DOCUMENT RESUME

ED 442 861 TM 031 281

AUTHOR	Hoffman, R. Gene; Wise, Lauress L.
TITLE	Establishing the Reliability of Student Proficiency Classifications: The Accuracy of Observed Classifications.
INSTITUTION	Human Resources Research Organization, Alexandria, VA.
SPONS AGENCY	Kentucky State Dept. of Education, Frankfort.
PUB DATE	2000-04-25
NOTE	17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
CONTRACT	M-0003669
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement; *Classification; *Observation; Probability; Raw Scores; *Reliability; Test Theory; *True Scores
IDENTIFIERS	*Accuracy

ABSTRACT

Classical test theory is based on the concept of a true score for each examinee, defined as the expected or average score across an infinite number of repeated parallel tests. In most cases, there is only a score from a single administration of the test in question. The difference between this single observed score and the underlying true score is error. This paper focuses on accuracy as a function of particular observed scores, questioning whether a student's unknown true score is likely to be in the same category as the student's observed score. A limited set of test items was retrieved from a state-wide examination. Sixteen multiple-choice mathematics items for 3,000 students were scaled using the three-parameter logistic option. The primary conclusion from this study is that classification accuracy functions based on observed scores look quite different from accuracy functions based on true scores. For some of the observed scores, the most likely true score is an adjacent classification category. A further exploration considered how observed scores are placed on the true score scale and whether using the same cut-points for true and observed scores is the best approach. The overall conclusion is that there is no way, short of a perfectly reliable test, of simultaneously maximizing observed score classification accuracy and the accuracy with which overall population distributions are estimated. Nonetheless, observed score classification accuracy curves do provide information about individual observed scores that is quite useful. These curves also provide a way of illustrating the consequences of particular decisions about the scaling and equating of performance category subscores. An appendix contains a visual depiction of the probability computations from the study. (SLD)

NCME
April 25, 2000

Establishing the reliability of student proficiency classifications: The accuracy of observed classifications

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 2000

R. Gene Hoffman

Lauress L. Wise

This presentation is an elaboration of work initially conducted for The Kentucky Department of Education under contract M-0003669.

Human Resources Research Organization
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

Establishing the Reliability of Student Proficiency Classifications: The Accuracy of Observed Classifications

R. Gene Hoffman
Lauress L. Wise

Human Resources Research Organization

Standards-based testing, which assigns students to a small number of discrete performance categories, has become a popular mode of student assessment with the National Assessment of Educational Progress and particularly with state school accountability programs. In a variety of states, students' categorical scores are used to assess schools' performance, so analysis of the potential for student classification errors is important for both students and schools.

Classical test theory is based on the concept of a *true* score for each examinee, defined as the expected or average score across an infinite number of repeated parallel tests. In most cases, we have only a score from a single administration of the test in question. The difference between this single *observed* score and the underlying *true* score is error. In this report, we are concerned not just with the size of these errors, but with the impact of these errors on classifying students into performance categories. Livingston and Lewis (1995) introduced the concepts of (1) classification *consistency*, which is the likelihood that repeated assessment will yield the same classification, and (2) classification *accuracy*, which is the likelihood that classification from an observed score is in the same classification of the corresponding true score. Both consistency and accuracy are often shown as a function of the true score, that is, different consistency or accuracy values are estimated and plotted for different possible true scores. We would like to introduce a somewhat different perspective. A teacher or parent is presented a student's observed test score and may wonder about the likelihood that their student's true score is in a proficiency classification that is the same or different from his/her observed score In this paper, we focus on accuracy as a function of particular observed scores. Our question is whether a student's *unknown* true score is likely be in the same category as the student's observed score. This perspective is important because it expresses error in a meaningful way for individual students. For need of a distinguishing term, we will refer to our perspective as a question about *observed score classification accuracy.*

