Estimating the Consistency and Accuracy of Classifications in a Standards-Referenced Assessment

CSE Technical Report 475

Michael James Young
Learning Research and Development Center,
University of Pittsburgh

and

Bokhee Yoon
Office of the President,
University of California

April 1998

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

Project 3.4 Dependability of Assessment Results. Lauren Resnick, Project Director, CRESST/ University of Pittsburgh; David E. Wiley, Project Dirctor, CRESST/Northwestern University

The work reported herein was supported under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

ESTIMATING CONSISTENCY AND ACCURACY OF CLASSIFICATIONS IN STANDARDS-REFERENCED ASSESSMENT[1]

Michael James Young
Learning Research and Development Center, University of Pittsburgh

Bokhee Yoon
Office of the President, University of California

Abstract

An important feature of recent large-scale performance assessments has been the reporting of pupil and school performance in terms of performance or proficiency categories. When an assessment uses such ordered categories as the primary means of reporting results, the natural way of reporting on the quality of the assessment is through the probabilities of consistent and correct classification of students. This paper applied a method introduced by Livingston and Lewis (1995) for calculating those probabilities. The use of this procedure to extend Lord's strong true score theory to tests containing other than multiple-choice items has created an important tool for test developers. The data used in the paper are from the New Standards Reference Examinations in Mathematics and English-Language Arts that were administered to students in Grades 4, 8, and 10 in spring 1996. The results of these analyses showed that for total score composites, the range of students accurately classified as having "met the standard" is from 85% to 98%, and the range of students consistently classified as having "met the standard" is from 77% to 96% across grades and content areas.

To prepare students for the challenges of the 21st century, a number of educators turned their attention to examinations that are referenced to performance standards. An important objective of these new examinations has been on redirecting instruction toward more challenging and appropriate learning goals. To this end, a higher proportion of tasks requiring students to construct their own responses to the questions being asked has appeared on these examinations. These assessment tasks are typically scored in a way that allows for partial credit, and student performance is classified on the basis of weighted or unweighted aggregates of the task scores. Finally, a determination is made on

[1] This a revised version of a paper presented at the 1997 meeting of the National Council on Measurement in Education in Chicago, IL.

whether a student is "up to the standard" as measured by his or her score on the examination. While the accuracy and consistency of students' scores in any examination are always important, of even greater concern is the accuracy and consistency of the *decisions* based on these scores.

The *accuracy* of a decision is the extent to which such a decision would agree with the decisions that would be made if each student could somehow be tested with all possible forms of the examination. The *consistency* of a decision is the extent to which such a decision would agree with the decisions that would have been made if the students had taken a different form of the examination than the one actually taken.

Most of the proposed methods for estimating the consistency or accuracy of such decisions have assumed that the test consists of equally weighted, dichotomously scored items with the total score equal to the number of correctly answered items (Hanson & Brennan, 1990; Huynh, 1976; Subkoviak, 1976). This paper applies a method introduced by Livingston and Lewis (1995) for estimating the accuracy and consistency of these decisions regardless of the scoring system used. 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. The conditional distribution of scores on an alternate form given the true score is estimated from a binomial distribution based on the estimated effective test length.

The 1996 Reference Examination Configuration

The data used in this study are taken from the spring 1996 administration of the New Standards Reference Examinations in Mathematics and English-Language Arts (ELA). Student performance was reported in three "clusters" or areas for Mathematics—Skills, Concepts, and Problem Solving—and in four clusters for English-Language Arts—Reading: Basic Understanding, Reading: Analysis and Interpretation, Writing, and Conventions.

The Mathematics examination consisted entirely of open-ended tasks that varied in length from 2 minutes for some of the tasks in the Skills cluster to 45 minutes for the tasks in the Problem-Solving cluster. The ELA examination integrated writing and reading tasks. The first day of the examination had students respond to a single writing prompt that was scored for rhetorical

effectiveness and use of conventions. On the second day, students' written responses to a text were scored using rubrics for writing, for students' understanding of the text, and on their analysis and interpretation of the text. The third day of the examination used multiple-choice items to assess the students' understandings and interpretations of several passages, and their mastery of written conventions and grammar in several editing passages. Table 1 summarizes the configurations of the 1996 Reference Examinations. The assignment of standards levels to students' examination results was done on the basis of their aggregated task scores within each cluster. Weights were assigned to tasks in each of the clusters, and the weighted averages were calculated. Sets of cutpoints on these weighted average scales were used to determine the standards

