The New Standards Reference Examination
Standards-Referenced Scoring System

CSE Technical Report 470

David E. Wiley
Northwestern University

February 1997

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 90095-1522
(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 Education Research and Improvement or the U.S. Department of Education.

The New Standards Reference Examination Standards-Referenced Scoring System

David E. Wiley
Northwestern University

Lauren Resnick
CRESST/LRDC, University of Pittsburgh

Contents

0. Overview of This Report

This report describes the procedures by which the New Standards Reference Examination system for reporting and interpreting results in terms of standards was designed and implemented.

First (Section 1), the basis for standards-referenced assessment is outlined. This includes characterizations of (a) standards, (b) standards-referenced assessment, and (c) the assessment development process. In addition, it also briefly describes the methods used to evaluate the accuracy of reported scores and performance levels.

Second (Section 2), the procedures used for actually setting performance standards/levels are outlined. Appendices I-VI exhibit the technical details of (a) assessment score definitions, (b) characteristics of expert judges participating in the process, (c) the process itself, and (d) the final outcomes of the process.

1. The Basis for Standards-Referenced Assessment

What Is a Standard?

"Standards" have become politically controversial. In general terms, a standard is a criterion for an acceptable outcome. However, in the context of education, learning outcomes are the consequences of goal-based instruction. Thus the criterion for a successful outcome involves *what* as well as *how much* is learned. In New Standards, standards specify the desired contents of learning and exemplify student performances that successfully meet such standards. Standards defined in this way, however, are neither measurement targets nor test blueprints. New Standards has had to create standards specifications that allow relative precision in the definition of such targets and in the formulation of such blueprints.

What is a Standards-Referenced Assessment?

For an assessment to be standards referenced, there must exist (a) a set of standards, (b) a definition of measurement targets—that is, constructs—that are derived from these standards, (c) a test blueprint for a test that yields scores for

estimating the status of test respondents with respect to these constructs, and (d) criteria for successful performance in terms of these scores.

How Do You Develop a Standards-Referenced Assessment?

A standards-referenced assessment must have a blueprint that allows task development and assembly into a complete examination. The standards-referenced blueprint structures the question "Has the student learned what he or she was supposed to learn?" into a measurement framework. In such a framework, specific measurement targets are formulated in terms of the standards. Using these targets, tasks are developed with scoring rubrics, procedures and benchmark performances that are explicitly related to performance standards. Scoring procedures and criteria for sets of tasks that map to the measurement targets and standards are established. It also requires a development plan for test tasks, which results in a collection of tasks and scoring rubrics that conform to the blueprint.

Estimating Misclassification Error for Standards-Referenced Assessment

In New Standards, student performances were reported in terms of one of five possible classification levels (Standards Levels) using students' performances on the exams to assess how they are doing relative to the New Standards Performance Standards. In classifying students into Standards Levels, one measure of accuracy is the probability of correct classification and/or probabilities of misclassification.

New Standards applies a method, introduced in Livingston and Lewis (1995), for estimating the accuracy and consistency of Standards Levels classifications based on weighted composite test scores. In this method, the reliability of the score is used to estimate the effective test length in terms of discrete items. The true-score distribution is estimated by fitting a 4-parameter beta distribution (Lord, 1965). The conditional distribution of scores on an alternate form given the true score is estimated using a compound binomial distribution (Livingston & Lewis, 1995). This procedure and the resulting accuracy estimates will be described and reported in the New Standards 1997 Technical Report.

Estimating Standard Error for Standards-referenced Assessment Using Generalizability Study Design

Many studies (Cronbach, Linn, Brennan, & Haertel, 1995; Gao, Shavelson, & Baxter, 1994; Shavelson, Mayberry, Li, & Webb, 1990) discussed generalizability of large-scale performance assessments in estimating sampling error and measurement error (e.g., task variability, rater variability, task x pupil, task x reader, and pupil x reader, etc.). All of these methods are based on the test that consists of a known number of equally weighted items. However, essay tests and performance assessments typically are scored in a way that allows for partial credit on each item. In New Standards, test takers are classified on the basis of a composite score—a weighted mean of scores on two or more tasks or subtests, which are unequally weighted. The generalizability analyses of these weighted mean scores will be described and reported in the New Standards 1997 Technical Report.

2. Setting the Examination Standard: The New Standards Standards-Referenced Scoring Decision Rules

Examinations

In 1996, the New Standards Reference Examinations in Mathematics and English Language Arts were taken by students nationwide in Grades 4, 8, and 10. The exams included tasks based on the New Standards *Performance Standards*. The *Performance Standards* specify what students at each of these grade levels should know and be able to do and examples of real student work that illustrate that level of achievement. The Reference Examinations are designed to assess reading, writing and mathematics, by asking students to apply problem-solving, critical thinking and fundamental content area skills.

The development of new systems of assessing what students know and are able to do requires a new set of thinking about student performance. Interpreting and reporting student performance in a standards-based system cannot rely on the same set of assumptions underlying traditional, norm-referenced multiple-choice tests. Standards-based assessments do not produce scores that separate and distribute students across a traditional scale. Standards-based assessments do treat item(s) and task(s) as interchangeable and do not produce scores that primarily compare one student against another student.

The new paradigm needed to understand student performance scores in a system based on standards has several unique characteristics which are an important departure from traditional testing.

First, standards-based assessments provide scores which describe student performance against a set of standards, not against other students.

