EXAMINING GUIDELINES FOR DEVELOPING ACCURATE PROFICIENCY LEVEL SCORES

Kadriye Ercikan

One attempt to make scores from large-scale assessments more interpretable has been to provide proficiency level scores to describe the meaning of student performance on tests. This study has examined the accuracy of Ercikan and Julian's (2002) guidelines for developing proficiency level scores and the classification accuracy of proficiency level scores from British Columbia's Foundation Skills Assessment tests. The guidelines were examined by comparing expected classification accuracies, based on these guidelines, to those estimated using a statistical procedure. The guidelines provided accurate expected classification accuracies to use in making decisions about assessment design.

Key words: proficiency level scores, classification accuracy, assessment design, reliability

L'une des façons utilisées pour faciliter l'interprétation des résultats d'épreuves communes a été de fournir des scores de rendement comparatifs en fonction de normes de référence. Dans cet article, les auteurs analysent la pertinence des directives d'Ercikan et Julian (2002) ayant trait à l'élaboration des scores de rendement et l'exactitude du classement des scores de rendement dans les tests d'évaluation des compétences fondamentales en Colombie-Britannique. L'analyse des directives a donné lieu à une comparaison entre l'exactitude du classement en fonction des directives et l'exactitude du classement obtenue par une méthode statistique. Les directives ont produit des classements exacts et conformes aux prévisions et peuvent servir dans les décisions à prendre au sujet de la conception des évaluations.

Mots clés : scores de rendement, exactitude du classement, conception de l'évaluation, fidélité

One attempt to make scores from large-scale assessments more interpretable has been to provide proficiency level scores to describe the meaning of student performance on tests. Proficiency level score reporting is widely used in national as well as provincial achievement tests such as the School Achievement Indicators Program (SAIP) and the British Columbia Foundation Skills Assessments (FSA). In these assessments, performance is represented by the classification of student performance to a number of proficiency levels determined by a standard setting process. These proficiency levels may have a set of labels such as Basic, Proficient, and Advanced, and descriptions of performance at each proficiency level. Once these scores are released, typical users, educators or policy makers do not question the accuracy of proficiency level scores. Yet these types of scores involve errors in classifying student performance to different levels, especially when the number of proficiency levels and how the classifications are obtained do not match the properties of the tests on which the scores are based. Misclassifications of student performance to different proficiency levels jeopardize the validity of inferences about achievement trends and the policy decisions these assessments are intended to inform.

Given the increased use of proficiency level classifications as important indicators of learning outcomes to describe student performance, it is important to examine the accuracy of these classifications. Classification accuracy refers to accuracy of decisions made based on test scores rather than the accuracy of scores. This notion of accuracy is typically interpreted as consistency of classifications based on the same or parallel tests. Several authors have discussed and demonstrated procedures to estimate accuracy or consistency of classifications based on test scores (Huynh, 1976; Livingston & Lewis, 1995; Livingston & Wingersky, 1979; Subkoviak, 1976; Swaminathan, Hambleton, & Algina, 1974; Traub, Haertel, & Shavelson, 1996; Wilcox, 1981). Previous research has shown that one factor that determines the accuracy of classifications is the measurement precision provided by the test (Hambleton & Slater, 1997; Livingston & Lewis, 1995; Traub & Rowley, 1980), particularly measurement precision at cut-score points. Specifically, measurement error near the cut-scores provides information about the likelihood of misclassification errors, i.e., false-positive and false-negative errors. One factor that affects classification accuracy is the distance between cut-scores. When the cut-scores are closer to each other, the likelihood of false-positive and false-negative misclassifications is higher. Higher numbers of proficiency levels typically result in cut-scores that are closer to each other than if a smaller number of proficiency levels were used. Therefore, the higher the number of proficiency levels, the higher the probability that students may be misclassified.

In practical large-scale assessment situations, procedures are available to estimate classification accuracy for proficiency scores based on a single test administration once the assessment results have been determined (Huynh, 1976, 1979; Livingston & Lewis, 1995; Subkoviak, 1976). However, for most assessment purposes, and especially for high stakes decision-making purposes, discovery of unreliable proficiency level scores after the completion of the assessment is problematic. Therefore, in the assessment design stage, guidelines are needed to answer questions, such as (a) Given the test length and reliability, and the desired level of classification accuracy, how many proficiency levels can be used for reporting assessment results?; (b) Given the test length and reliability, and for a specific number of proficiency levels, what type of classification accuracy can be expected?; (c) For a certain number of proficiency levels with an identified level of classification accuracy, what type of reliability, or test length, is needed?