Table 1

1996 New Standards Reference Examination Configuration: Mathematics and English-Language Arts

Cluster	Numbers of scores produced		
	Elementary	Middle school	High school
Mathematics			
Skills	8 OE	8 OE	12 OE
Concepts	10 OE	12 OE	16 OE
Problem solving	3 OE	4 OE	6 OE
Mathematics composite	21 OE	24 OE	34 OE
English-language arts			
Reading: Basic understanding	17 (1 OE/16 MC)	15 (1 OE/14 MC)	10 (1 OE/9 MC)
Reading: Analysis and interpretation	14 (1 OE/13 MC)	12 (1 OE/11 MC)	12 (1 OE/11 MC)
Reading composite	31 (2 OE/29 MC)	27 (2 OE/25 MC)	22 (2 OE/20 MC)
Writing	2 OE	2 OE	2 OE
Conventions	11 (1 OE/10 MC)	6 (1 OE/5 MC)	11 (1 OE/10 MC)
Writing composite	13 (3 OE/10 MC)	8 (3 OE/5 MC)	13 (3 OE/10 MC)

Note. OE = open-ended tasks; MC = multiple-choice items.

levels that were reported for individual students. The five categories determined by the cutpoints were labeled:

- Achieved the Standard with Honors

- Achieved the Standard
- Nearly Achieved the Standard
- Below the Standard
- Little Evidence of Achievement

Further information on the processes of setting and assigning the standards levels can be found in the *1996 New Standards Reference Examination Technical Summary* (New Standards, 1997).

Analysis

Calculating the Reliability of Reference Examination Scores

The 1996 Reference Examinations consisted of tasks that varied substantially in both the rubrics used to score them and the length of time needed to answer them. The time needed for a student to respond to a task ranged from 2 minutes to the entire length of a examination session, and both multiple-choice and constructed response tasks could be found in the English-Language Arts examination. In addition, the New Standards Reference Examinations produced scores that were weighted averages of tasks or items for each cluster. When these factors of scoring rubric, length of time needed to respond, task weight, and task type are confounded, then tasks can vary in their relative contribution to an overall score: The tasks differ in their "functional lengths." In addition, when different task types measure unique rather than common skills, careful attention must be paid to the reliability coefficients used in order to represent adequately the part scores based on these different tasks (Qualls, 1995). The reliabilities calculated for New Standards examinations used a variety of procedures to address these factors. The formulas for these different coefficients are shown in Figure 1.

Coefficient Alpha:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_t^2}\right]$$

where, σ_i^2 = variance of weighted score on item *I*;

σ_t^2 = variance of total weighted score;

n = number of items.

Feldt-Raju reliability coefficient:

$$_{F-R}\rho_{tt'} = \frac{\sigma_t^2 - \sum\sigma_i^2}{(1 - \sum\hat{\lambda}_i^2)\sigma_t^2}$$

where, σ_i^2 = variance of weighted score on a composite *i* in a cluster (i.e., an open-ended item or multiple-choice item total);

σ_t^2 = variance of total weighted score;

$\lambda_i = \sigma_{it} / \sigma_t^2$ functional contribution of composite *I*.

Stratified Cronbach Alpha:

$$_{Strat}\rho_{tt'} = 1 - \frac{\sum\sigma_i^2(1-\rho_{ii'})}{\sigma_t^2}$$

where, σ_i^2 = variance of weighted score on Cluster *i* for Mathematics and item type for English-Language Arts;

σ_t^2 = variance of total weighted score;

σ_t^2 = reliability coefficient of weighted score on Cluster *i* or item type *i* (i.e., open-ended or multiple-choice).

Figure 1. Reliability formulas used in New Standards Reference Examinations.

In Mathematics, only open-ended response tasks were used, and results were reported for three clusters. Weights were assigned to each task within a cluster in Mathematics, and a weighted mean was calculated. Because the weights for individual items were fairly homogeneous within clusters, and the tasks were all open-ended, the reliability for each Mathematics cluster was estimated using Cronbach's Alpha on the weighted task scores. In order to

estimate the overall reliability of the Mathematics Reference Examination, a simple composite was formed by adding together the three weighted averages of the clusters. The reliability for this "total score" was estimated using Stratified Cronbach Alpha, where the tasks were stratified by cluster.