Second, standards based assessments are made up of tasks that are clustered together to assess several academic areas and skills. These clusters circumscribe the constructs that form the basis for the examination scores reported by New Standards. (See Appendix I for definitions of these constructs.)

Third, standards-based assessments produce scores that are based on a system of scoring that weights (see Appendix III) and averages task scores to produce construct scores and establishes cut points (see Appendix IV) on these scores to report what students know and are able to do.

Task Scoring

Each of the tasks on the examinations is objectively scored by well-trained, qualified scorers. The scorers use carefully developed scoring guides that were extensively pilot-tested to judge each student's response to the tasks. After being trained to use the scoring guides, scorers are continually monitored during the scoring process to assure consistent application of the New Standards scoring criteria. This yields fair, consistent and accurate scores on each task for each student.

Reporting

Because no single task can adequately represent a standard, reporting a student's performance against a standard requires summarizing information from several tasks. When New Standards reports reference exam results, the eight mathematics standards are clustered into three areas: Concepts, Skills, and Problem Solving; and the five (or, at high school, seven) English Language Arts standards are clustered into four areas: Writing, Reading for Basic Understanding, Reading Interpretation and Analysis, and Conventions. In each cluster of standards, student performance is reported in terms of one of five possible classification levels.

Decision Rules

The decision rules for determining student performance on New Standards' 1996 Reference Examinations in Mathematics and English Language Arts were established using a three-part process. These rules were set by a diverse group of qualified experts in each content area. The procedures used by these experts meet standards established by the assessment and measurement profession and reflect the values and goals of New Standards. The process ensured that the resulting decision rules are aligned with the standards of accomplishment described in the New Standards *Performance Standards*.

Step I: Preliminary Decision Rules Are Created by the Exam Developers

The first step in the process was the establishment of preliminary decision rules by the developers of the exams. These developers were selected for their knowledge of the content and pedagogical issues confronted by students and teachers at each age level and had supervised the development of each task and scoring rubric used. Bringing this familiarity to bear, they established preliminary decision rules for combining and classifying student scores on the tasks that make up each of the subtests for which final scores are reported.

Step II: Decision Rules Are Created by a Panel of Expert Judges

Qualified experts. The next step, a two-and-a-half-day standard-setting workshop, is the central part of the process. Qualified experts in each age and content area were selected and brought together to establish decision rules of their own, without having knowledge of the preliminary decision rules. See Appendix II for a description of the characteristics of these experts. There were three to five such experts employed for each of the six examinations; 26 experts participated in the workshop. These included classroom teachers, curriculum specialists, and university-based educators specializing in teaching mathematics or English language arts. Though some of these individuals had reviewed and commented on draft test tasks and rubrics and thus had familiarity with aspects of New Standards' development process, none played formal roles or had formal responsibilities for development. Many were introduced to New Standards for the first time during the course of the standard-setting workshop.

Training. These experts, or judges, as they were called during the workshop, were put through a training regimen to understand the goals and methods of the standard-setting workshop, the tasks of the assessment, and the scoring rubrics for each task. Because it was critical that each judge have a thorough understanding of the demands of the tasks and their relationship to the *Performance Standards*, judges were required to read the *Performance Standards* and work the tasks prior to attending the workshop. On-site training was conducted by the developers of the exams and focused on understanding the meaning of each of the score points for each task. During this training, judges reviewed multiple student work samples for each task.

Setting weights and cut points. Following this training, the judges worked through standardized procedures to allow them to come first to individual judgments and then to consensus judgments about the way the scores on the exam should be weighted and how the resulting aggregate scores should be classified. That is, after setting weights for the tasks, they set cut points for each of the five levels New Standards uses to classify student performance. For both weights and cut points, the process allowed judges first to come to individual judgments, followed by rich discussion of the pedagogical and content issues that went into making such judgments. The discussion was facilitated by trained leaders with content area qualifications who had not had any role in establishing the preliminary decision rules. Judges were asked to anchor their arguments in the *Performance Standards*, as well as their own expert opinions. Each discussion was documented by a designated note-taker. At the close of each discussion, consensus was reached by the judges on the weights or cut points that had been the subject of the discussion. (See Appendices III, IV, and V.)

Impact and revision. The decision rules that resulted from the weights and cut points provided by the judges were then applied to possible score profiles of students so that judges could see the impact of the decision rule on hypothetical cases. At this time, judges had an opportunity to amend their initial decision rule so that it was more in keeping with their understanding of the *Performance Standards*.

Evaluation. The process described above was followed for each of the areas for which New Standards reports scores, for each exam. At the close of the standard-setting workshop, judges completed independent evaluation questionnaires, in which they rated their comfort with the decision rules they had created. The overwhelming sense of the judges, with respect to all exams and all decision rules, was that these were rules they could defend.

Step III: Judges' Decision Rules Are Reviewed to Ensure Validity and Alignment With the Performance Standards

Following the standard-setting workshop, the decision rules were carefully reviewed to ensure their validity, fairness, and adherence to the *Performance Standards*. In most cases, the judges' decision rules were accepted without change as the final decision rules. In a few cases, minor modifications were made to the rules, but only if one of the following criteria was met:

1. There was evidence that the judges misunderstood the demands of the task or the scoring rubrics. Such evidence could be found in the discussion that the judges had in arriving at the components of the decision rule in question.
2. The judges' decision rule could be enhanced in its adherence to the principles of standard-setting endorsed by New Standards. Though judges were trained in these principles, they sometimes did not adhere as closely to them as would be necessary to ensure a valid and fair classification system.