Ercikan and Julian (2002) presented guidelines to answer these questions. The purpose of the present study is to examine the accuracy of these guidelines by comparing expected classification accuracy based on the guidelines to the estimated classification accuracy using a statistical method to estimate classification accuracy using a single test administration. The classification accuracy is estimated for a large-scale assessment, namely the British Columbia Foundation Skills Assessment (FSA) using Huynh's Beta-nomial classification accuracy estimation procedure (Huynh, 1979), and these estimates are compared to classification accuracies based on the Ercikan and Julian (2002) guidelines. In addition to providing results regarding the accuracy of the Ercikan and Julian guidelines, the estimation of classification accuracy for the FSA serves an additional purpose. Because the FSA is similar to many provincial assessments in Canada in terms of its scope and characteristics, the classification accuracies obtained for the FSA may provide information about the kinds of classification accuracies that may be expected from other assessments with similar test length and measurement accuracies.

ERCIKAN AND JULIAN GUIDELINES

Ercikan and Julian (2002) based their guidelines for test design on a simulation study. In this study, they examined classification accuracy as a function of three factors: measurement precision, number of proficiency levels, and score level. They examined, separately as well as jointly, the effects of each of these factors on classification accuracy by varying the levels of these factors and observing the effect on classification accuracy. They defined classification accuracy as the agreement of classifications based on true and observed scores. The agreement indicators p_0, per cent agreement across classification categories, and Cohen's κ (Cohen, 1960) were used as measures of agreement. The variation in measurement precision was provided by simulating observed and true scores, using parameters from ten tests whose reliabilities ranged from 0.70 to 0.93. The number of proficiency levels varied between two and five, and the analyses were repeated for two different sets of cut-scores. The results from this simulation study can be summarized as follows: Classification accuracy is affected by measurement precision, as would be expected, and decreases as the number of proficiency levels increases. For a given reliability level, the classification accuracy, as would be estimated by p_0 and κ, decreased on average by 10 per cent for an increase of one proficiency level, 20 per cent for an increase of two proficiency levels, and 20 per cent to 30 per cent for an increase of three proficiency levels. In addition, classification accuracy was more sensitive to measurement precision when larger numbers of proficiency levels were considered. In other words, change in classification accuracy with changes in reliability is greater when higher numbers of proficiency levels are considered.

The minimum required test reliabilities presented in Ercikan and Julian (2002) for a desired level of classification accuracy are summarized in Table 1 for two, three, four, and five proficiency levels. These guidelines suggest that when a reliability estimate of 0.85 may be sufficient for obtaining classification accuracy of 0.90 for two proficiency levels, a reliability estimate of 0.95 or higher would be needed for three proficiency levels. A classification accuracy of 0.90 would be highly unlikely for four or larger numbers of proficiency levels. To obtain a classification accuracy level of 0.80, tests with reliabilities of at least 0.70, 0.80, and 0.95 would be needed for two, three, and four proficiency levels respectively; if larger numbers of proficiency levels, such as four or five, are needed, more modest classification accuracies such as 0.50 to 0.70 should be expected even with reliabilities as high as 0.90.

Table 1
Required Minimum Reliability Estimates for the Desired Classification Accuracy for 2, 3, 4 and 5 Proficiency Levels

Desired Classification Accuracy (p_0)	Number of Proficiency levels			
	2	3	4	5
0.90	0.85	0.95	Not likely	Not likely
0.80	0.70	0.80	0.95	Not likely
0.70	-	-	0.80	0.90
0.60	-	-	0.70	0.75
0.50	-	-	-	0.70

VERIFICATION OF GUIDELINES

Using data from the FSA 2000 tests in this study, Ercikan examined the accuracy of the guidelines presented in Ercikan and Julian (2002) by comparing these guidelines to classification accuracy estimates using the Huynh's (1979) Beta-nomial procedure.

