school); African-American students (15.3% elementary school, 11.7% middle school) and Hispanic/Latino students (9.5% elementary school, 9.6% middle school) comprised the next largest ethnic groups. The majority of students who took the examinations were English Proficient. Only two percent of the students were Limited English Proficient in both elementary and middle schools (2.4% in elementary school; 2.0% in middle school).

In parental education, at least 30% of the parents of elementary school students, and 43% of the parents of middle school students, had graduated from college or had an advanced degree. Note that a large proportion of responses on parental education was “Unknown” (36.5% in elementary school; 18.0% in middle school). The large percentage in elementary school may be because teachers responded for their students while middle school students responded to the question themselves.

Note that the total numbers of students completing the SIFs in elementary and middle school (Table 1.4) were less than the numbers of students who completed both booklets of the examination in elementary and middle school (Tables 1.2 and 1.3). The percents of students who completed both examination booklets given that they completed the SIFs were 95.3% and 88.9% for elementary and middle schools respectively.

Table 1.1 Forms and task names used in the 1994 Mathematics Reference Examination.

Elementary		Middle School	
Form	**Task Names**	**Form**	**Task Names**
	Multi-Task Forms		
E 1	Sharing a Cake	**M 1**	Cubes
	Fair Game?		Odd or Even
	Recycling		Life Expectancy (1 of 3)
	Marble Estimation		Life Expectancy (2 of 3)
	Rule It Out		Life Expectancy (3 of 3)
E 2	Helping Martin	**M 2**	Graphing
	Willie's Patterns		Comparing Triangles (1 of 3)
	Fourths?		Comparing Triangles (2 of 3)
	Coins (1 of 3)		Comparing Triangles (3 of 3)
	Coins (2 of 3)		Comparing Drawings
	Coins (3 of 3)		Thermometer
	Secret Number		Bake Sale
E 3	Counting Raisins Part 1(1 of 2)	**M 3**	Chocolate Chip Cookies (1 of 2)
	Counting Raisins Part 1(2 of 2)		Chocolate Chip Cookies (2 of 2)
	Counting Raisins Part 2		Changing Percents
	Jana's Number		Area
	It's a Half		Copy Shop
	Building With Squares		
	Jessie's Marbles		
E 4	Symmetry in Shapes	**M 4**	Moe's Patterns (1 of 3)
	Favorite Fruit Part 1		Moe's Patterns (2 of 3)
	Favorite Fruit Part 2		Moe's Patterns (3 of 3)
	Carpet Job		Eating Pizza
	Marvelous Marbles		The Better Airline Deal
			Science Fair
			How Cold Is It?
	Single Task Forms		
E 5	Order/Reorder	**M 5**	Carnival Game
E 6	Partitions	**M 6**	Design a Box
E 7	School Supplies	**M 7**	Sports Bag
E 8	A Trip to the State Park	**M 8**	Truth in Advertising

Table 1.2 Number of elementary school students taking each form of the 1994 Mathematics Reference Examination (by partner).

	Multi-Task Forms					Single-Task Forms				
Partner	**E1**	**E2**	**E3**	**E4**	**Total[1]**	**E5**	**E6**	**E7**	**E8**	**Total[1]**
California	101	93	101	97	392	217	113	62	0	392
Colorado	214	217	202	211	844	209	111	182	342	844
Delaware	65	61	58	63	247	0	129	118	0	247
Iowa	499	495	485	482	1,961	400	500	398	663	1,961
Kentucky	49	49	50	50	198	63	135	0	0	198
Maine	137	143	142	141	563	51	87	319	106	563
Massachusetts	216	222	221	213	872	193	215	151	313	872
New York	457	462	442	439	1,800	386	360	628	426	1,800
New York City, NY	84	81	84	88	337	0	181	156	0	337
Oregon	50	46	43	48	187	0	187	0	0	187
Pennsylvania	376	380	374	375	1,505	356	339	415	395	1,505
Pittsburgh, PA	181	173	179	170	703	151	136	173	243	703
Rochester, NY	14	14	16	14	58	58	0	0	0	58
San Diego, CA	32	25	29	31	117	0	0	117	0	117
Vermont	412	407	402	398	1,619	481	434	454	250	1,619
Washington	582	571	577	568	2,298	635	446	399	818	2,298
White Plains, NY	89	87	85	86	347	0	0	0	347	347
National Alliance	340	339	341	333	1,353	442	392	289	230	1,353
Total	3,898	3,865	3,831	3,807	15,401	3,642	3,765	3,861	4,133	15,401

[1] All students took one multi-task form and one single-task form; therefore the totals are the same.

Table 1.3 Number of middle school students taking each form of the 1994 Mathematics Reference Examination (by partner).

	Multi-Task Forms					Single-Task Forms				
Partner	**M1**	**M2**	**M3**	**M4**	**Total[1]**	**M5**	**M6**	**M7**	**M8**	**Total[1]**
California	157	157	164	152	630	150	280	200	0	630
Colorado	255	253	244	252	1,004	298	212	214	280	1,004
Delaware	21	20	21	20	82	0	82	0	0	82
Iowa	647	636	639	643	2,565	497	764	759	545	2,565
Kentucky	41	41	40	42	164	0	77	87	0	164
Maine	94	96	90	97	377	13	73	128	163	377
Massachusetts	300	297	304	301	1,202	268	406	336	192	1,202
New York	388	364	396	404	1,552	344	315	484	409	1,552
New York City, NY	55	62	59	61	237	80	0	0	157	237
Oregon	205	201	204	200	810	196	30	382	202	810
Pennsylvania	397	419	414	408	1,638	348	377	549	364	1,638
Pittsburgh, PA	154	158	144	162	618	178	202	125	113	618
Rochester, NY	22	18	19	19	78	25	53	0	0	78
San Diego, CA	7	6	6	6	25	25	0	0	0	25
Vermont	462	458	473	441	1,834	323	343	457	711	1,834
Washington	754	753	738	740	2,985	709	735	773	768	2,985
White Plains, NY	0	0	0	0	0	0	0	0	0	0
National Alliance	516	524	519	528	2,087	760	322	230	775	2,087
Total	4,475	4,463	4,474	4,476	17,888	4,214	4,271	4,724	4,679	17,888

1. All students took one multi-task form and one single-task form; therefore the totals are the same.

Table 1.4 Demographic summary of students administered the 1994 Mathematics Reference Examination.

	Elementary School		Middle School	
	Number	Percent[1]	Number	Percent
Gender				
Male	8,088	50.6	9,621	49.4
Female	7,892	49.4	9,836	50.5
Missing[2]	182		650	
Ethnicity				
African-American	2,172	15.3	2,203	11.7
Native American	184	1.3	378	2.0
Asian	367	2.6	601	3.2
Filipino	88	0.6	238	1.3
Hispanic/Latino	1,349	9.5	1,804	9.6
Pacific Islander	70	0.5	91	0.5
White (not Hispanic)	9,909	69.9	131,376	71.3
Multiple Marks	29	0.2	61	0.3
Missing	1,994		1,361	
English Proficiency Status				
English Proficient	13,935	97.6	16,073	98.0
Limited English Proficiency	342	2.4	313	2.0
Missing	1,885		3,717	
Parental Education				
Not a HS graduate	569	4.1	733	4.0
HS graduate	2,429	17.6	3,274	17.8
Some college	1,605	11.7	3,116	16.9
College degree	2,882	20.9	4,792	26.0
Advanced degree	1,255	9.1	3,189	17.3
Unknown	5,027	36.5	3,309	18.0
Missing	2,395		1,700	
Total Tested	**16,162**		**20,113**	

[1] Percents do not include missing.

[2] Missing denotes students who did not respond to a specific item on the Student Information Form (SIF).