Observed scores are assumed to vary in lawful ways around theoretical true scores or around domain scores with the variation calculated as the standard error of measurement (σ_e). In traditional reliability and generalizability theory, σ_e is a simple function of reliability (r_{tt}) and total test variability (σ_T):

$$\sigma_e = \sigma_T \sqrt{(1-r_{tt})} \qquad (1)$$

Error bands around estimated scores often accompany reports of students' test scores. These error bands are typically based on an estimate of σ_e with the assumption that errors of measurement are normally distributed. This approach used to estimate error bands can also be used to estimate traditional (true score) classification accuracy functions. For any given true score, the conditional distribution of observed scores is modeled as a normal distribution and the proportion of that distribution falling within the achievement category of the true score is taken as the classification accuracy value for that true score. Note, however, that the distribution of true scores for a given observed score is **not** necessarily normal, it is typically skewed, with the most likely true score being closer to the overall mean than the observed score. As a consequence, this same approach cannot be used directly to estimate observed score classification accuracy.

Item Response Theory (IRT) is a common scaling method that produces estimates of standard errors of measurement that vary along the true ability scale (Lord & Novick, 1968; Feldt & Brennan, 1989). Standard errors of measurement are conditioned on student ability, tending to be smallest near the center of the distribution and increasing toward the extremes. Estimates of conditional standard errors of measurement, i.e., $\sigma(x|\theta)$, increase precision in understanding the relationship between estimated scores and true scores. They also create a complication for estimating classification accuracy. *Observed score accuracy* is based on the distribution of true scores around observed scores, i.e., $f(\theta|x)$, whereas $f(x|\theta)$ represents the opposite – the distribution of observed scores around true scores.

Analytic Approach

A relatively straightforward solution for estimating *observed score accuracy* is available by conceptualizing observed score accuracy as a probability problem and using Bayes' Theorem.

Bayes' Theorem, as applied to continuous variables, states that:

$$f(\theta|x_j) = \frac{f(\theta)P(x_j|\theta)}{P(x_j)}, \tag{2}$$

where $P(x_j) = \int P(x_j|\theta)\, f(\theta)\, d\theta$ and $f(\theta)$ is the density function for the distribution of true scale scores.

To simplify the computation in our proposed approach, true scores and observed scores are treated as discrete variables. Observed scores, based on a finite number of items, are necessarily discrete. When scoring is based on raw (number right) scores from n items, there are a fixed number of possible scores even though the raw scores may be mapped onto a more continuous scale. When more complex pattern scoring is used, the number of possible values is greater, but still finite. For true scores, a set of score intervals can be used to define discrete values. By using discrete values, the probability of different true scores, θ_i for any given observed score, x_j, can be rewritten as

$$P(\theta_i|x_j) = \frac{P(x_j|\theta_i)P(\theta_i)}{P(x_j|\theta_1)P(\theta_1)+P(x_j|\theta_2)P(\theta_2)+P(x_j|\theta_3)P(\theta_3)+\ldots+P(x_j|\theta_k)P(\theta_k)} \tag{3}$$

where x_j = observed scale score at level *j*, and θ_i = true ability at level *i*, with *k* levels of ability included in the analysis.

For each θ_i, the probability of obtaining a given x_j, denoted $P(x_j|\theta_i)$, can be calculated directly from the IRT item parameter estimates or estimated by using $\sigma(x|\theta_i)$ and assuming a normal distribution of errors.

$P(x_j|\theta_i)$ is estimated for each of the combinations of possible observed scores and true scores by first assuming that each discrete scale score includes a hypothetical range of scale values from half of the distance to the next lower possible value to half the distance to the next higher value, i.e.,

$$\text{Score Range for } x_j = \frac{x_j + x_{(j-1)}}{2} \text{ to } \frac{x_j + x_{(j+1)}}{2} \quad (4)$$

For each θ_i and its associated $\sigma(x|\theta_i)$, the cumulative probability of scores within the score range for x_j can be calculated to estimate $P(x_j|\theta_i)$, assuming a normal distribution of errors. For the lowest score level, $P(x_1|\theta_i)$ is calculated as the cumulative probability of $\frac{x_1 + x_2}{2}$, given θ_i and $\sigma(x|\theta_i)$. For the highest score level, $P(x_k|\theta_i)$ is calculated as 1 minus the cumulative probability of $\frac{x_{k-1} + x_k}{2}$, given θ_i and $\sigma(x|\theta_i)$.