A detailed description of the complete Standard Setting process is given in Appendix V (A: Mathematics; B: English/Language Arts).

The final decision rules (weights and cuts) are given in Appendix VI.

Appendix I. Construct Definitions for Summary Scores and Test Structure

The tasks or parts of tasks that are selected to provide a balanced assessment of Performance Standards 1 through 7 are classified according to one of three constructs: Mathematical Skills, Conceptual Understanding, and Mathematical Problem Solving. A single rubric is developed for each task or for each part of a task that warrants a separate score. No single rubric contributes to more than one score.

Mathematical skill tasks are broadly described as those that create the opportunity for students to apply a well practiced and important routine or algorithm. Tasks or parts of tasks designed to assess skills are routine, always short, and rarely cast in a context. Accomplishment of this type of task draws heavily on recall and solutions are characterized largely but not entirely by manipulation. Tasks designed to assess mathematical skill generally have a single correct answer. It is likely that students will have learned how to solve mathematical skills tasks in class.

Conceptual understanding tasks are broadly described as those that usually create the opportunity for students to analyze an idea, to reformulate it, and to express it in their own terms. Tasks designed to assess conceptual understanding are usually nonroutine, short, and cast in a context. Conceptual understanding tasks can be thought of as "idea probes." Usually the accomplishment of a conceptual understanding task draws heavily on reconstruction rather than on recall; solutions are characterized by representation or explanation rather than by the manipulation. Often a short written explanation is sufficient to accomplish the task. These are the kinds of tasks that students can do easily if they understand the mathematics involved.

Profile of mathematical skills tasks is that they:

- assess recall of basic procedures or/and manipulations outlined in the HS Performance Standards.
- have a medium to low strategic hurdle.
- have a medium to low conceptual hurdle.
- are routine.
- are cast using a simple context or no context at all.
- take 1-2 minutes to complete.

Profile of conceptual understanding tasks is that they:

- assess use of mathematical concepts outlined in the HS Performance Standards. "Use" that is defined here is not confined to procedural use, but requires use with reformulation.
- Not more than 25% of tasks that qualify as conceptual understanding tasks in any one exam can ask students to explain a concept.
- have a medium to low strategic hurdle.
- have a medium to low skills hurdle.
- take 1-5 minutes to complete.

Problem-solving tasks are described as those that create the opportunity for students to select and deploy problem-solving strategies. Tasks designed to assess problem solving are usually nonroutine, long, and cast in a context. One way of thinking about the assessment of problem solving that is useful to us is to think of it as a measure of what students can do with the mathematics that they have learned one or even two years previously. Our most successful problem-solving tasks make high-level use of well-assimilated facts, concepts, and skills. Appropriately cast problem-solving tasks ensure that students are given the opportunity to formulate an approach to the problem and implement a solution.

Profile of mathematical problem-solving tasks is that they:

- assess use of mathematical concepts and skills to formulate and implement a solution to a problem.
- At least 25% of the tasks that qualify as problem-solving tasks in any one exam require a form of conclusion that provides a generalization or form of summary.
- have a medium to low conceptual hurdle.
- have a medium to low skills hurdle.
- are nonroutine.
- take 10-25 minutes to complete.

I. A. Design of the Test Instruments for Mathematics

Number of items by form, construct, and grade

	Elementary				Middle				High			
# Items	S	C	PS	Total	S	C	PS	Total	S	C	PS	Total
A3	8	7	0	15	8	11	0	19	8	9	2	19
B3	0	3	2	5	0	1	3	4	3	7	3	13
C3	0	0	1	1	0	0	1	1	1	0	1	2
Total	8	10	3	21	8	12	4	24	12	16	6	34

I. B. Design of the Test Instruments for English/Language Arts

Day One

Description: Open-ended writing prompt. Results in genre-specific writing sample.

Time Allotted: 45 Minutes

Scores Received: The entire response receives two scores: Writing; Conventions.

Clusters Represented: Scores map to two clusters: Writing; Conventions.

Standards Represented: The clusters map to two ELA standards: E2 Writing; E4 Conventions, Grammar, and Usage of the English Language. The possibility of mapping to E5 Literature, or at High School only to E6 Public Documents or E7 Functional Documents, exists as determined by the writing prompt.

Day Two

Description: Reading passage and prompts. Results in three short answers and one longer, text-based essay.

Time Allotted: 45 Minutes

Scores Received: All four answers receive two scores: Reading—Basic Understanding; Reading—Analysis and Interpretation.

The fourth answer, the text-based essay, receives one score: Text-Based Writing.

Clusters Represented: Scores map to three clusters: Reading—Basic Understanding; Reading—Analysis and Interpretation; Writing.

Standards Represented: The clusters map to two ELA Standards: E1 Reading; E2 Writing. The possibility of mapping to E5 Literature, or at High School only to E6 Public Documents or E7 Functional Documents, exists as determined by the reading passage.

Day Three

Description: Three reading passages and two editing passages. Each reading passage is followed by 5–8 multiple-choice items. Each editing passage is followed by 46 multiple-choice items.

Time Allotted: 45 Minutes.

Scores Received: Items scored by machine.

Clusters Represented: The items map to one of three clusters: Reading—Basic Understanding; Reading—Analysis and Interpretation; Conventions.

Standards Represented: The clusters map to two ELA Standards: E1 Reading; E4 Conventions, Grammar, and Usage of the English Language.