2. Scoring the Tasks of the 1994 Mathematics Reference Examination

2.1 Criteria for Scoring Tasks

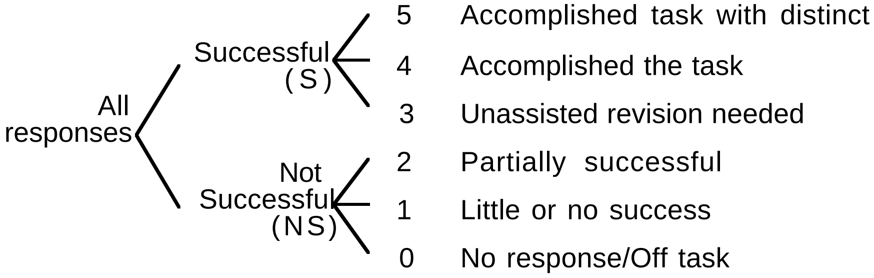
The discipline of mathematics values careful checking and rechecking, reviews of drafts by colleagues, and many revisions followed by careful editing. These refinements are not possible within the framework of on-demand testing. New Standards avoids undermining these values, however, by scoring assessments not as finished pieces, but as mathematical work generated in an on-demand setting.

The criteria for scoring are called rubrics. The New Standards rubrics were designed for use in a professional collaborative setting where teachers discuss and score together. This type of scoring has proved to be a powerful tool when used by teachers and students in self assessment. The New Standards rubrics focus on the performance rather than the performer. Scorers were directed by the rubric to the evidence in the response. To help make distinctions, scorers were asked to consider what feedback to the student would be appropriate based on the evidence in the response. While this assessment will not actually send feedback to individual students, the scoring of these on-demand tasks based, in part, on the feedback idea, should prove helpful to teachers who seek useful and accurate scores based on sound classroom practice.

Using the rubrics and previously scored examples called anchors, scorers judged each response. The first decision was whether the student had in fact accomplished the task successfully in terms of mathematics. Those that were successful were further subdivided into those that accomplished the task without need of revision, and those for which an unassisted revision was necessary and sufficient. Among those who had already accomplished the task, some were nominated for distinction.

Responses that were not successful were also further subdivided. Those that were partially successful were distinguished from those that engaged the task with little success and those that failed to engage the task (no response or off-task response).

This two-stage approach is illustrated by the following decision tree which summarizes the application of the rubric:



The decision tree was used to produce two kinds of scores. The 5-minute tasks and some of the 15-minute tasks in the multi-task booklets were scored using the first stage of the decision tree. Each 5-minute task and some parts of the 15-minute tasks were dichotomously scored as either successful (S) or unsuccessful (NS). The remaining 15-minute tasks in the multi-task booklets and the tasks in the single-task booklets were scored using the full decision tree on a 0 through 5 scale.

2.2 Scoring Procedures and Accuracy

Scores were assigned using the rubrics, which are guidelines for assigning scores to task-specific performances. The scoring process was guided by benchmark papers (anchors) selected by subject matter experts to exemplify particular scores in the rubric. Scorers were trained to use the rubric and benchmarks according to a standardized training procedure developed by New Standards (Figure 2.1).

The actual scoring operation was performed under a contract with the Psychological Corporation. Chief scorers and table leaders participated in the New Standards benchmarking meeting where the scoring references were prepared. The management of papers was supervised by the contractor.

During training and scoring, the scorers' understanding of the rubrics was calibrated through reading and discussing papers. Scorers ensured their ability to score to the rubric via calibration rounds before the actual scoring began. In assessing student performance, particularly in such a massive undertaking as the Mathematics Reference Examination, ensuring the accuracy of scores is critical. Therefore, during scoring, selected papers were scored twice to control the process.

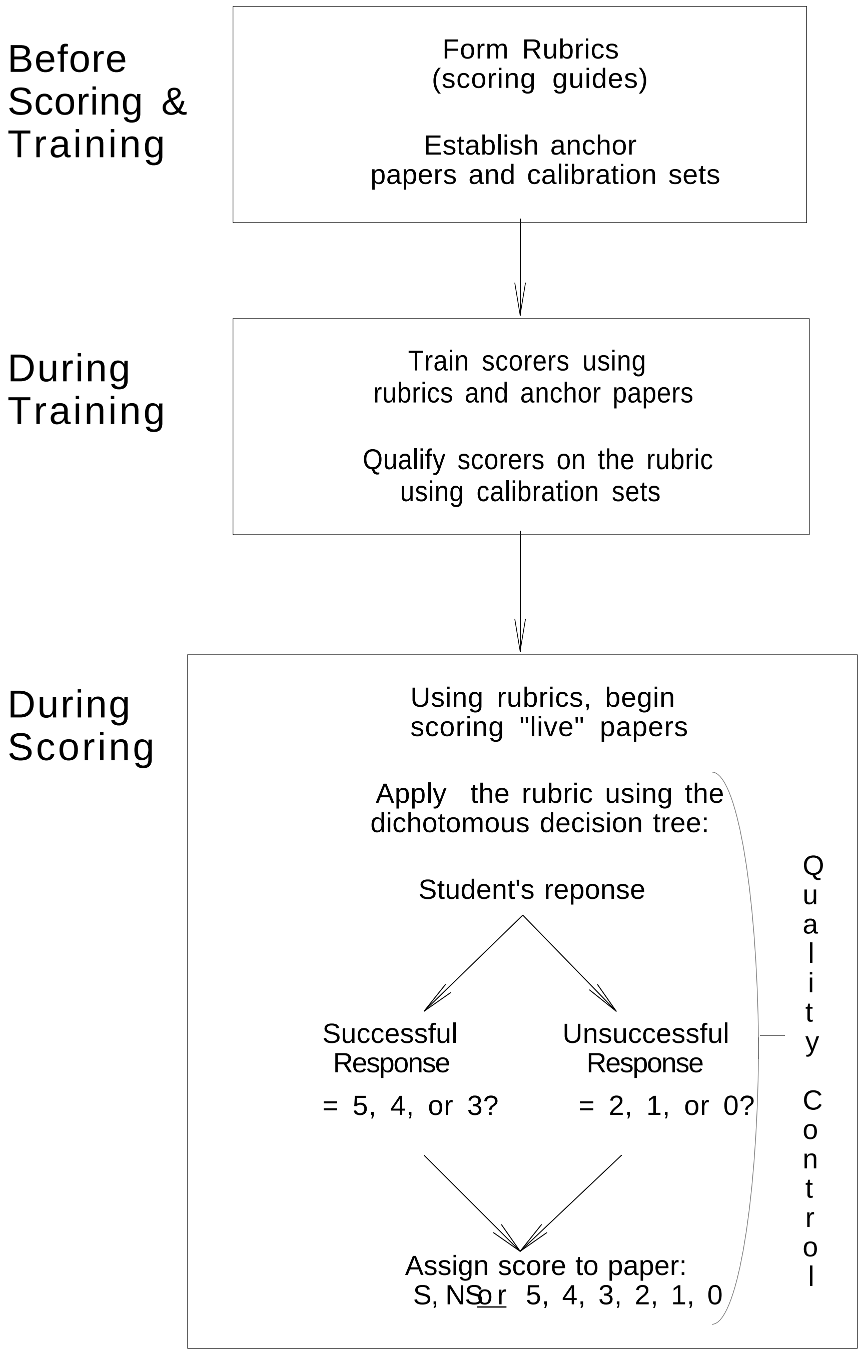

Before Scoring & Training
Form Rubrics (scoring guides)
Establish anchor papers and calibration sets
During Training
Train scorers using rubrics and anchor papers
Qualify scorers on the rubric using calibration sets
During Scoring
Using rubrics, begin scoring "live" papers
Apply the rubric using the dichotomous decision tree:
Student's reponse
Successful Response
= 5, 4, or 3?
Unsuccessful Response
= 2, 1, or 0?
Assign score to paper:
S, NS or 5, 4, 3, 2, 1, 0
Quality Control

Figure 2.1 Outline of the task scoring process.

Table 2.1 shows median percent perfect agreements among alternate scores of the same responses to tasks, and Table 2.2 shows the interquartile ranges of these percents within task types.

Many assessments define percent agreement between scores as the percentage of scores falling within plus-or-minus one point of each other. To compare the 1994 Mathematics Reference Examination to other assessments that use this measure of agreement, the percent of agreement within one score point has been calculated for the 0 – 5 point tasks (Table 2.3).