Foundation Skills Assessment (FSA) Tests

The FSA is part of the British Columbia provincial assessments. The performance on the FSA tests are reported in terms of three proficiency levels: Not Yet Within Expectations, Meeting Expectations, and Exceeding Expectations. Students who have not attained the "meets expectations" standard are considered to be "not yet within expectations." The top two proficiency levels were defined as follows:

> *Meets expectations.* The level of performance at which a student meets or exceeds the widely held expectations for the grade on this test. With no other information, this is the level below which a teacher would want to know more about the reasons for a student's low performance.
>
> *Exceeds expectations.* The level of a student's performance that is beyond that at which a teacher would say the student has fully met the expectations of the grade on this test. Students' performance would be considered excellent for the grade on this test. (British Columbia Ministry of Education, 2001, p. 23)

For this study, a representative 10 per cent sample of data for each of the grades 4, 7 and 10 from the Year 2000 assessment was obtained. Students who took the tests in French, ranging from 9 to 25 students for each grade, were eliminated from the sample because the properties of the tests may vary for this group. The numbers of students, means, and standard deviations for each test and other descriptive statistics are presented in Tables 2 and 3. The FSA tests contained both multiple-choice and constructed-response items and the maximum possible scores ranged from 48 to 56. These tests had moderate to high difficulty levels, with the per cent of maximum score ranging from 0.50 to 0.71. The coefficient-alpha reliability estimate was used to estimate the reliabilities of these tests, given both dichotomously and polytomously scored item types. The reliabilities of the scores ranged from 0.84 to 0.88.

Classification Accuracy

Two classification accuracy indices were used in the study, p_0 and κ. The most commonly used measure of classification accuracy is a simple measure of agreement, p_0, defined as the total proportion of examinees who were classified into the same proficiency level according to their true score and observed score across all possible proficiency levels.

Another commonly used classification accuracy indicator is Cohen's κ coefficient (Cohen, 1960). This statistic is similar to the proportion agreement p_0, except that it is corrected for the agreement that is due to chance. Neither of these classification indices distinguishes among different degrees of misclassifications such as misclassifying examinees by one proficiency level versus two proficiency levels. For the purposes of this study, all misclassifications are treated as equally important.

Table 2: Foundation Skills Assessment (FSA) 2000 Sample Data and Tests

Subject	Grade	# Items (Max score)	Sample size
Reading	4	39 (51)	4710
	7	44 (56)	4724
	10	43 (55)	4648
Numeracy	4	36 (48)	4705
	7	36 (48)	4685
	10	36 (48)	4737

Table 3: Descriptive Statistics Based on the Foundation Skills Assessment (FSA) 2000 Sample Data

Subject	Grade	Average % of max.	Cut-scores[1]	Mean	SD	Coefficient-α
Reading	4	67	27, 41 (232, 361)	33.86	8.79	0.86
	7	68	31, 48 (225, 425)	37.81	8.59	0.84
	10	71	32, 49 (230, 420)	38.62	9.06	0.87
Numeracy	4	52	17, 39 (239, 465)	25.12	9.73	0.87
	7	58	19, 41 (237, 473)	27.50	9.86	0.87
	10	50	17, 39 (218, 424)	24.10	9.88	0.88

[1] On raw-score scale (on scale-score scale)

Measurement Precision at Cut-score Points

Classification accuracy is closely related to measurement precision provided at the cut-score points. To examine the measurement precision provided at the two cut-score points in FSA, the six FSA tests were calibrated using an item response theory (IRT) based approach. The multiple-choice items were calibrated using the 3-Parameter Logistic (3PL) model (Lord, 1980) and the constructed-response items were calibrated using the 2-Parameter Partial Credit (2PPC) model (Yen, 1993). The estimations were conducted using PARDUX (Burket, 1991). The standard error of measurement (SEM) for each θ score was computed based on the item parameter estimates using FLUX (Burket, 1993).

Beta-nomial Procedure

Analyses focused on estimating classification accuracy for each of the six FSA tests using Huynh's Beta-nomial procedure and comparing these estimates to those classification accuracies that would be expected based on the Ercikan and Julian guidelines. The Beta-nomial procedure uses the mean and the standard deviation of raw scores, the reliability estimate, maximum possible score points, the number of proficiency levels, and the cut-scores based on the raw score scale to estimate classification accuracy estimates p_0 and κ. Raw scores are defined as the the sum of scores across all items. The reliability was estimated using Cronbach-α (Cronbach, 1951). Cut-scores are scores that are used for classifying examinees to different proficiency levels. The Huynh method assumes that the test scores on each test follow a Beta-nomial model. The classification accuracy indicators, p_0 and κ, are then computed using the Beta-nomial distribution. The Beta-nomial procedure was implemented using a DOS based software developed by Huynh (1979).