Appendix II. Characteristics of Judges Participating in Standard Setting

Participants

Summary of 1996 Standard-Setting Judges' Background Information

Variables	Levels of variables	Math ES (*N*=5)	Math MS (*N*=4)	Math HS (*N*=3)	Math % Total (*N*=12)	ELA ES (*N*=4)	ELA MS (*N*=6)	ELA HS (*N*=4)	ELA % Total (*N*=14)
Gender	Male	2	1	1	33%	0	2	2	28.6%
	Female	3	3	2	67%	4	4	2	71.4%
Ethnicity[a]	African-American	1	0	0	8%	1	1	2	28.6%
	Asian	0	1	0	8%	1	1	0	14.3%
	Filipino	0	0	0	0%	0	0	0	0.0%
	Hispanic	0	0	0	0%	0	1	0	7.1%
	Native American	0	0	0	0%	0	0	0	0.0%
	Pacific Islander	0	0	0	0%	0	0	0	0.0%
	White	4	3	3	83%	2	1	2	35.7%
	Other	0	0	0	0%	0	1	0	7.1%
Occupation[b]	Teacher	2	2	3	50%	2	2	2	37.5%
	School admin.	1	0	0	7%	2	2	1	31.3%
	University faculty	2	1	0	21%	0	0	1	6.3%
	Content specialist	2	1	0	21%	0	2	2	25.0%
Highest degree[a]	B.A. or B.S.	1	0	0	8%	1	0	0	7.1%
	M.A. or M.S.	2	2	2	50%	3	4	3	71.4%
	Ph.D.	2	2	0	33%	0	2	1	21.4%

[a] One participant did not answer these questions so the variables did not add up to 100%.

[b] Participants marked all that apply for these variables.

Appendix III. Instructions to Judges on Item Weighting

Principles of Weighting

Just as there are principles of standard-setting, there are principles to follow in developing weights for a decision rule:

1. The governing principle behind weighting of tasks or scores is the **centrality** of the task or score to the performance standards being assessed. Think about the information concerning a student's knowledge, skills, and abilities that would be yielded by the task. How central to the performance standard(s) targeted by this set of tasks is this information?
2. Weights should **not be determined by the relative difficulty of tasks**. Some things in the performance standards are difficult to learn, and some things are easy to learn. But if something is included in the performance standard, then it has been judged to be important to learn, regardless of its easiness or difficulty.
3. Weights that are applied to tasks or scores within clusters should be treated **within cluster** only. That is, how much a task or score is weighted within one cluster does not stand in relation to task or score weightings within other clusters. Clusters are treated separately.
4. **No one task or score should have too much weight.** Applying too much weight is equivalent to introducing a conjunctive element into the standard. (See standard-setting principle #3.)
5. Differences in weights should not be so small that they lose meaning for people trying to draw inferences about the exam. Also, the relationships between the weights should allow for simple comparisons between weights. The table below provides examples of reasonable and unreasonable weights applied to a hypothetical set of tasks:

	Reasonable *weights*	Unreasonable *weights*
Task 1	10%	9%
Task 2	30%	34%
Task 3	20%	17%
Task 4	20%	19%
Task 5	20%	21%
Total	100%	100%

Though both sets of weights sum to 100% and are fairly close for each of the tasks, there are two problems with the set of weights on the right. First, the distinctions between the weights applied to tasks 3, 4, and 5 are so small that they strain credulity. Is it really possible that the centrality of these tasks can be distinguished at such a fine level? It is important to provide weights that are plausible in the level of discernment between tasks.

The second problem with the set of weights on the right is that it is difficult to see the relationships between the weights. Task 3 is weighted .86 of Task 5, but .53 of Task 2. Besides the fact that this level of refinement is incredible, it is also confusing to consumers of the exam.

To avoid both these types of problems, it is a good idea to follow these guidelines when assigning weights:

- Think about what weight would be assigned to each task or score if the weights were all equal. On a five-task cluster, that would be 20%; on a three-task cluster that would be 33.3%. (Though 33.3 is a fractional number, it is easily understood as 1/3rd by most test users.)
- If equal weighting doesn't make sense, assign weights that are simple multiples of each other. Try sticking with multiples of 10. Or think of the 100 points in terms of halves, quarters, and eighths. Steer clear of weights that don't represent a simple fraction like 1/2, 1/3, 1/4, etc. Be careful of weights that end in digits that are not 0 or 5.

Appendix IV. Instructions to Judges on Scale Score Cut Points

Principles for Setting Cut Points

Cut points tell us which "bucket" a weighted average belongs in. It is important that judges have a shared understanding of some basic concepts as they set cut points:

1. The **performance standards are the driving principle behind cut points.** We are trying to align the weighted average with the level of accomplishment established in the performance standards.
2. The five buckets along the continuum of performance (Meets the standard with *H*onors, *M*eets the standard, *N*ear the standard, *P*rogressing to the standard, and *N*o *P*rogress to the standard) are ***not*** **on the same scale as the tasks.** It is merely a coincidence that many of the tasks have a score scale that has five points (0 through 4 for many tasks). That means that a weighted average of 3.0 does not necessarily "map" to the third bucket down. *It is up to you to use your professional judgment to make a decision about where the weighted averages should be mapped*. You may be lenient, stringent, or in between. But have a rationale for the mapping strategy you employ that is couched in your knowledge of the tasks and your understanding of the performance standards.
3. The five buckets are **not necessarily evenly spaced**. For example, you may decide that only the top three weighted averages belong in the "Honors" bucket, and that the next seventeen belong in the "Meets" bucket, and that the next thirty in the "Near" bucket. Or, you may decide that the buckets should be evenly distributed. It is up to you.
4. **You must not skip any buckets**. Each bucket must have at least one weighted average assigned to it.
5. The implication of #4, above, is that the highest weighted average must be assigned to the "Honors" bucket, and the lowest weighted average (0.0) must be assigned to the "No Progress" bucket.