Generally, the longer the task, the less consistent the scoring — both in terms of scoring agreement and variation over tasks. However, typical agreement percents are quite high indicating that scoring errors do not contribute greatly to total measurement error.

2.3 Task Scoring Results

The result of the task scoring was rubric scores for each task, yielding a percentage distribution over score points. These percentage distributions for tasks scored using the first stage of the decision tree and for those tasks scored using the full scoring rubric are shown in Tables 2.4 through 2.7 for elementary and middle school. Note that rubric scores of five were rare. Scores of four or above generally constituted less than 20 percent of the responses for each task.

Table 2.1 Percent perfect agreement between first and second ("read behind") scores, by grade level and task type.

	Grade Level	
Task Type	**Elementary**	**Middle**
Dichotomously scored[1]; Multi-task booklet	94.5%	95.0%
0 to 5 scoring; Multi-task booklet	80.1%	91.3%
0 to 5 scoring; Single-task booklet	77.4%	92.2%

[1] Successful/Not Successful

Table 2.2 Interquartile ranges of percent perfect agreement between first and second ("read behind") scores, by grade level and task type.

	Grade Level	
Task Type	**Elementary**	**Middle**
Dichotomously scored[1]; Multi-task booklet	4.5%	2.8%
0 to 5 scoring; Multi-task booklet	2.1%	2.4%
0 to 5 scoring; Single-task booklet	6.3%	8.6%

[1] Successful/Not Successful

Table 2.3 Percent close agreement between typical scores and "read behind" scores, by grade level and task type.

	Grade Level	
Task Type	**Elementary**	**Middle**
0 to 5 scoring; Multi-task booklet	98.7%	99.7%
0 to 5 scoring; Single-task booklet	99.3%	99.9%

Table 2.4 Partner-wide summary of elementary school tasks scored using first part of decision tree.

Task Name	Form	N[1]	Percent at Rubric Score	
			S[2]	NS[3]
Building with Squares	E3	3,951	32.3	67.7
Coins (1)	E2	3,990	49.5	50.5
Coins (2)	E2	3,990	53.9	46.1
Coins (3)	E2	3,990	62.1	37.9
Counting Raisins (1:1)	E3	3,951	37.2	62.8
Counting Raisins (1:2)	E3	3,951	24.7	75.3
Counting Raisins (2)	E3	3,951	36.4	63.6
Favorite Fruit (1)	E4	3,946	57.3	42.7
Favorite Fruit (2)	E4	3,946	52.2	47.8
Helping Martin	E2	3,990	64.2	35.8
Jana's Numbers	E3	3,951	25.2	74.8
Jessie's Marbles	E3	3,951	14.1	85.9
Marble Estimation	E1	4,004	44.2	55.8
Marvelous Marbles	E4	3,946	47.9	52.1
Rule It Out	E1	4,004	43.5	56.5
Secret Number	E2	3,990	35.2	64.8
Sharing a Cake	E1	4,004	40.1	59.9
Symmetry in Shapes	E4	3,946	89.1	10.9
Willie's Patterns	E2	3,990	57.1	42.9

1. Number of students taking task.
2. Percent successfully completed the task.
3. Percent who did not successfully complete the task.

Table 2.5 Partner-wide summary of elementary school tasks scored using full decision tree.

Task Name	Form	N[1]	Percent at Rubric Score						Mean
			5	4	3	2	1	0	
Carpet Job	E4	3,946	0.1	25.8	9.6	33.5	25.2	5.8	2.2
Fair Game?	E1	4,004	0.0	8.3	21.4	11.4	52.4	6.5	1.7
Fourths?	E2	3,990	0.5	6.0	4.7	11.5	72.0	5.2	1.4
It's a Half	E3	3,951	0.1	27.7	17.3	12.7	38.9	3.4	2.3
Recycling	E1	4,004	0.0	19.2	18.7	20.8	38.0	3.2	2.1
Order/Reorder	E5	3,939	0.0	8.0	11.5	20.4	57.0	3.0	1.6
Partitions	E6	4,206	0.0	0.7	6.1	60.2	32.3	0.6	1.7
School Supplies	E7	3,811	0.0	3.2	8.8	32.2	54.1	1.7	1.6
A Trip to the State Park	E8	3,712	0.1	2.8	9.1	25.5	57.8	4.7	1.5

1. Number of students taking task.

Table 2.6 Partner-wide summary of middle school tasks scored using first part of decision tree.

			Percent at Rubric Score	
Task Name	Form	N[1]	S[2]	NS[3]
Bake Sale	M2	4,869	32.6	67.4
Better Airline Deal	M4	4,864	33.3	66.7
Changing Percents	M3	4,866	36.6	63.4
Chocolate Chip Cookies (1)	M3	4,866	49.8	50.2
Chocolate Chip Cookies (2)	M3	4,866	25.6	74.4
Comparing Drawings	M2	4,869	8.9	91.1
Comparing Triangles (1)	M2	4,869	35.0	65.0
Comparing Triangles (2)	M2	4,869	12.7	87.3
Comparing Triangles (3)	M2	4,869	12.0	88.0
Eating Pizza	M4	4,864	35.8	64.2
How Cold Is It?	M4	4,864	24.8	75.2
Life Expectancy (1)	M1	4,875	21.4	78.6
Life Expectancy (2)	M1	4,875	48.7	51.3
Life Expectancy (3)	M1	4,875	36.3	63.7
Moe's Patterns (1)	M4	4,864	16.7	83.3
Moe's Patterns (2)	M4	4,864	26.0	74.0
Moe's Patterns (3)	M4	4,864	9.8	90.2
Thermometer	M2	4,869	8.9	91.1

[1] Number of students taking task.
[2] Percent successfully completed the task.
[3] Percent who did not successfully complete the task.

Table 2.7 Partner-wide summary of middle school tasks scored using full decision tree.

			Percent at Rubric Score						
Task Name	Form	N[1]	5	4	3	2	1	0	Mean
Area	M3	4,866	0.0	6.9	14.9	12.4	52.2	13.6	1.5
Copy Shop	M3	4,866	0.0	1.6	4.7	38.9	39.4	15.3	1.4
Cubes	M1	4,875	0.0	3.8	4.5	31.2	53.5	7.1	1.4
Graphing	M2	4,869	0.0	1.0	19.9	38.1	24.8	16.1	1.6
Odd or Even	M1	4,875	0.0	14.2	11.7	9.6	58.1	6.2	1.7
Science Fair	M4	4,864	0.0	3.5	5.2	20.6	57.8	12.8	1.3
Carnival Game	M5	4,378	0.0	0.6	3.9	19.9	72.6	3.0	1.3
Design a Box	M6	4,897	0.0	7.3	9.2	29.2	50.4	3.9	1.7
Sports Bag	M7	4,904	0.0	0.4	11.4	21.6	60.9	5.7	1.4
Truth in Advertising	M8	4,346	0.0	0.0	5.6	30.7	48.5	15.1	1.3

[1] Number of students taking task.

3. Combining Task Scores to Examination Scores

3.1 Performance Standards and the Examination Content

Each of the New Standards grade-level performance standards in mathematics is defined largely by descriptions of what the student does, supported by illustrations of student work which is “at standard” (New Standards, 1995). It was important to tie each examination task to the standards it reflected because the examination was referenced to standards. In order to map the examination tasks to these standards, a group of mathematics educators (Section 2.1) was charged to use expert judgment, based on those standards and the tasks themselves (including task rubrics, mathematical commentary, and student responses).

The mapping was done by assigning a weighted value to each cell in a two-dimensional grid representing tasks by standards. Weight values reflect the strength of the relationship between task and standard. The weights were absolute rather than relative indices of strength for the total task. Weights of 1, 2, or 3 were assigned to each cell. A weight of 1 indicates that the task draws weakly upon the standard, and a weight of 3 indicates the task draws strongly upon that standard. A blank cell indicates that the task bears no significant relationship to the standard. The weights for the 1994 Mathematics Reference Examination tasks are given in Tables 3.1 (elementary school) and 3.2 (middle school).