RESULTS

In this study, Ercikan used the Beta-nomial procedure (Huynh, 1979), which can be used to estimate classification accuracy based on a single test administration, to examine the reasonableness of the guidelines provided in Ercikan and Julian (2002). In addition, classification accuracy in the British Columbia Foundation Skills Assessment (FSA) for grades 4, 7, and 10 on reading and numeracy was examined. The sections below describe results of the analyses investigating properties of the FSA tests, classification accuracy of proficiency level scores from these tests, and compare estimates of classification accuracies to those that would be expected based on the Ercikan and Julian guidelines.

Cut-scores

Two cut-scores are associated with the three proficiency levels reported by the FSA. The cut-scores were originally set on the raw-score scale by the British Columbia Ministry of Education. However, analyses included IRT calibrations that allowed examining measurement precision at different score points. Therefore, the description here includes both the cut-scores in terms of raw score points as well as scale scores and measurement precision based on IRT calibrations.

After calibration using the 3PL and 2PPC models, for dichotomous and polytomous items, respectively, scales were created for each test. The θ scale was transformed to range between 0 and 600 by multiplying θ scores that ranged between -4.00 to +4.00 by 75, the desired standard deviation, and adding 300, the desired mean. The scale scores that corresponded to the cut-scores on the raw score scale were determined by using the test characteristic curves that map raw scores onto θ score scale, which in return can be converted to a scale score. One factor that affects classification accuracy is the distance between the cut-scores. When cut-scores are closer to each other, the classification accuracy is expected to be lower. As can be seen in Table 3, the difference between the two cut scores ranged from 14 (for reading grade 4) to 22 raw score points (for all numeracy tests). The shortest distance between the cut-scores on the FSA tests correspond to 1.6 standard deviation of raw scores on the reading grade-4 test and the largest was approximately 2.2 standard deviation of raw scores on the three numeracy tests. The scale score cut-score differences ranged from 129 (for reading grade 4) to 236 (for numeracy grade 7).

Measurement Precision at Cut-score Points

The SEM was calculated for each scale score point. Based on the IRT methodology, the SEM is on the θ scale. Using the test characteristic curves, the scale scores and their corresponding SEM values on the θ

scale for each cut-score point were determined and are presented in Table 4. The SEM values at the first cut-score point were similar for reading and numeracy tests. However, the SEM values at the second cut-score were considerably larger for reading tests than numeracy tests. In addition, the second set of cut-scores was on parts of the scale where measurement precision was lower for all tests except for the numeracy grade-10 test.

Table 4: Standard error of measurement at cut-score points

Subject	Grade	Cut-Scores	
		1	2
Reading	4	25	40
	7	27	54
	10	24	54
Numeracy	4	27	34
	7	30	35
	10	30	25

Expected Classification Accuracy for the FSA based on the Ercikan and Julian Guidelines

The Ercikan and Julian guidelines require two types of information about the assessment to determine the expected classification accuracy levels: the reliability of the tests and the number of desirable proficiency levels. The FSA reported individual student performances as well as group level performances using three proficiency levels (Not Yet Within Expectations, Meeting Expectations, and Exceeding Expectations). The reliability as estimated by coefficient-α ranged from 0.84 to 0.88. Using the reliability estimates and the number of proficiency levels in the FSA tests, the Ercikan and Julian guidelines were used to determine the expected proficiency levels for the six tests. Given that the six FSA tests had similar reliabilities ranging from 0.84 to 0.88, the expected classification accuracy ranges were determined to be the same for all six tests. These expected ranges of classification accuracy are presented in Table 5. The expected classification accuracy p_0 ranged from 0.80 to 0.90 and the expected classification accuracy κ ranged from 0.65 to 0.75. The expected classification accuracy has a wide range because of variability in where the cut-scores are placed on the score scale and the measurement precision associated with these cut-scores.