Appendix V. Detailed Outline of "Standard Setting" Process

V. A. Mathematics: The following procedures are repeated, one cluster at a time, for the Skills, Concepts, and Problem Solving.

	Procedure	Data Collection	Feedback Provided
	Weight Task Scores		
1	Obtain initial score weights	Each judge's set of weights for each task score obtained within a cluster	
2	Provide feedback on initial score weights		• All judges' weights are entered on the overhead form; • Mean, range across judges by task score are entered; • Largest ranges across tasks are discussed first
3	Obtain consensus weights	*Consensus weights recorded for further analysis*	Consensus weights taken down and written on overhead
	Obtain Cut points		
4	Obtain initial cut points on weighted mean scale	Each judge's set of H, M, N, P, NP decisions for each point on the weighted mean scale	
5	Provide feedback on initial cut points		• Side-by-side comparison of judges' cut points; • Overhead has "heavy line" drawn to indicate "bucket thresholds"; • Judges draw in thresholds on their copies; • Discuss largest threshold differences first starting with H/M, ending with P/NP
6	Obtain consensus cut points	*Consensus cut points recorded for further analysis*	Consensus cut points taken down and written on overhead
	Present Impact Decision Rule on Selected Profiles		
7	Show decision rule results	*Retain judges' weight change/cut point change comments for later analysis by development team*	Includes consensus weights used for decision and bucket thresholds. • For 100 selected profiles, the following results are shown: Weighted mean • Bucket under decision rule • Bucket assigned by development team • Difference flag

V. B. English Language Arts: The following procedures are repeated, one cluster at a time, for the Writing, Reading Comprehension, Reading Interpretation, and Conventions scores.

	Procedure	Data Collection	Feedback Provided
	Weight Task Scores		
1	Obtain initial component weights	Each judge's set of weights for each component obtained within a cluster	
2	Provide feedback on initial score weights		• All judges' weights are entered on the overhead form; • Mean, range across judges by task score are entered; • Largest ranges across tasks are discussed first
3	Obtain consensus weights	*Consensus weights recorded for further analysis*	Consensus weights taken down and written on overhead
	Obtain Cut points		
4	Obtain initial cut points on weighted mean scale	Each judge's set of H, M, N, P, NP decisions for each point on the weighted mean scale	
5	Provide feedback on initial cut points		• Side-by-side comparison of judges' cut points; • Overhead has "heavy line" drawn to indicate "bucket thresholds"; • Judges draw in thresholds on their copies; • Discuss largest threshold differences first starting with H/M, ending with P/NP
6	Obtain consensus cut points	*Consensus cut points recorded for further analysis*	Consensus cut points taken down and written on overhead
	Present Mapping Table of Decision Rule and Obtain Revised Table		
7	Show decision rule results and provide feedback		Includes consensus weights used for decision and bucket results for all pairwise combination of component scores in *mapping tables*: • Table of buckets under judges' decision rule • Table of buckets assigned by development team • Table showing differences flag
8	Obtain revised consensus mapping table	*Retain judges' revised mapping table results and comments for later analysis by development team*	Consensus mapping table taken down and written on overhead

The Process We Will Use In This Standard-Setting Session

English Language Arts

I. Familiarize judges with the **goals and methods** of the standard-setting session

II. Familiarize judges with the **overall structure** of the exam—**how tasks, scores, and clusters link to the performance standards**

III. Set a **decision rule** for the TEXTUAL UNDERSTANDING cluster

A. Judges receive training on the TEXTUAL UNDERSTANDING **tasks and scoring** system.

B. Judges set **weights** for TEXTUAL UNDERSTANDING scores.

C. Judges set **cut points** for TEXTUAL UNDERSTANDING weighted averages.

D. Judges discuss **feedback** showing the impact of the TEXTUAL UNDERSTANDING decision rule on all possible score pairs and comparing results to those achieved with preliminary decision rule.

IV. Set a **decision rule** for the ANALYSIS & INTERPRETATION cluster

A. Judges receive training on the ANALYSIS & INTERPRETATION **tasks and scoring** system.

B. Judges set **weights** for ANALYSIS & INTERPRETATION scores.

C. Judges set **cut points** for ANALYSIS & INTERPRETATION weighted averages.

D. Judges discuss **feedback** showing the impact of the ANALYSIS & INTERPRETATION decision rule on all possible score pairs and comparing results to those achieved with preliminary decision rule.

V. Set a **decision rule** for the WRITING cluster

A. Judges receive training on the WRITING **tasks and scoring system.**

B. Judges set **weights** for WRITING scores.

C. Judges set **cut points** for WRITING weighted averages.

D. Judges discuss **feedback** showing the impact of the WRITING decision rule on all possible score pairs and comparing results to those achieved with preliminary decision rule.

VI. Set a **decision rule** for the CONVENTIONS cluster

A. Judges receive training on the CONVENTIONS **tasks and scoring system.**

B. Judges set **weights** for CONVENTIONS scores.

C. Judges set **cut points** for CONVENTIONS weighted averages.

D. Judges discuss feedback showing the impact of the CONVENTIONS decision rule on all possible score pairs and comparing results to those achieved with preliminary decision rule.