Many tasks involve more than one standard. Where more than one is relevant to a task, the strength of the relationship between standards and the task may vary. Different tasks based on the same standard vary in two ways — the degree in which they elicit performances that are central to that standard, and the depth of understanding that good performances on those tasks can be expected to exhibit.

This mapping process was completed in two phases.

Phase 1. Participants asked themselves these questions about each task:

- What is the main challenge in this task? Which standard(s) are at its core? Problem Solving? Skills and Tools? Concepts?

Table 3.1 Weights relating elementary examination tasks to standards[1].

	Standard[2]							
Tasks by Form	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Multi-Task Form E1								
Sharing a Cake	3				2		1	
Fair Game?	1			3			1	
Recycling	3					2	1	
Marble Estimation	3	1					1	
Rule It Out	1	3				1		
Multi-Task Form E2								
Helping Martin	3					3	1	
Willie’s Patterns			3					
Fourths?	2	3					1	
Coins (1 of 3)	1			3		1		
Coins (2 of 3)	3					3		
Coins (3 of 3)	1					3		
Secret Number	3		2		1			
Multi-Task Form E3								
Counting Raisins Part 1(1 of 2)	3					1		
Counting Raisins Part 1(2 of 2)	1			3		1		
Counting Raisins Part 2	3						1	
Jana’s Number	3		2		1			
It’s a Half	3	1			1	1		
Building With Squares		3		2	1			
Jessie’s Marbles	3	2						
Multi-Task Form E4								
Symmetry in Shapes		2						
Favorite Fruit Part 1				2				
Favorite Fruit Part 2				2			1	
Carpet Job		3			2		1	
Marvelous Marbles			3				1	
Single-Task Form E5								
Order/Reorder				3	3	2	3	
Single-Task Form E6								
Partitions		2	2		3		3	
Single-Task Form E7								
School Supplies	3				3	3	3	
Single-Task Form E8								
A Trip to the State Park	3				3	3	3	

1. Weight values reflect the strength of the relationships between the task and standards. A weight of 1 indicates that the task draws weakly upon the standard, and a weight of 3 indicates the task draws strongly upon that standard.

2. The standards are:

1. Number and Operation Concepts	5. Problem Solving and Mathematical Reasoning
2. Geometry and Measurement Concepts	6. Mathematical Skills and Tools
3. Function and Algebra Concepts	7. Mathematical Communication
4. Statistics and Probability Concepts	8. Putting Mathematics to Work (Not assessed in the Reference Examination).

Table 3.2 Weights relating middle school examination tasks to standards[1].

	Standard [2]							
Tasks by Form	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Multi-Task Form M1								
Cubes		3	1			2	2	
Odd or Even				3	1		2	
Life Expectancy (1 of 3)				2		3	1	
Life Expectancy (2 of 3)				3		2	1	
Life Expectancy (3 of 3)				3		1	1	
Multi-Task Form M2			3		2	1	3	
Graphing		1				3		
Comparing Triangles (1 of 3)		1				3		
Comparing Triangles (2 of 3)		1				3		
Comparing Triangles (3 of 3)		3			2	2	2	
Comparing Drawings	2		2				2	
Thermometer	1					3		
Bake Sale	1					3		
Multi-Task Form M3								
Chocolate Chip Cookies (1 of 2)	2					3		
Chocolate Chip Cookies (2 of 2)	2					3		
Changing Percents	3					1	2	
Area		3			2	2	2	
Copy Shop	1		2		3	2	2	
Multi-Task Form M4								
Moe's Patterns (1 of 3)			3				2	
Moe's Patterns (2 of 3)			3				2	
Moe's Patterns (3 of 3)			3				2	
Eating Pizza	2					2	1	
The Better Airline Deal	2					2	1	
Science Fair	3					3		
How Cold Is It?	2		1			2	1	
Single-Task Form M5								
Carnival Game				3	3	2	3	
Single-Task Form M6								
Design a Box		2			3	2	3	
Single-Task Form M7								
Sports Bag		3			3	2	3	
Single-Task Form M8								
Truth in Advertising				2	3	1	3	

[1] Weight values reflect the strength of the relationships between the task and standards. A weight of 1 indicates that the task draws weakly upon the standard, and a weight of 3 indicates the task draws strongly upon that standard.

[2] The standards are:

1. Number and Operation Concepts
2. Geometry and Measurement Concepts
3. Function and Algebra Concepts
4. Statistics and Probability Concepts
5. Problem Solving and Mathematical Reasoning
6. Mathematical Skills and Tools
7. Mathematical Communication
8. Putting Mathematics to Work (Not assessed in the Reference Examination).

- What other standards does the task draw upon? Perhaps these are subsidiary skills and tools that are still essential to a good performance.

Working from resources that included the standards and the task itself (including the mathematical commentary and the rubric), each participant completed a preliminary mapping of standard by task.

All maps were compared and differences among the maps were discussed task by task. If some participants identified a relationship and others did not, those who did defended their decision using the task and standards resources. If they were convincing, the cell defined a relationship, but if they were convinced otherwise, no relation was identified or a decision was deferred until Phase 2.

Phase 2. Using the same resources as in Phase 1 plus about 50 samples of student work, the mathematics experts first returned to any tasks for which the mapping was not completed during Phase 1. For these tasks, student work was now considered. Then the group used all the resources at hand, particularly the student work, as the basis for one more round of discussion to decide whether or not to map any of the cells for which there was not consensus during Phase 1.

To complete the map, weights were assigned. In order to assign the weights, individuals examined student work and reviewed the other resources used during Phase 1. In assigning the weights, participants asked themselves how centrally or deeply the task draws upon the core strategies, understanding, or skills and tools elaborated in the standard. Finally, all weights were compared and differences were discussed task by task. As in Phase 1, the end result of the discussion was a consensus based on the evidence in the standards and the task and its rubric, mathematical commentary, and student work.

3.2 Examination Scores

New Standards examination scores are referenced to standards. The examination is constructed to provide information about performance keyed to a specific set of standards. The performance of a single task on a test does not contain enough information to infer unambiguously whether or not a student has met a standard, but reference examinations include a series of tasks that together are sufficient to draw such inferences. The basis for an inference about a specific standard is a collection of performances on tasks which address that standard. A score for a standard is generated by combining scores on such tasks. Not all standards in a content area can be scored from an on-demand reference examination, primarily because of time limitations on performance. Even those that can be scored may have small numbers of tasks and therefore, large measurement errors. Therefore, by combining standards into conceptually coherent clusters, measurement errors may be decreased by increasing the number of tasks contributing to a score. Because of this, results were reported on clusters that consisted of one or more standards (see Section 1.1). The weights shown in the maps in Tables 3.1 and 3.2 were summed to produce task weights for Conceptual Understanding (the sum of the task weights for Standards 1 to 4), Skills (the task weights for Standard 6), and Problem Solving (the sum of the task weights for Standards 5 and 7). These sets of task weights for each cluster are shown in Tables 3.3 and 3.4. Note that the weights vary depending on the specific multi-task and long -task booklets taken by a student.

The process used to report the 1994 Reference Examination results had the following steps (Figure 3.1):

Step 1: Task Scoring

An individual student's tasks were <u>scored</u> using the rubric. This step was discussed in Sections 2.1 and 2.2.

Step 2a: Recoding Task Scores

In this step, the student's rubric scores for tasks **were** <u>recoded</u> to value the evidence from a task about a standard. These values were created to recode each rubric score point for a specific task into a decimal fraction between zero and one. Fractions closer to one represent more evidence than fractions closer to zero. These fractions reflect the degree of evidence contained in the performance of the task about a student's likelihood of being **Above, At, and Below** the standard (Table 3.5).

The metric of the values was set so that averages are interpretable as probabilities or proportions. The rubric scores of tasks were recoded to value the evidence from the task about a standard. This table was a consensus among a group of mathematics educators who were responsible for production and assembly of the test tasks and the task scoring rubrics.