In English-Language Arts, various combinations of open-ended tasks and multiple-choice items were used, and this required a modification of the procedures above. Within each cluster, the weights of the open-ended and multiple-choice totals used in forming weighted averages were less homogeneous than for Mathematics. Therefore, tasks were stratified within the cluster as to whether they were open-ended or multiple-choice. However, because three of the four clusters reported in English-Language Arts combined a single open-ended task with a set of multiple-choice items in producing a weighted average, this produced strata containing only single tasks. Because the Feldt-Raju reliability coefficient can be used with strata consisting of single tasks, this coefficient rather than Stratified Cronbach Alpha was used to estimate the reliabilities for the clusters.

Composite scores were also calculated for English-Language Arts. The first composite was formed by adding together the weighted averages of the two Reading clusters; the second composite was formed by adding together the weighted averages for the Writing and Conventions clusters. This produced two composites that could be interpreted as "total scores" for reading and writing. All of the tasks within a composite were stratified as to whether they were open-ended or multiple-choice. Since these composites did not involve strata with single tasks, the reliabilities were estimated using Stratified Cronbach Alpha.

Defining Decision Consistency and Accuracy

A set of analyses were performed to estimate the accuracy and consistency of decisions based on standards levels. The *accuracy* of the decisions is the extent to which they would agree with the decisions that would be made if each student could somehow be tested with all possible forms of the examination. The *consistency* of the decisions is the extent to which they would agree with the decisions that would have been made if the students had taken a different form of the New Standards examination, equal in difficulty and covering the same content as the form they actually took. These ideas are shown schematically in Figures 2 and 3.

		Decision made on form actually taken	
		Below the Standard	Above the Standard
"True status" on the basis of the all forms average	Below the Standard	*Correct Classification*	*Misclassification*
	Above the Standard	*Misclassification*	*Correct Classification*

Figure 2. Classification accuracy.

		Decision made on the basis of the second form taken	
		Below the Standard	Above the Standard
Decision made on the basis of the first form taken	Below the Standard	*Consistent Classification*	*Inconsistent Classification*
	Above the Standard	*Inconsistent Classification*	*Consistent Classification*

Figure 3. Classification consistency.

In Figure 2, correct classifications occur when the decision made on the basis of the all-forms-average (or true score) agrees with the decision made on the basis of the form actually taken. Misclassifications occur when, for example, a student who is actually "Below the Standard" on the basis of his or her all-forms-average is classified incorrectly as being "Above the Standard." Consistent classifications occur (Figure 3) when two forms agree on the classification of a student as either being "Above the Standard" or "Below the Standard"; inconsistent classifications occur when the decisions made by the forms differ.

Estimating Decision Consistency and Accuracy

These analyses make use of the techniques outlined and implemented by Livingston and Lewis (1995) and Haertel (1996). Estimates of decision accuracy

and consistency were made for the "Achieved the Standard" cutpoint for each cluster reported in the Mathematics and English-Language Arts Reference Examinations. Additional analyses were performed to examine the decision accuracy and consistency of the composite "total scores" for Mathematics, Writing, and Reading. The "Meets the Standard" cutpoint for these composites was defined as the sum of the "Achieved the Standard" cutpoints of the clusters within the composite. The outline below lists the steps followed in each analysis.

Step 1: Estimate the effective test length (*n*)

For each test level and score (e.g., the Elementary Reading composite score) the *effective length* of the test was estimated. The estimate of the effective length of the test was based on the reliability of the cluster or composite score. The effective test length is defined as "the number of discrete, dichotomously scored, locally independent, equally difficult items required to produce a total score of the same reliability" (Livingston & Lewis, 1995, p. 186).

$$n = \frac{(\mu_X - X_{\min})(X_{\max} - \mu_X) - r\sigma_X^2}{\sigma_X^2(1-r)}$$

where r is the reliability coefficient, μ_X is the mean of the raw score, and σ_X^2 is the variance of the raw score. The lowest and highest possible scores are denoted by X_{min} and X_{max} respectively.

Step 2: Transform the original raw score

Given the effective test length n found in Step 1, the original raw score scale was transformed onto a new scale extending 0 to n by

$$X' = n\frac{X - X_{min}}{X - X_{max}} = np$$

where X' represents the transformed score, and X is the original raw score, rounded to the nearest integer. The lowest and highest possible scores are denoted by X_{min} and X_{max} respectively, and p represents the score on the 0 to 1 scale.