A More Detailed Look at the Steps of Setting a Decision Rule

The following steps are repeated, one cluster at a time, for the Textual Understanding, Interpretation & Analysis, Writing, and Conventions clusters.

A	**Training on Tasks and Scores**	Exam developers provide training to judges on the tasks and scores in the cluster. Developers show and discuss benchmark performances that illustrate each score point for each task.
B	**Set Weights for Tasks**	
	1 Obtain initial score weights	Each judge records his/her set of weights for each score within a cluster.
	2 Provide feedback on initial weights	All judges' weights are entered on an overhead form. Next, the mean and range across judges by score are entered. Those scores showing the largest ranges across judges are discussed first.
	3 Obtain consensus weights	Judges arrive at consensus weights. These are recorded on the overhead and in Excel for later analyses.
C	**Set Cut Points**	
	4 Obtain initial cut points on weighted mean scale	Each judge records his/her set of H, M, N, P, NP decisions for each point on the weighted average scale.
	5 Provide feedback on initial cut points	Judges' cut points are entered into an Excel spreadsheet. Side-by-side comparisons of the judges' cut points will be provided to judges and on an overhead. Overhead has "heavy line" drawn to indicate "bucket thresholds." Judges draw in thresholds on their copies. Discuss largest threshold differences first starting with H/M, ending with P/NP.
	6 Obtain consensus cut points	Judges arrive at consensus cut points. These are recorded on the overhead and in Excel for later analyses.

D Discuss Feedback Showing Impact of Decision Rule on All Possible Score Pairs Within the Cluster

7	**Present feedback to judges**	Feedback includes the judges' consensus weights and cut points. For all possible score pairs, four tables are shown: • The weighted average score for the pair, using judge weights; • The bucket assigned by judges' decision rule; • The bucket assigned by the preliminary decision rule; • The "bucket span" of different classifications, if any.
8	**Judges discuss feedback**	Judges discuss classifications that are problematic. Judges may "tweak" classifications without "jumping any buckets." Judges' changes to the classification table are recorded for later analysis by developers.

Appendix VI. Final Decision Rules

This document contains the final decision rules for the 1996 English Language Arts and Mathematics Reference Examinations.

When necessary, the following abbreviations of the names of the different categories are used:

H	Achieved the Standard with Honors
S	Achieved the Standard
N	Nearly Achieved the Standard
B	Below the Standard
L	Little Evidence of Achievement

There are four scores reported for **English Language Arts**:

- Writing
- Reading—Basic Understanding
- Reading—Interpretation and Analysis
- Conventions

In every case two scores from the exam are used to arrive at the scores listed above (i.e., the open-ended score of rhetorical effectiveness and the text-based writing score are used for the overall "writing" score). The two scores used for each specific score above are shown on the top row and the far left column of the English Language Arts matrices. These matrices show all the possible combinations of scores and their classification into the different "buckets."

For example, if an elementary student got scores of 2 on rhetorical effectiveness and 4 on the text-based writing, then the student would be classified into "N" bucket, that is, Nearly Achieved the Standards.

There are three scores reported in **Mathematics**:

- Skills
- Concepts
- Problem Solving

Each of these scores is informed by several tasks. The information presented under the mathematics section shows the "weights" given to each task. The weights are given based on how "central" the information gained from a particular task is to the standard(s) being assessed by that task. These weights are applied to the scores for each task. (Note that the weights are percents that add up to 100% for each set of weights.)

A weighted average is calculated for each group of student performances on the tasks informing a score. **This weighted average is rounded to the nearest tenth** (i.e., the nearest 0.1). Therefore, the cut points that follow the weights are the boundaries of the performance levels based on these "weighted averages."

For example, suppose a student received scores of **2, 3, 3, 4, 4, 2** across tasks in high school Problem Solving. The weights for each task in problem solving are given in the following table.

Task name	Final weight
Adding Odd Numbers	18
Marbles in the Medicine Cabinet	22
Paper Clips (rubric 1)	18
Two Situations for Percent 1 (rubric 1)	10
Two Situations for Percent 2 (rubric 1)	10
Mailing Costs Made Easy 1 (rubric 1)	22

A weighted average for these scores is computed as follows:

$\{(2 \times 18) + (3 \times 22) + (3 \times 18) + (4 \times 10) + (4 \times 10) + (2 \times 22)\} / 100 =$ **2.8**

Now, we have a weighted average of 2.8, look at the table of cut points below.

Category	Final cut point
Achieved the Standard with Honors	3.3
Achieved the Standard	2.5
Nearly Achieved the Standard	2.1
Below the Standard	1.2
Little Evidence of Achievement	0.0

In this table, the category "Achieved the Standard with Honors" would include weighted means of 3.3 and above, the category "Achieved the Standard" would include the weighted means from 2.5 through 3.2 inclusively, etc. Therefore, a student with a weighted average of 2.8 would be categorized as "Achieved the Standard."

English Language Arts

Elementary School Grades

Elementary: Writing

Text-Based Writing Score

Open-Ended Score (RE‡)	0	1	2	3	4	5
0	L	L	L	B	B	B
1	L	L	B	B	B	N
2	B	B	B	N	N	N
3	N	N	N	S	S	S
4	N	N	S	S	S	H
5	N	N	S	H	H	H

‡RE= Rhetorical effectiveness.