Step 2b: Creating Task Weights

Weights **were** used to show the relationship of each task to each cluster of standards. Recall that the weights relating the tasks to the standards were summed to produce sets of task weights for Concepts, Skills, and Problem Solving. This step was described in Sections 3.1 and 3.2.

Step 3: Producing Individual Cluster Scores

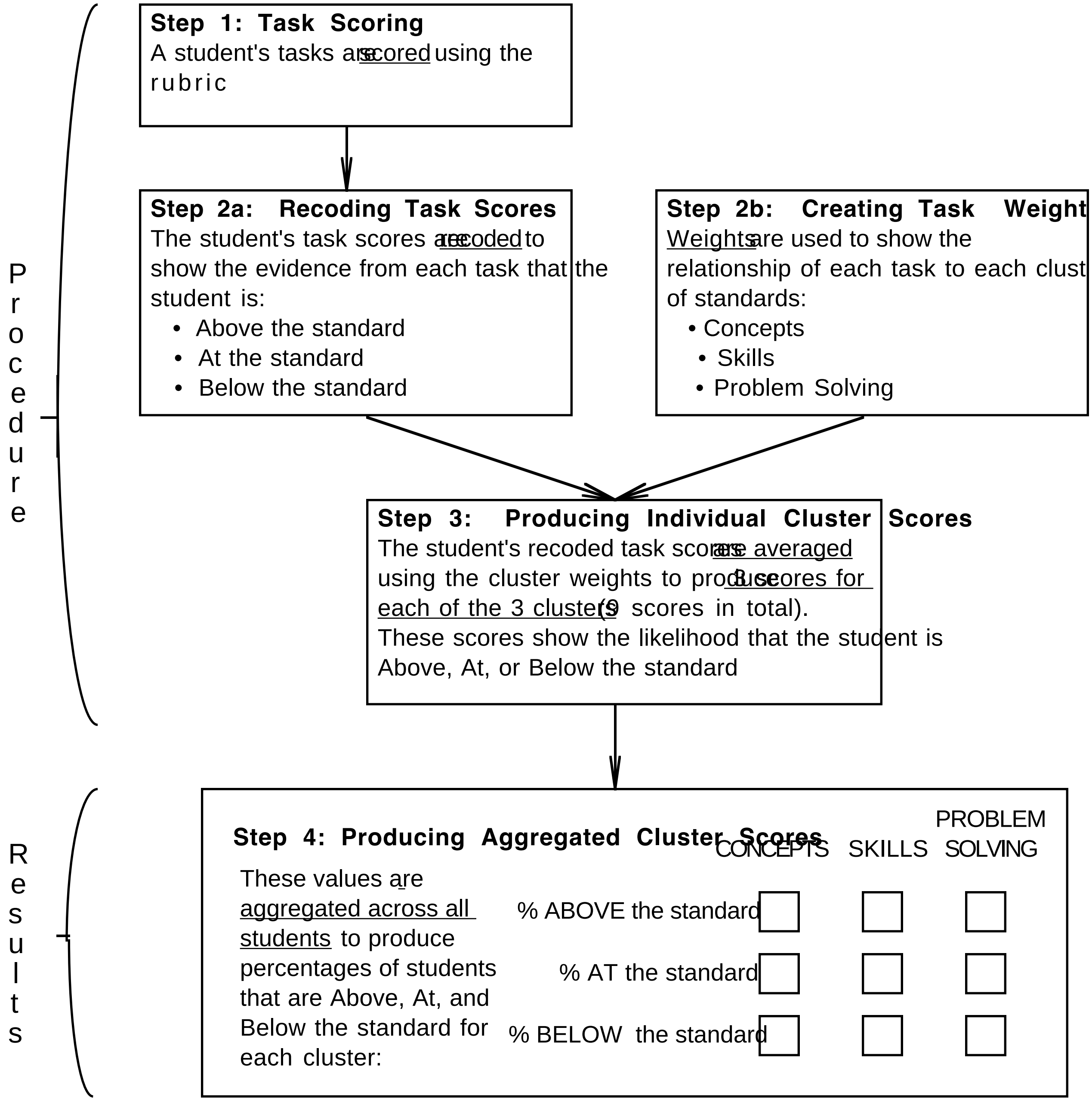
The recoded task scores from Step 2a **were** averaged using the task weights for each cluster that were created in Step 2b. This step produced three scores for each of the three clusters. That is, scores **were** produced that show the likelihood of the student being **Above, At, and Below** the standard in Concepts, Skills, and Problem Solving.

Table 3.3 Weights relating elementary school examination tasks to clusters (summed over standards).

	Standards Cluster		
Tasks by Form	**Concepts**	**Skills**	**Problem Solving**
Multi-Task Form E1			
Sharing a Cake	3	0	3
Fair Game?	4	0	1
Recycling	3	2	1
Marble Estimation	4	0	1
Rule It Out	4	1	0
Multi-Task Form E2			
Helping Martin	3	3	1
Willie’s Patterns	3	0	0
Fourths?	5	0	1
Coins (1 of 3)	4	1	0
Coins (2 of 3)	3	3	0
Coins (3 of 3)	1	3	0
Secret Number	5	0	1
Multi-Task Form E3			
Counting Raisins Part 1(1 of 2)	3	1	0
Counting Raisins Part 1(2 of 2)	4	1	0
Counting Raisins Part 2	3	0	1
Jana’s Number	5	0	1
It’s a Half	4	1	1
Building With Squares	5	0	1
Jessie’s Marbles	5	0	0
Multi-Task Form E4			
Symmetry in Shapes	2	0	0
Favorite Fruit Part 1	2	0	0
Favorite Fruit Part 2	2	0	1
Carpet Job	3	0	3
Marvelous Marbles	3	0	1
Single-Task Form E5			
Order/Reorder	3	2	6
Single-Task Form E6			
Partitions	4	0	6
Single-Task Form E7			
School Supplies	3	3	6
Single-Task Form E8			
A Trip to the State Park	3	3	6

Table 3.4 Weights relating middle school examination tasks to clusters (summed over standards).

	Standards Cluster		
Tasks by Form	**Concepts**	**Skills**	**Problem Solving**
Multi-Task Form M1			
Cubes	4	2	2
Odd or Even	3	0	3
Life Expectancy (1 of 3)	2	3	1
Life Expectancy (2 of 3)	3	2	1
Life Expectancy (3 of 3)	3	1	1
Multi-Task Form M2			
Graphing	3	1	5
Comparing Triangles (1 of 3)	1	3	0
Comparing Triangles (2 of 3)	1	3	0
Comparing Triangles (3 of 3)	1	3	0
Comparing Drawings	3	2	4
Thermometer	4	0	2
Bake Sale	1	3	0
Multi-Task Form M3			
Chocolate Chip Cookies (1 of 2)	2	3	0
Chocolate Chip Cookies (2 of 2)	2	3	0
Changing Percents	3	1	2
Area	3	2	4
Copy Shop	3	2	5
Multi-Task Form M4			
Moe’s Patterns (1 of 3)	3	0	2
Moe’s Patterns (2 of 3)	3	0	2
Moe’s Patterns (3 of 3)	3	0	2
Eating Pizza	2	2	1
The Better Airline Deal	2	2	1
Science Fair	3	3	0
How Cold Is It?	3	2	1
Single-Task Form M5			
Carnival Game	3	2	6
Single-Task Form M6			
Design a Box	2	2	6
Single-Task Form M7			
Sports Bag	3	2	6
Single-Task Form M8			
Truth in Advertising	2	1	6



Step 1: Task Scoring
A student's tasks are scored using the rubric

Step 2a: Recoding Task Scores
The student's task scores are recoded to show the evidence from each task that the student is:

- Above the standard
- At the standard
- Below the standard

Step 2b: Creating Task Weight
Weights are used to show the relationship of each task to each clust of standards:

- Concepts
- Skills
- Problem Solving

Step 3: Producing Individual Cluster Scores
The student's recoded task scores are averaged using the cluster weights to produce 3 scores for each of the 3 clusters (9 scores in total).
These scores show the likelihood that the student is Above, At, or Below the standard

Step 4: Producing Aggregated Cluster Scores
These values are aggregated across all students to produce percentages of students that are Above, At, and Below the standard for each cluster:

	CONCEPTS	SKILLS	PROBLEM SOLVING
% ABOVE the standard	□	□	□
% AT the standard	□	□	□
% BELOW the standard	□	□	□

Figure 3.1 Outline of the process used to report the 1994 Reference Examination results.