Step 3: Estimate the distribution of the proportional true scores (T_p)

The transformed observed scores (X') were used to estimate the distribution of proportional true scores T_p using Lord's (1965) strong true score theory. This theory assumes that the proportional true score distribution has the form of a four-parameter beta distribution with density

$$g\left(T_p \mid \alpha,\beta,a,b\right) = \frac{1}{Beta\left(\alpha+1,\beta+1\right)} \frac{\left(T_p - a\right)^{\alpha}\left(b - T_p\right)^{\beta}}{\left(b-a\right)^{\alpha+\beta+1}},$$

where *Beta* is the beta function. This formula can be obtained by taking a random variable having a (two-parameter) beta distribution on (0,1), with parameters (α+1) and (β+1), and transforming it linearly onto the interval (a, b), where $0 \leq a < b \leq 1$. The additional parameters a and b make the model more flexible, by allowing zero frequencies for extremely low or extremely high true-score levels (Hanson & Brennan, 1990).

The parameters of the distribution were estimated using Hanson's (1995) USmooth program. The program uses the transformed score distribution as its input and calculates the method-of-moments estimates as outlined in Hanson (1991). These estimated parameters were used to generate a discrete version of the distribution of proportional true scores, by dividing the (0,1) range into steps of .01 and estimating the proportion of the distribution at each level of T_p.

Step 4: Estimate the conditional and joint distributions of classifications (Decision Accuracy)

For each level of the proportional true score distribution generated in Step 3, a binomial distribution was generated with parameters n and p. Each of these binomial distributions represents the distribution of scores on a hypothetical test of n independent dichotomous items conditional on true score level of the test takers.

The cumulative probabilities of the binomial distributions (i.e., on the X' scale) were found for each true score level, and cutpoint on the original score scale for "Meets the standard" was transformed so that

$$\text{Original scale } (X) \text{ cutpoint } x^* \rightarrow \begin{cases} \text{Transformed scale } (X') \text{cutpoint } x'^* \\ \text{True score scale } (T_p) \text{cutpoint } t_p^* \end{cases}$$

Using the distribution of true scores generated from Step 3, and the cumulative distributions conditional on true scores, the *joint* distribution of the classifications based on the true scores and the test scores was estimated, and a table of joint probabilities was produced (see Figure 4).

	Observed Score (Form Taken)		
True Score (All-forms-average)	Below x'	Above x'	
Below t'_p	P_{11}	P_{12}	$P_{1.}$
Above t'_p	P_{21}	P_{22}	$P_{2.}$
	$P_{.1}$	$P_{.2}$	1

Figure 4. True vs. observed score probability matrix for classification accuracy decisions.

Here the overall correct classification probability is given by the sum of the probabilities of being correctly classified as being "Below the Standard" (P_{11}) and "Above the Standard" (P_{22}) by both their true *and* observed scores. For students whose true scores are "Below the Standard," the probability of correct classification is given by the conditional probability $P_{11}/P_{1.}$. Similarly, the probability of correct classification for students whose true scores are "Above the Standard" is given by the conditional probability $P_{22}/P_{2.}$.

Step 5: Estimate the joint distribution of classifications on another form of the test and on the form actually administered.

Given the 2 x 2 table above, the probabilities of correct classification for the "Achieved the Standard" cutpoint were calculated for the 1996 Reference Examination. A similar table using the x' cutpoint for both margins was used to calculate the probability of consistently classifying a student as having "Achieved the Standard" on two independent administrations of the examinations. These calculations were made by assuming the

conditional independence of alternate test forms given the true score. A table of these joint probabilities is shown in Figure 5.

		Form 2 Score	
Form 1 Score	Below x'	Above x'	
Below x'	P_{11}	P_{12}	$P_{1.}$
Above x'	P_{21}	P_{22}	$P_{2.}$
	$P_{.1}$	$P_{.2}$	1

Note: $P_{12} = P_{21}$.

Figure 5. Form 1 vs. Form 2 score probability matrix for classification consistency decisions.

The overall probability of consistent classification is given by the sum of the probabilities of being classified as being "Below the Standard" (P_{11}) and "Above the Standard" (P_{22}) by both forms of the test. (Note that the probabilities referred to in Figure 5 are different than those in Figure 4.)

Results

Reliability of the Cluster Scores

Table 2 presents reliability coefficients and standard errors of measurement for English-Language Arts and Mathematics examinations.

The reliability coefficients and standard errors of measurement are estimated based on the weighted mean scores. In English-Language Arts, the weighted mean scores ranged from 0 to 5 for the cluster scores and from 0 to 10 for the total composite scores. In Mathematics, the weighted mean scores ranged from 0 to 4 for cluster scores and from 0 to 12 for the total composite scores.