Table 5: Verification of Classification Accuracy Based on the Foundation Skills Assessment (FSA) 2000 Sample Data

Subject	Grade	Expected[1] p_0 (κ)	Estimated p_0 (κ)	Adjusted p_0 (κ)
Reading	4	0.80 – 0.90 (0.65-0.75)	0.75 (0.59)	0.77 (0.69)
	7	0.80 – 0.90 (0.65-0.75)	0.78 (0.56)	0.80 (0.66)
	10	0.80 – 0.90 (0.65-0.75)	0.79 (0.60)	0.81 (0.70)
Numeracy	4	0.80 – 0.90 (0.65-0.75)	0.83 (0.63)	0.85 (0.73)
	7	0.80 – 0.90 (0.65-0.75)	0.84 (0.64)	0.86 (0.74)
	10	0.80 – 0.90 (0.65-0.75)	0.83 (0.65)	0.85 (0.75)

[1] Based on Ercikan and Julian guidelines

Classification Accuracy Estimates for FSA tests

For the Beta-nomial model, the main assumption that scores be distributed Beta-nomially was verified by visual examination of the graphical display of the distribution of scores from the six FSA tests. These distributions indicated that the Beta-nomial distribution would be a reasonable assumption. The Beta-nomial procedure was applied to the six FSA tests and classification accuracies were estimated. The results are presented in Table 5. The estimated classification accuracy, p_0, ranged from 0.75 (for the reading grade-4 test) to 0.84 (for the numeracy grade-7 test). The κ estimates ranged from 0.56 (for the reading grade 8 test) to 0.65 (for the numeracy grade-10 test). Previous research on the Beta-nomial classification accuracy estimates indicated that p_0 estimates had -2% bias, and the κ estimates had –10% bias (Huynh & Saunders, 1980). These bias estimates mean that using the Beta-nomial procedure, on the average, p_0 would be estimated to be two per cent less than it actually is, κ would be estimated to be 10 per cent less than it actually is. Therefore, the classification estimates were corrected for these biases to get more accurate estimates. For example, the estimated p_0 for reading grade 4 increased from 0.75 to 0.77 after an adjustment for -2% bias. An adjustment for the -10% bias on estimated κ for reading grade 4 increased the estimate from 0.59 to 0.69. The adjusted estimates for p_0 and κ are presented in Table 5. These adjusted estimates ranged from 0.77 to 0.86, for p_0, and they ranged from 0.69 to 0.75, for κ. The reading tests had consistently lower classification accuracies than the numeracy tests. Although measurement precision at the first cut-score points were similar for all tests, numeracy tests had higher measurement precision at the second cut-score and they had greater distances between the cut-scores, which may have led to the higher classification accuracies for these tests.

Comparison of Estimated versus Expected Classification Accuracies

The expected classification accuracy ranges based on the Ercikan and Julian guidelines were compared to the estimated classification accuracies. The *estimated* p_0 was within the range of expected values for the numeracy tests. However, they were lower than the expected ranges for the reading tests. *Estimated* κ, on the other hand, was lower than those predicted by the guidelines for all tests. When the estimates were adjusted for the expected negative bias, by -2% for p_0 and by -10% for the κ estimates, all *estimated* p_0 and κ fell within the range of *expected* p_0 and κ, for all tests except for the grade 4 reading test. For this test, even though the *estimated* κ fell within the range of *expected* κ values, the *expected* p_0 value was approximately 3 per cent less than the lower bound of the range of *expected* p_0.

SUMMARY AND DISCUSSION

In this article, the author summarized the guidelines provided in the Ercikan and Julian (2002) regarding the classification accuracy of proficiency levels and examined the accuracy of these guidelines. The accuracy of the guidelines was evaluated by comparing the expected classification accuracy based on these guidelines to the estimated classification accuracy using Huynh's Beta-nomial classification accuracy estimation procedure. These comparisons were conducted using the FSA 2000 assessments as examples. The results of the estimation procedure showed that the FSA assessments had moderate classification accuracy levels that had p_0 ranging between 0.77 and 0.86.

The classification accuracies estimated based on the statistical procedure were all within the expected range of classification accuracies based on the guidelines. The only exception was the estimated p_0 for the reading grade-4 test which had an estimated p_0 that was 3 per cent less than the lower bound of the range of the expected p_0. The small inconsistency between the expected and estimated classification accuracy p_0 for this test may be due to potential error in the guidelines because they do not take the distance between cut-scores and measurement precision into account, as well as possible bias greater than -2 per cent in the statistical estimation procedure. Overall, the findings indicate that the guidelines provided by Ercikan and Julian (2002) are reasonable rule of thumb to follow at the planning stage of an assessment design, when test developers do not have data needed to estimate these classification accuracies.