Table 3.5 Recoding scheme for task scores to show evidence that a student is above, at, or below the standard for a standards cluster.

	Recoded Value for a Student's Likelihood of Being:		
Task Score	**Above the Standard**	**At the Standard**	**Below the Standard**
5	1.000	0	0
4	.950	.050	0
3	.800	.150	.050
2	.100	.200	.700
1	.010	.090	.900
0	.005	.045	.950

As an example, Table 3.6 summarizes the calculations for estimating the likelihood that a student who received scores of 0, 2, 3, 2, and 4 on the elementary school form E1 tasks was Above, At, or Below the standard for the Concepts cluster. Note that the one set of tasks scores produces three sets of recoded values.

Table 3.6 Calculations for estimating the likelihood of being above, at, or below standard.

			Recoded Value[2]		
Task	**Weights[1] for Concepts**	**Task Score**	**Above the Standard**	**At the Standard**	**Below the Standard**
T1	3	0	.005	.045	.950
T2	4	2	.100	.200	.700
T3	3	3	.800	.150	.050
T4	3	2	.100	.200	.700
T5	4	4	.950	.050	.000
		Mean[3]	**.407**	**.129**	**.465**

[1] Values from Table 3.3.

[2] Values from Table 3.5.

[3] Weighted mean. Computed as the sum of products of weights (w) and values (s) divided by the sum of weights. E.g., Weighted mean $= \sum(ws) \Big/ \sum w$.

Step 4: Producing Aggregated Cluster Scores
The values produced in Step 3 were aggregated across all students and test forms to produce percentages of students that were Above, At, and Below the standard in Concepts, Skills, and Problem Solving clusters.

3.3 Highlights of Examination Results

Using the procedures outlined in Section 4, cluster scores were computed for elementary and middle schools (Table 3.7).

For elementary school students, these highlights emerge from the data:

- About one-third of the students were above the standard in Concepts and Skills and Tools, but just under one-quarter were above standard in Problem Solving and Communication.
- About 15% of students were at standard in all clusters.
- In all clusters, one-half or more of the students were below standard, but in Problem Solving and Communication the number below standard was more than 60%.

Against the standards set for middle school students, the examination results indicate an even weaker performance than for the elementary students:

- Under one-quarter of the students were above the standard in Concepts and Skills and Tools, and only 18% of the students were above the standard in Problem Solving and Communication.
- About 15% performed at standard across all students.
- More than two-thirds of the students were below standard in Problem Solving and Communication.

Cluster scores were also compared by gender, ethnicity, and English language proficiency (Table 3.8. The scores of males and females were close in all clusters at both the elementary and middle schools.

- Among elementary students, whites and Asians scored better in all clusters, with the scores of Filipinos slightly lower. All other groups scored lower across clusters.

- In middle school, Asians and whites scored higher than all other groups, but scores of Filipinos.

Table 3.7 Estimated percent of elementary and middle school students above, at, and below the standard (by standards cluster).

	Estimated Percent of Students at Different Levels by Standards Cluster[1]		
Standards Level	**Concepts[1]**	**Skills & Tools[2]**	**Problem Solving & Communication[3]**
Elementary School			
Above the Standard	33.9	35.8	24.1
At the Standard	15.9	14.9	14.1
Below the Standard	50.3	49.3	61.9
Middle School			
Above the Standard	22.9	24.7	18.0
At the Standard	15.9	16.2	14.2
Below the Standard	61.3	59.1	67.9

1. Conceptual Understanding ....Standards 1, 2, 3, 4
2. Mathematical Skills and Tools ....Standard 6
3. Problem Solving and Mathematical
   Reasoning/Mathematical Communication.............Standards 5, 7

Table 3.8 Estimated percent of elementary and middle school students above, at, and below the standard (by standards cluster and subgroup).

		Estimated Percent of Students by Standard Level by Standards Cluster and Subgroup								
		Concepts[2]			Skills & Tools[3]			Problem Solving & Communication[4]		
Subgroup	**N[1]**	**Above**	**At**	**Below**	**Above**	**At**	**Below**	**Above**	**At**	**Below**
Elementary School										
Gender										
Male	7,698	33.7	15.9	50.4	35.2	14.8	50.0	23.6	14.0	62.5
Female	7,562	34.0	15.9	50.1	36.5	14.9	48.5	24.6	14.2	61.2
Ethnicity										
African-American	2,003	23.1	16.1	60.8	25.4	14.9	59.7	14.2	13.5	72.4
Native American	174	25.6	16.2	58.2	27.8	15.6	56.6	16.9	14.2	68.9
Asian	361	37.0	15.8	47.2	38.1	14.7	47.2	28.1	14.1	57.7
Filipino	84	35.5	15.6	48.8	32.8	14.0	53.1	23.8	13.8	62.4
Hispanic/Latino	1,253	23.8	16.2	60.0	26.1	15.0	58.9	14.6	13.8	71.6
Pacific Islander	66	25.2	16.3	58.4	26.4	15.1	58.4	17.3	14.3	68.4
White (not Hispanic)	9,571	37.5	15.8	46.6	39.9	14.9	45.2	27.5	14.3	58.1
English Proficiency Status										
English Proficient	13,299	34.3	15.9	49.8	36.5	14.9	48.6	24.5	14.2	61.3
Limited English Proficiency	325	22.6	16.0	61.4	23.0	14.6	62.5	14.4	13.1	72.5
Middle School										
Gender										
Male	8,648	23.4	15.8	60.8	25.4	16.2	58.4	18.3	14.1	67.6
Female	8,803	22.9	15.9	61.2	24.3	16.3	59.4	18.0	14.3	67.8
Ethnicity										
African-American	1,764	13.1	15.6	71.3	14.2	16.1	69.6	8.9	13.0	78.2
Native American	317	15.7	15.6	68.7	16.9	16.2	66.9	11.0	13.3	75.7
Asian	548	27.1	15.7	57.2	29.1	16.2	54.8	21.1	14.4	64.6
Filipino	224	19.2	15.8	65.0	21.3	16.4	62.3	14.2	13.8	71.9
Hispanic/Latino	1,611	14.1	15.6	70.3	15.3	16.2	68.6	9.9	13.2	76.9
Pacific Islander	75	16.2	15.7	68.1	16.8	15.8	67.4	11.3	13.3	75.4
White (not Hispanic)	12,384	25.9	15.9	58.1	27.9	16.3	55.9	20.8	14.5	64.7
English Proficiency Status										
English Proficient	14,790	23.7	15.9	60.5	25.4	16.2	58.4	18.7	14.3	67.0
Limited English Proficiency	280	17.8	15.6	66.5	19.2	16.3	64.5	13.1	13.6	73.3

[1] Number of students in subgroup

[2]. Conceptual Understanding

[3]. Mathematical Skills and Tools

[4]. Problem Solving and Mathematical Reasoning/Mathematical Communication

4. Accuracy of the Reference Examination

The methodology for assessing the accuracy of large-scale assessments that use only a few reporting categories has changed dramatically in the last few years. The impetus for these changes has come mainly from recommendations taken from the Select Committee Report on the statistical procedures used in the 1993 California Learning Assessment System (CLAS) (Cronbach, Bradburn, and Horvitz, 1994) and implemented in the 1994 CLAS administration (CTB, 1994a; CTB, 1994b; Wiley, 1994). The current methodology requires the calculation of standard errors of reported estimates. The standard errors assess the level of accuracy of these estimates.

The standard errors must take into account multiple sources of variability of student performance arising from parts of the measurement process which are idiosyncratic (i.e., inconsistencies in student and school performance from task to task and form to form). Standard errors were calculated using the framework of generalizability theory (G-theory) (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Brennan, 1983; Shavelson and Webb, 1991).