The reliability coefficients for the cluster scores ranged from .70 to .76 in Reading, from .54 to .71 in Writing, and from .66 to .90 in Mathematics across grades. The reliability coefficients for the total composite scores ranged from .97 to .98 in Reading, from .72 to .83 in Writing, and from .89 to .95 in Mathematics across grades. In Mathematics, the reliability coefficient was higher for Skills than for Concepts even though the Concepts cluster had more scorable parts than

Table 2

Reliability Coefficients and Standard Errors of Measurement for English-Language Arts and Mathematics Examinations

	Elementary school		Middle school		High school	
	Reliability	SEM	Reliability	SEM	Reliability	SEM
English-language arts						
Reading: Basic understanding[a]	0.71	0.47	0.71	0.65	0.76	0.59
Reading: Analysis & interpretation[a]	0.70	0.59	0.71	0.61	0.75	0.61
Reading composite[b]	0.97	0.33	0.97	0.40	0.98	0.37
Writing[a]	0.59	0.51	0.54	0.63	0.62	0.61
Convention[a]	0.61	0.62	0.63	0.68	0.71	0.64
Writing composite[b]	0.83	0.69	0.72	0.97	0.78	0.94
Mathematics						
Concepts[c]	0.73	0.35	0.74	0.35	0.84	0.33
Skills[c]	0.79	0.33	0.85	0.37	0.90	0.35
Problem solving[c]	0.70	0.32	0.66	0.28	0.81	0.32
Math composite[b]	0.89	0.71	0.90	0.77	0.95	0.56

Note. SEM = Standard Error of Measurement. English-language arts: N = 4,028 for elementary, N = 1,439 for middle, N = 889 for high school; mathematics: N = 14,816 for elementary, N = 11,178 for middle and N = 6,356 for high school.

[a] Reliability calculated using Feldt-Raju coefficient.

[b] Reliability calculated using Stratified Cronbach Alpha.

[c] Reliability calculated using Cronbach Alpha.

the Skills cluster. This is because tasks (items) in each cluster have different scoring rubrics and task weights that varied their functional lengths; therefore, a cluster with a larger number of items did not necessarily show a higher reliability coefficient.

Accuracy and Consistency of Standards Level Decisions

Table 3 reports the estimates of decision accuracy and consistency for the New Standards examinations with respect to the cutpoint "Meets the Standard" for cluster scores and total composite scores. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores.

Table 3

Estimated Accuracy and Consistency of Decisions (Percentage)

	Elementary school		Middle school		High school	
	Accuracy	Consistency	Accuracy	Consistency	Accuracy	Consistency
English-language arts						
Reading: Basic understanding	85	77	87	82	91	85
Reading: Analysis & interpretation	87	83	84	86	94	90
Reading composite	96	94	96	94	98	96
Writing	88	89	85	84	92	82
Convention	78	70	75	64	81	74
Writing composite	88	83	86	77	85	84
Mathematics						
Concepts	92	89	88	83	89	84
Skills	85	80	90	84	92	89
Problem solving	93	90	91	86	95	92
Mathematics composite	94	91	92	89	95	93

Decision consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate form.

For total composite scores, decision accuracy ranged from 96% to 98% in Reading, from 85% to 88% in Writing and from 92% to 95% in Mathematics across grades. Decision consistency ranged from 94% to 96 in Reading, from 77% to 84% in Writing, and from 89% to 93% in Mathematics across grades. Both decision accuracy and decision consistency for the total composites were the highest in Reading and the lowest in Writing.

For cluster scores, decision accuracy ranged from 84% to 94% in Reading, from 75% to 92% in Writing, and from 85% to 95% in Mathematics across grades. Decision consistency ranged from 77% to 90% in Reading, from 64% to 89% in Writing, and from 83% to 92% in Mathematics across grades. As expected, both decision accuracy and consistency for the total composites were higher than those for the cluster scores in Reading and Mathematics. In Writing, decision consistency for Writing cluster (89% in elementary; 84% in middle) was higher

than decision consistency for Writing composite in elementary (83%) and middle (77%) schools. Decision accuracy was higher for Writing cluster (92%) than for Writing composite (85%) in high school.

In comparing decision accuracy and consistency, notice that the agreement in consistency is higher than the agreement in accuracy in all estimates except for the Writing cluster in elementary school and the Reading cluster (Analysis & Interpretation) in middle school. This is because in decision consistency, both variables are random while only one of the two variables includes a random component in decision accuracy.