The Ercikan and Julian guidelines are expected to inform decisions about number of proficiency levels to use in an assessment, expected level of classification accuracy for an assessment with predetermined number of proficiency levels, and test length for a desired level of classification accuracy and number of proficiency levels. To determine the number of proficiency levels for an assessment, assessment developers need first to decide the minimum classification accuracy that would be acceptable for the consumers of the assessment results, such as educators and policy makers. Deciding on a level of classification accuracy is not the same as deciding on an appropriate level of reliability for a test. The developers need to consider the acceptable level of misclassifications, both false-positive and false-negative, and the costs associated with such misclassifications. For example, when classification accuracy is expected to be 0.80-0.90, assessment developers need to consider the implications of misclassifying 20 per cent of students into a wrong proficiency level, as well as on decisions such as resource allocation and remediation programs. The desirable classification accuracy level can be combined with the information about the reliability of the test to determine the number of proficiency levels based on the guidelines. Once the number of proficiency levels is determined, where cut-scores are established on the score scale and how far apart the cut-scores are, will affect the actual classification accuracy of the proficiency level scores. To achieve optimal levels of classification accuracy, the cut-scores should be established on points of the score scale where measurement precision is maximized. They should also be set as far apart on the score scale as possible, in addition to considerations given to criteria that may include behavioural expectations regarding performance on different parts of the scale.

Similarly to determine the expected level of classification accuracy for an assessment with a predetermined number of proficiency levels, the main information needed is the measurement precision provided by the test. However, the further apart the cut-scores are from each other, the higher is the likelihood that the expected classification accuracy level will be close to the actual classification accuracy.

The Ercikan and Julian guidelines also provide information about the number of test items needed for a desired level of classification accuracy and number of proficiency levels. It is important to highlight that test items that contribute to measurement precision on parts of the scale that are likely to have the cut-scores should be prioritized in constructing tests. On parts of the scale with high levels of measurement precision, examinees at different ability levels are better discriminated and, therefore, are less likely to be misclassified.

The Ercikan and Julian guidelines were evaluated based on a classification estimation procedure that itself has some error associated with it. The classification accuracy, consistency of classifications, of examinees can be examined more validly using two test administrations of parallel tests or the same test. The next step in evaluating the Ercikan and Julian guidelines should focus on comparing the guidelines to classification consistency based on two test administrations.

ACKNOWLEDGEMENTS

I thank Dr. Kim Koh for her assistance in data analysis, Professor Huynh Huynh for providing the software for estimating classification accuracy estimates, and the British Columbia Ministry of Education for providing the FSA data.

REFERENCES

British Columbia Ministry of Education. (2001). Interpreting and communicating British Columbia Foundation Skills Assessment Results, 2000. Vancouver, BC: The Author

Burket, G. (1991). PARDUX [Computer program]. Unpublished. Monterey, CA: CTB/McGraw-Hill.

Burket, G. (1993). FLUX [Computer program]. An unpublished IRT estimation and graphic analysis program. Monterey, CA: CTB/McGraw-Hill.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297-334.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education,* 15, 269-294.

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education,* 10, 19-39.

Huynh, H. (1976). On the reliability of domain-referenced testing. *Journal of Educational Measurement,* 13, 253-359.

Huynh, H. (1979). Computation and inference for two reliability indices in mastery testing based on the Beta-nomial model. *Journal of Educational Statistics,* 4, 231-246.

Huynh, H., & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement,* 4, 351-358.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement,* 32, 179-198.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement,* 16, 247-260.

Lord, F. M. (1980). *Applications of item response Theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Swaminathan, H., Hambleton, R. H., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement,* 11, 263-267.

Traub, R. E., Haertel, E. H., & Shavelson, R. J. (1996, April). The effects of measurement error on the trustworthiness of examinee classification. Paper presented at the 1996 annual conference of the American Educational Research Association (Session 54.40), New York.

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement,* 4, 517-545.

Wilcox, R. R. (1981). A review of the Beta-binomial model and its extensions. *Journal of Educational Statistics,* 6, 3-32.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement,* 30, 187-213.