Elementary: Reading—Basic Understanding

Minimum Number of Correct Multiple-Choice Items From 16 Possible Items

TU‡	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	L	L	L	L	L	L	L	L	L	L	B	B	B	B	B	B	B
1	L	L	L	L	L	B	B	B	B	B	B	B	B	B	B	N	N
2	B	B	B	B	B	B	B	B	B	B	N	N	N	N	N	N	N
3	N	N	N	N	N	N	N	N	N	N	S	S	S	S	S	S	S
4	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
5	S	S	S	S	S	S	S	S	S	S	S	H	H	H	H	H	H

‡TU= Textual understanding.

Elementary: Reading—Interpretation and Analysis

Minimum Number of Correct Multiple-Choice Items From 13 Possible Items

RI‡	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	L	L	L	L	L	L	L	L	L	B	B	B	B	B
1	L	L	L	L	B	B	B	B	B	B	B	B	N	N
2	B	B	B	B	B	B	B	N	N	N	N	N	N	N
3	N	N	N	N	N	N	N	N	N	N	S	S	S	S
4	S	S	S	S	S	S	S	S	S	S	S	S	S	S
5	S	S	S	S	S	S	S	S	S	H	H	H	H	H

‡RI= Reading interpretation and analysis.

Elementary: Conventions

Minimum Number of Correct Multiple-Choice Items From 10 Possible Items

Conv.‡	0	1	2	3	4	5	6	7	8	9	10
0	L	L	L	L	L	B	B	B	B	B	B
1	L	L	L	L	B	B	B	B	B	B	B
2	B	B	B	B	B	N	N	N	N	N	N
3	NOT APPLICABLE										
4	N	N	N	N	N	S	S	S	S	S	S
5	S	S	S	S	S	S	S	H	H	H	H

‡Conv. = Conventions (Writing).

Middle School Grades

Middle School: Writing

Text-Based Writing Score

Open-Ended Score (RE‡)	0	1	2	3	4	5
0	L	L	L	B	B	N
1	L	L	B	B	N	N
2	B	B	B	N	N	N
3	N	N	N	S	S	S
4	S	S	S	S	S	H
5	S	S	S	H	H	H

‡RE= Rhetorical effectiveness.

Middle School: Reading—Basic Understanding

Minimum Number of Correct Multiple-Choice Items From 14 Possible Items

TU‡	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	L	L	L	L	L	L	L	L	L	L	B	B	B	B	B
1	L	L	L	L	L	L	L	B	B	B	B	B	B	B	N
2	L	B	B	B	B	B	B	B	B	B	N	N	N	N	N
3	B	B	B	B	N	N	N	N	N	N	S	S	S	S	S
4	N	N	N	N	N	N	N	S	S	S	S	S	S	S	S
5	N	N	N	N	S	S	S	S	S	S	S	H	H	H	H

‡TU= Textual understanding.

Middle School: Reading—Interpretation and Analysis

Minimum Number of Correct Multiple-Choice Items From 11 Possible Items

RI‡	0	1	2	3	4	5	6	7	8	9	10	11
0	L	L	L	L	L	L	L	L	L	B	B	B
1	L	L	B	B	B	B	B	B	B	B	B	B
2	B	B	B	B	B	B	B	N	N	N	N	N
3	N	N	N	N	N	N	N	N	S	S	S	S
4	N	N	S	S	S	S	S	S	S	S	S	S
5	S	S	S	S	S	H	H	H	H	H	H	H

‡RI= Reading interpretation and analysis.

Middle School: Conventions

Minimum Number of Correct Multiple-Choice Items
From 5 Possible Items

Conv.‡	0	1	2	3	4	5
0	L	L	L	L	L	L
1	L	L	L	B	B	B
2	B	B	B	N	N	N
3	NOT APPLICABLE					
4	N	N	N	S	S	S
5	S	S	S	S	H	H

‡Conv.= Conventions (Writing).

High School Grades

High School: Writing

Text-Based Writing Score

(RE‡)	0	1	2	3	4	5
0	L	L	L	B	B	N
1	L	L	B	B	N	N
2	L	B	B	N	N	S
3	B	B	N	N	S	S
4	B	N	N	S	S	H
5	N	N	S	S	H	H

‡RE= Rhetorical effectiveness.

High School: Reading—Basic Understanding

Minimum Number of Correct Multiple-Choice Items From 9 Possible Items

TU‡	0	1	2	3	4	5	6	7	8	9
0	L	L	L	L	L	L	L	L	L	B
1	L	L	L	L	L	B	B	B	B	B
2	B	B	B	B	B	N	N	N	N	N
3	B	B	B	N	N	N	N	S	S	S
4	N	N	N	N	S	S	S	S	S	S
5	S	S	S	S	S	S	H	H	H	H

‡TU= Textual understanding.

High School: Reading—Interpretation and Analysis

Minimum Number of Correct Multiple-Choice Items From 11 Possible Items

RI‡	0	1	2	3	4	5	6	7	8	9	10	11
0	L	L	L	L	L	L	L	L	L	L	B	B
1	L	L	L	L	L	B	B	B	B	B	B	B
2	B	B	B	B	B	B	B	B	B	B	N	N
3	B	B	B	B	B	N	N	N	N	N	S	S
4	N	N	N	N	N	S	S	S	S	S	S	S
5	S	S	S	S	S	S	S	S	S	H	H	H

‡RI= Reading interpretation and analysis.