Discussion

Because the method described here is a relatively new one, a consensus on the required levels of decision consistency and accuracy has not yet emerged. Although "rules of thumb" exist for determining the level of reliability needed for a multiple-choice test, there has been very little guidance for choosing the levels of decision consistency and accuracy needed for educational assessments.

In general, one should follow the basic principle that the more important the educational decision to be made, the higher the consistency and accuracy rate should be. A test of minimum competency for high school graduation would clearly require much higher consistency and accuracy rates because of the importance of the decisions to be made.

The consistency and accuracy rates of those assessments that have used the Livingston and Lewis procedure can be consulted to provide some benchmarks. Table 4 presents results for the composite scores on the New Standards Reference Examinations in Mathematics together with the results of the 1986 Advanced Placement Calculus AB and BC composite scores. When comparing these results, it is important to remember that the examinations differ in their purposes, content tested, and intended examinees, their cutpoint scores, the numbers and kinds of items administered, and in the total testing time allowed. Given these caveats, the composite scores on New Standards Mathematics Examinations have decision consistency and accuracy rates at levels comparable to the results reported for this set of AP Calculus examinations.

Table 4

Comparison of New Standards (NS) and Advanced Placement (AP) Examinations by Number of Items, Testing Time, Reliability, and Decision Accuracy and Consistency

	No. tasks		Test time		Accuracy[b]	Consistency[c]
	MC	OE	(min)	Reliability[a]	(%)	(%)
NS mathematics composite						
Elementary		21	120	.89	94	91
Middle		24	120	.90	92	89
High school		34	120	.95	95	93
Advanced Placement composite						
Calculus AB (1986)	45	6	180	.93+	93	90
Calculus BC (1986)	45	6	180	.91+	92	88

Note. MC = multiple-choice; OE = New Standards (NS) open-ended tasks or AP "free response" tasks.

[a] Composite of MC and OE items for AP and OE items only for NS; AP reports range of reliabilities.

[b] Overall accuracy at: Near/Achieves the Standard cut for NS; at 3/4 grade cut for AP.

[c] NS consistency cuts same as for accuracy in 3.

Conclusion

An important feature of recent large-scale assessments, such as the state assessments of California and Kentucky, and the National Assessment of Educational Progress (NAEP), has been the reporting of pupil and school performance in terms of ordered categories. When an assessment uses performance or proficiency categories as the primary means of reporting results, the natural way of reporting on the quality of the assessment is through the probabilities of consistent and correct classification of students.

This paper applied one method for calculating those probabilities, the procedure introduced by Livingston and Lewis (1995). The use of the "effective *n*-size" to extend Lord's strong true score theory to tests containing other than multiple-choice items has created a important tool for test developers.

Other approaches to examining classification errors include the AP Reliability-of-Classification Procedure, which is a variant of the Livingston and Lewis procedure (College Entrance Examination Board, 1988, Appendix A), and

applications of classical test theory (Rogosa, 1994) and extensions of generalizability theory (Rogosa and Kupermintz, 1996). Our future research will examine these alternatives to the Livingston and Lewis procedure.

References

College Entrance Examination Board. (1988). *The College Board technical manual for the Advanced Placement program.* Author: New York.

Haertel, E. H. (1996). *Estimating the decision consistency from a single administration of a performance assessment battery. A report on the National Board of Professional Teaching Standards McGEN Assessment.* Palo Alto, CA: Stanford University.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (ACT Res. Rep. 91-5). Iowa City, IA: American College Testing.

Hanson, B. A. (1995). USmooth: A program for smoothing univariate test score distributions (Version 1.5) [Computer software]. Iowa City, IA: American College Testing.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Huynh, H. (1976). On the reliability of domain-referenced testing. *Journal of Educational Measurement, 13,* 253-264.

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.

New Standards. (1997). *1996 New Standards Reference Examination technical summary.* Pittsburgh, PA: University of Pittsburgh.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. Applied Measurement in Education, 8, 111-120.

Rogosa, D. R. (1994, April). *Misclassification in student performance categories. Appendix to CLAS Technical Report.* Monterey, CA: CTB/McGraw-Hill.

Rogosa, D. R., & Kupermintz, H. (1996, June). *Examples of performance of G-theory extensions for estimating error.* A paper presented at the CCSSO (Council of Chief State School Officers) Conference on Large-Scale Assessment, Phoenix, AZ.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265-276.