4.1 G-study Design for the Assessment

In order to assess the accuracy of the 1994 Mathematics Reference examination, the following prepatory steps were made:

- Each school is assigned to one of four groups, depending upon which single-task booklet was used at that school;

- For each school within one of these groups, a set of four “psuedoforms” were created by crossing the single-task booklet that the school used with the four multi-task booklets that were spiraled within the school;

Note that under this arrangement

1. For each group, schools *(s)* and psuedoforms *(f)* were completely crossed – that is, each school within a group was administered all four of the pseudoforms for that group;

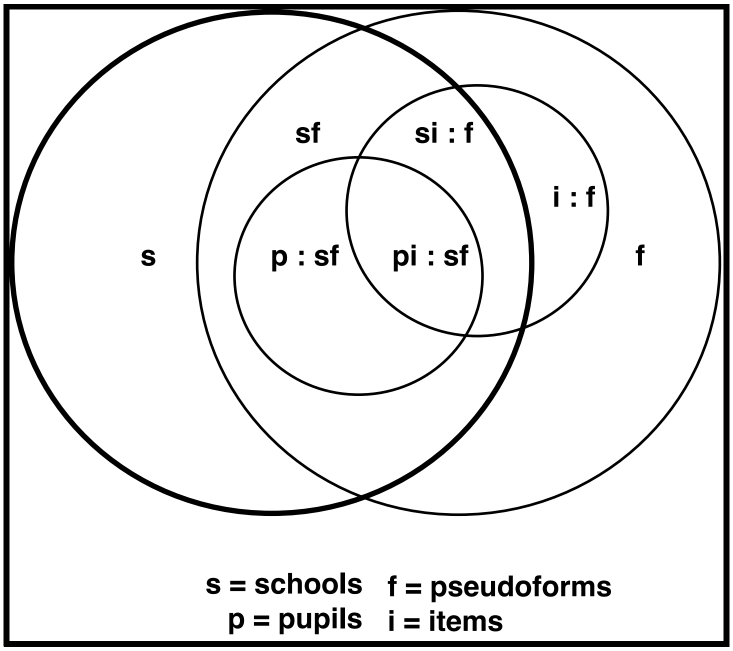


Figure 4.1 The $\boxed{p{:}(s \times i{:}f)}$ design

2. Except for the common 45 minute task taken from the single-task booklet, all of the other tasks *(i)* were nested within one of the psuedoforms;

3. Given the previous two conditions, this implies that schools and tasks-within-psuedoforms were completely crossed as well;

4. Each student *(p)* was nested within a school;

5. Each student was also nested within a pseudoform – that is, each student completed only one of the four psuedoforms;

6. Combining 4 and 5, each student was also nested within each combination of school and pseudoform;

7. Each student completed each of the tasks within a pseudoform.

This complicated pattern of crossing and nesting was denoted as a $p{:}(s \times i{:}f)$ G-study design outlined in the Venn diagram in Figure 4.1. The object of measurement (school *s*) is represented by the thick black circle; the facets or conditions of measurement (pseudoform *f*, item *i*, person *p*) are represented by the lighter black circles. The Venn diagram also shows which components are estimable within the context of the design. Specifically, the seven components are school *(s)*, pseudoform *(f)*, school by pseudoform (*sf)*, task nested within pseudoform (*i:f)*, person nested within school-by-psuedoform (p:sf), school crossed with task nested within psuedoform (*si:f)*, and the person crossed with task, nested within school crossed with psuedoform (*pi:sf*). The last term also contains the residual error (*e*).

The estimates of percent of pupils above, at, or below standard for each cluster were the numbers for which standard errors were required. The recoded task scores (Section 3.2) were used as the dependent variables for the G-study. An *analysis of variance* (*ANOVA*) was performed using the design shown in Figure 4.1. From this design, estimates of variance components were calculated.

Only a sample of data for each set of psuedoforms was used in the analysis. Schools with at least two students taking each of the psuedoforms were eligible for inclusion in the sample. The variance components were calculated by setting the *expected mean square* (EMS) equations for the design equal to the sample mean squares (MS) values taken from the ANOVA table (for details see Cornfield and Tukey, 1956; Searle and Fawcett, 1970; Kirk, 1982; Brennan, 1983).

In the implementation of this design, several assumptions and simplifications were made:

- The facets for schools, students, items, and psuedoforms were assumed to be random;
- No finite correction factors have been used;
- In the elementary schools, two students per psuedoform were sampled and in the middle schools, three students per psuedoform were sampled.

This study produced estimates of variability (variance components) for the sources listed above for each cluster. These estimates for the percent above standard are shown in Table 4.1 through 4.3 for elementary school, and 4.4 through 4.6 for middle school.

The estimated standard errors of the percent above standard for each cluster at the partnership level are shown in Tables 4.7 through 4.9 for elementary school and 4.10 through 4.12 for middle school. These tables show the varying contributions of the variance components to the overall error variance and the standard errors (**SE**).

Table 4.1 Variance component estimates for elementary school Concepts cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**E5**	**E6**	**E7**	**E8**	**Mean**
s	.0083	.0083	.0061	.0050	.0069
f	.0012	.0013	.0019	.0006	.0013
i:f	.0206	.0196	.0182	.0223	.0202
sf	**0**[1]	**0**	.0002	.0010	.0003
si:f	.0048	.0033	.0022	.0042	.0036
p:sf	.0158	.0170	.0180	.0152	.0165
pi:sf,e	.0788	.0846	.0895	.0836	.0841

[1] Denotes a negative variance component set to zero.

Table 4.2 Variance component estimates for elementary school Skills cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**E5**	**E6**	**E7**	**E8**	**Mean**
s	.0088	.0077	.0069	.0065	.0075
f	.0036	.0051	.0021	.0049	.0039
i:f	.0191	.0200	.0167	.0031	.0147
sf	**0**[1]	**0**	.0016	**0**	.0004
si:f	.0082	.0038	.0019	.0040	.0045
p:sf	.0144	.0162	.0182	.0236	.0181
pi:sf,e	.0811	.0861	.0956	.0966	.0899

[1] Denotes a negative variance component set to zero.

Table 4.3 Variance component estimates for elementary school Problem Solving cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**E5**	**E6**	**E7**	**E8**	**Mean**
s	.0108	.0107	.0075	.0055	.0086
f	**0**[1]	**0**	**0**	**0**	**0**
i:f	.0169	.0197	.0164	.0224	.0189
sf	**0**	**0**	**0**	.0014	.0004
si:f	.0050	.0045	.0016	.0050	.0040
p:sf	.0194	.0189	.0209	.0140	.0183
pi:sf,e	.0772	.0840	.0934	.0859	.0851

[1] Denotes a negative variance component set to zero.

Table 4.4 Variance component estimates for middle school Concepts cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**M5**	**M6**	**M7**	**M8**	**Mean**
s	.0097	.0104	.0086	.0118	.0101
f	**0**	**0**	**0**	**0**	**0**
i:f	.0137	.0167	.0107	.0124	.0134
sf	**0**	.0014	.0001	**0**	.0004
si:f	.0024	.0033	.0045	.0033	.0034
p:sf	.0135	.0143	.0154	.0166	.0150
pi:sf,e	.0576	.0571	.0598	.0618	.0591

1 Denotes a negative variance component set to zero.

Table 4.5 Variance component estimates for middle school Skills cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**M5**	**M6**	**M7**	**M8**	**Mean**
s	.0106	.0110	.0090	.0126	.0108
f	**0**	**0**	**0**	**0**	**0**
i:f	.0168	.0203	.0130	.0146	.0162
sf	**0**	.0012	**0**	**0**	.0003
si:f	.0023	.0036	.0047	.0027	.0033
p:sf	.0153	.0147	.0162	.0171	.0158
pi:sf,e	.0560	.0555	.0605	.0616	.0584

1 Denotes a negative variance component set to zero.

Table 4.6 Variance component estimates for middle school Problem Solving cluster (Above the Standard).