High School: Conventions

Minimum Number of Correct Multiple-Choice Items From 10 Possible Items

Conv.‡	0	1	2	3	4	5	6	7	8	9	10
0	L	L	L	L	L	L	L	L	L	L	L
1	L	L	L	L	L	L	B	B	B	B	B
2	B	B	B	B	B	B	N	N	N	N	N
3	NOT APPLICABLE										
4	N	N	N	N	N	N	S	S	S	S	S
5	S	S	S	S	S	S	H	H	H	H	H

‡Conv.= Conventions (Writing).

Mathematics

Elementary School Grades

Elementary: Skills

Weights

Task name	Final weight
How Many Cups?	10
David's Discoveries	15
Student's Choices	10
Bike Rides	10
Irma's Wake Up	15
Camping Showers	15
To the Nearest Whole cm	15
100-Year-Old Brothers	10

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.7
Achieved the Standard	2.9
Nearly Achieved the Standard	2.1
Below the Standard	0.9
Little Evidence of Achievement	0.0

Elementary: Concepts

Weights

Task name	Final weight
Same Size, Same Shape	5
Gilberto's Cranes	5
Where Were They Born	10
Square Rug	15
What's the Rule?	15
Spinner Prediction	10
How Many Beads?	10
Car Trip (rubric 2)	10
Going to School	15
Barney's Sandwich Shop (rubric 2)	5

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.7
Achieved the Standard	2.9
Nearly Achieved the Standard	2.1
Below the Standard	1.1
Little Evidence of Achievement	0.0

Elementary: Problem Solving

Weights

Task name	Final weight
Car Trip (rubric 1)	25
Barney's Sandwich Shop (rubric 1)	35
House of Cards	40

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.5
Achieved the Standard	2.8
Nearly Achieved the Standard	2.2
Below the Standard	1.1
Little Evidence of Achievement	0.0

Middle School Grades

Middle School: Skills

Weights

Task name	Final weight
12-mile trip	8
How many girls?	16
Bulletin board border	12
Two toppings	16
Math-a-thon	16
Car travel	8
"Muscle city"	16
Buying a stereo	8

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.5
Achieved the Standard	2.8
Nearly Achieved the Standard	2.2
Below the Standard	1.5
Little Evidence of Achievement	0.0

Middle School: Concepts

Weights

Task name	Final weight
Pentagons perimeter	10
A square	5
Ordering Pizza #1	5
Ordering Pizza #2	5
Ordering Pizza #3	5
Folding a Cube #1	8
Folding a Cube #2	6
Probability on the Line	15
Leon's Phone Bill #1	5
Leon's Phone Bill #2	10
The Quilt	14
Foreign Language #1	12

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.6
Achieved the Standard	2.8
Nearly Achieved the Standard	2.2
Below the Standard	1.6
Little Evidence of Achievement	0.0

Middle School: Problem Solving

Weights

Task name	Final weight
Foreign Language #2	27
Which Game?	15
Tony's Walk	20
Truth in Advertising	38

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.6
Achieved the Standard	2.8
Nearly Achieved the Standard	2.1
Below the Standard	1.4
Little Evidence of Achievement	0.0

High School Grades

High School: Skills

Weights

Task name	Final weight
Wins and Losses	8.33
Parallel Lines	8.33
Oops!	8.33
Pick a Card	8.33
What Slope?	8.33
Movie Survey 1-a	8.33
Movie Survey 1-b	8.33
Movie Survey 1-c	8.33
Paper Clips (rubric 2)	8.33
Two situations for percent-1 (rubric 2)	8.33
Two situations for percent-2 (rubric 2)	8.33
Mailing Costs Made Easy-1 (rubric 2)	8.33

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.5
Achieved the Standard	2.4
Nearly Achieved the Standard	2.0
Below the Standard	1.1
Little Evidence of Achievement	0.0

High School: Concepts

Weights

Task name	Final weight
Numbers Between Numbers	7
Not a Right Angled Triangle	5
Pole Height	7
Adults and Teenagers	8
This is Always True	5
Movie Survey 2	8
Graphs of Rope (a)	8
Graphs of Rope (b)	7
Graphs of Rope (c)	6
Two Situations for Percent 3	6
Two Situations for Percent 4	5
Two Situations for Percent 5	5
Smoking 1	5
Smoking 2	7
Smoking 3	7
Smoking 4	4

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.3
Achieved the Standard	2.3
Nearly Achieved the Standard	1.8
Below the Standard	1.2
Little Evidence of Achievement	0.0

High School: Problem Solving

Weights

Task name	Final weight
Adding Odd Numbers	18
Marbles in the Medicine Cabinet	22
Paper Clips (rubric 1)	18
Two Situations for Percent 1 (rubric 1)	10
Two Situations for Percent 2 (rubric 1)	10
Mailing Costs Made Easy 1 (rubric 1)	22

Cut Points

Category	Final cut point
Achieved the Standard with Honors	3.3
Achieved the Standard	2.5
Nearly Achieved the Standard	2.1
Below the Standard	1.2
Little Evidence of Achievement	0.0

References

Cronbach, L., Linn, R. L., Brennen R. L., & Haertel, E. (1995, Summer). Generalizability analysis for educational assessments. *Evaluation Comment*, pp. 1-29.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education 7*, 323-342.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of educational measurement, 32*, 179-197.

Lord, F. M. (1965). A strong true score theory, with applications, *Psychometrika, 30*, 239-270.

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. *Military Psychology, 2*, 129-144.

New Standards. (1997). *Technical report on the 1996 reference examinations*. Pittsburgh, PA: Author