	Variance Component by Pseudoform Set				
Source	**M5**	**M6**	**M7**	**M8**	**Mean**
s	.0087	.0099	.0095	.0114	.0099
f	**0**	**0**	.0001	**0**	<.0001
i:f	.0132	.0155	.0088	.0110	.0121
sf	**0**	.0011	.0003	.0000	.0004
si:f	.0023	.0037	.0049	.0029	.0035
p:sf	.0116	.0115	.0138	.0152	.0130
pi:sf,e	.0582	.0589	.0618	.0634	.0606

1 Denotes a negative variance component set to zero.

Table 4.7 Standard error estimate for elementary school Concepts cluster (Above the Standard).

Source	Variance Component	n	Contribution
s	.0069	240	.000029
f	.0013	16	.000081
i:f	.0202	28	.000721
sf	.0003	960	.000000
si:f	.0036	6,720	.000001
p:sf	.0165	3,840	.000004
pi:sf,e	.0841	107,520	.000001
		Variance	.000837
		SE	.0289

Table 4.8 Standard error estimate for elementary school Skills cluster (Above the Standard).

Source	Variance Component	n	Contribution
s	.0075	240	.000031
f	.0039	16	.000244
i:f	.0147	12	.001225
sf	.0004	960	.000000
si:f	.0045	2,880	.000002
p:sf	.0181	3,840	.000005
pi:sf,e	.0899	46,080	.000002
		Variance	.001509
		SE	.0388

Table 4.9 Standard error estimate for elementary school Problem Solving cluster (Above the Standard).

Source	Variance Component	n	Contribution
s	.0086	240	.000036
f	**0**[1]	16	.000000
i:f	.0189	18	.001050
sf	.0004	960	.000000
si:f	.0040	4,320	.000001
p:sf	.0183	3,840	.000005
pi:sf,e	.0851	69,120	.000001
		Variance	.001093
		SE	.0331

[1] Denotes the value of a negative variance component that was set to zero.

Table 4.10 Standard error estimate for middle school Concepts cluster (Above the Standard).

Variance Component	Mean	n	Contribution
s	0.0101	240	.000042
f	**0**[1]	16	.000000
i:f	0.0134	28	.000479
sf	0.0004	960	.000000
si:f	0.0034	6,720	.000001
p:sf	0.015	3,840	.000004
pi:sf,e	0.0591	10,7520	.000001
		Variance	.000526
		SE	.0229

[1] Denotes the value of a negative variance component that was set to zero.

Table 4.11 Standard error estimate for middle school Skills cluster (Above the Standard).

Variance Component	Mean	n	Contribution
s	0.0108	240	.000045
f	**0**[1]	16	.000000
i:f	0.0162	23	.000704
sf	0.0003	960	.000000
si:f	0.0033	5,520	.000001
p:sf	0.0158	3,840	.000004
pi:sf,e	0.0584	88,320	.000001
		Variance	.000755
		SE	.0275

[1] Denotes the value of a negative variance component that was set to zero.

Table 4.12 Standard error estimate for middle school Problem Solving cluster (Above the Standard).

Variance Component	Mean	n	Contribution
s	0.0099	240	.000041
f	0.0001	16	.000006
i:f	0.0121	21	.000576
sf	0.0004	960	.000000
si:f	0.0035	5,040	.000001
p:sf	0.013	3,840	.000003
pi:sf,e	0.0606	80,640	.000001
		Variance	.000629
		SE	.0251

5. CONCLUSION

The New Standards partnership is committed to improving educational achievement for all American children. We believe that clear standards and standards-based assessments are essential to this process. When expectations are clear to students and teachers, learning and teaching efforts can be effectively directed and achievement will grow.

The results of the 1994 Mathematics Reference Examination provide important lessons for our efforts to improve achievement. In future years — beginning with the 1995-96 school year — the plans are to modify the methodology used to enable us to calculate scores for individual students. Reference examinations will be available that will allow schools to track their progress in bringing more students "up to standard." Individual students and their parents will be able to see how they are doing against a nationally and internationally benchmarked standard.

The standards embodied in the New Standards Reference Examinations clearly represent the fundamental knowledge that Americans agree all students need to master. Communities will develop many different ways to help their children meet the standards. But all who adopt these standards will be able to assure their children that they will be prepared for successful participation in further education, civic life, and the workplaces of the 21st century.

REFERENCES

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Commission on Standards for School Mathematics. (1989). *Curriculum and Evaluation: Standards for School Mathematics*. USA: National Council of Teachers of Mathematics.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics 27*, 907-949.

Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1994, July). Sampling and statistical procedures used in the California Learning Assessment System: Report of the Select Committee.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York, NY: John Wiley and Sons.

CTB/McGraw-Hill. (1994). I. Sampling and stratification for the variance component calculations in CLAS 94. 1994 CLAS Technical Specifications.

CTB/McGraw-Hill. (1994). II. Estimation of school-level standard errors (SEs) in CLAS 94. 1994 CLAS Technical Specifications.

New Standards. (1995). *Performance standards (volumes 1, 2, 3)*. Rochester, NY: National Center on Education and the Economy.

Kirk, R. E. (1982). *Experimental design (2nd ed.).* Belmont, CA: Brooks/Cole Publishing Co.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Searle, S. R., & Fawcett, R. F. (1970). Expected mean squares in variance component models having finite populations. *Biometrics 26*, 243-254.

Wiley, D. E. (1994). Estimating standard errors of "percent above cut" for schools, comparison groups, districts, and the state for CLAS 1994. 1994 CLAS Technical Specifications.

APPENDIX A: THE NEW STANDARDS SYSTEM

New Standards is a collaboration of the Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy, in partnership with states and urban school districts. The partners are building an assessment system to measure student progress toward meeting national standards at levels that are internationally benchmarked.

The Governing Board includes chief state school officers, governors and their representatives, and others representing the diversity of the partnership, whose jurisdictions enroll nearly half of the nation's students.

Founded by Lauren Resnick, Director of the Learning Research and Development Center (LRDC), and Marc Tucker, President of the National Center on Education and the Economy, New Standards staff is based at these organizations as well as the American Association for the Advancement of Science, the Fort Worth Independent School District, the National Council of Teachers of English, and the University of California Office of the President. Technical studies are based at LRDC and Northwestern University, advised by leading psychometricians from across the nation.

The New Standards assessment system has three interrelated components: performance standards, an on-demand examination, and a portfolio system.

The **performance standards** are derived from the national content standards developed by professional organizations, for example, the National Council of Teachers of Mathematics standards in Mathematics, and consist of two parts:

- *Performance descriptions* describe what students should know and the ways they should demonstrate the knowledge and skills they have acquired in the four areas assessed by New Standards — English Language Arts, Mathematics, Science, and Applied Learning — at elementary, middle, and high school levels.

- *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.

The performance standards were endorsed unanimously by the New Standards Governing Board in June 1995 for widespread consultation in 1995–96.

The on-demand examination, called the **reference examination** because it provides a point of reference to national standards, is currently available in English Language Arts and Mathematics at grades four, eight, and ten. Those aspects of the performance standards that can be assessed in a limited time frame under standardized conditions are covered in the reference examination. In English Language Arts, this means reading short passages and answering questions, writing first drafts, and editing. In Mathematics, this includes short exercises or problems that take 5 to 15 minutes and longer problems of up to 45 minutes. The reference examination stops short of being able to accommodate longer pieces of work — reading several books, writing with revision, conducting investigations in Mathematics and Science, and completing projects in Applied Learning — that are required by New Standards performance standards and the national consensus content standards from which they are derived.

The **portfolio system** complements the reference examination, providing evidence of the performance standards that depend on extended pieces of work (especially those that show revision) and accumulation of evidence over time. In 1994–95, using draft portfolio handbooks in English Language Arts and Mathematics, 3,000 teachers and almost 60,000 students participated in a field trial of the portfolio system. In addition to handbooks for students, teachers, and administrators, the current system provides example portfolios that contain concrete examples of expectations for students and teachers.

In 1995-96, the portfolio system trial is being extended to include Science and Applied Learning. The system has been revised in light of the first year's experience, with the goal of making it easier to understand and implement.