Equation 3 also requires a distribution for θ, in the form of $P(\theta_i)$, for i = 1 to k. Note that the true score distribution is not the same as the observed score distribution so the probability of true scores in each interval cannot be estimated directly from observed probabilities. Instead, we assume that true scores are normally distributed with variance given by $\sigma_t^2 = \sigma_X^2 - \sigma_e^2$, where σ_e is the standard error from Equation 1 above. We can then estimate the proportion of this distribution falling in each discrete score range.

Observed score accuracy is the idea that students with a given observed score, x_j, could have a true score in a proficiency category that is the same as or different from the level that contains that x_j. Thus, for any observed score, we can construct the probability that the true score is in each of the category levels. These probabilities can be calculated as

$P(\theta$ is in proficiency category a, given $x_j)$

$$= \sum_{i=m}^{n} P(\theta_i|x_j) \quad (5)$$

where m represents the lowest value of θ for a, the target true score proficiency category, and n represents the highest level of θ for category a.

When the target proficiency category (as bounded by m and n) for possible true scores is the same as the category for x_j, then the sum of the conditional probabilities in Equation 5 gives the probability that a student's true score is in the same category that he/she has been assigned by his/her observed score. This, then, is the probability that the student's observed classification is correct. Likewise, when the target category for possible true scores is different from the category containing x_j, the probability results from Equation 5 estimates the chances that a student's true category is different from his/her assigned category. By appropriately summing probabilities, we can obtain estimates of the probabilities for true scores being in any higher category, in any lower category, in the category one above the assigned category, in the category one below the assigned category, etc.

In addition to student level accuracy, judgments about the classification efficiency of a testing program as a whole depend on a system level estimate of the proportion of all students expected to be classified in congruence with their (unknown) true scores. This can be calculated by weighting the results of Equation (5) for each score level with the proportion of the sample who receive that score, and then summing over all score levels. That is,

The proportion of all students expected to be accurately classified =

$$\sum_{j=a}^{b} \Big[\sum_{i=m}^{n} \Big(P(\theta_i|x_j) \ * \frac{Freq_j}{\text{Total of All Students}}\Big)\Big], \qquad (6)$$

when the category boundaries $a = m$ and $b = n$. Proportions of misclassification can be calculated by setting a and b and m and n to reference different categories.

Figures A and B in the appendix provide visual representations of Equations 3 through 6.

An illustration

To illustrate the computations described above, a limited set of test items was retrieved from a state-wide exam. Specifically, for 3000 students, 16 multiple-choice mathematics items were scaled using Multilog's (Thissen, 1991)three-parameter logistic (3PL) option. Table 1 shows the item parameter estimates for each of the 16 items. Students were then scored with Multilog, producing estimated thetas (observed scores) and standard errors of measurement for each student. In order to display the data, the range of the student scores was divided into 15 equally spaced scores. Standard errors for each of these 15 scores were estimated by a simple least squares cubic function predicting individual SEMs from polynomials of theta ($R^2 = .997$). Cronbach's alpha for these 16 items is .73.

In other work, that is not yet released, we have conducted similar analyses with operational test data, including multiple choice and constructed response item that had been scaled and scored with CTB's PARDUX and FLUX programs (Burkett, 1995). In this program, raw scores (total correct with 72 points possible) are computed and converted to scale scores using the inverse of the test characteristic function. The test characteristic function gives the expected raw score as a function of the true score. We also used the IRT model to compute exact probabilities for every possible raw score for each of the discrete true score values. The pattern of results was similar to those described below.

Table 1
3PL Item Parameters for Sample Data

Item Number	*a*	*b*	*c*
1	0.56	-0.47	0.27
2	1.44	0.00	0.21
3	1.15	1.14	0.24
4	0.41	-0.85	0.27
5	0.96	-0.55	0.15
6	0.62	-0.22	0.10
7	1.26	0.85	0.35
8	0.68	0.97	0.18
9	0.79	0.09	0.19
10	1.19	0.52	0.19
11	0.74	1.16	0.35
12	1.14	0.25	0.11
13	1.15	0.22	0.09
14	0.73	0.34	0.13
15	0.50	1.17	0.30
16	0.47	-0.71	0.16

Table 2 presents example estimates for Prob($x_j|\theta_i$), the probability of different observed scores for different ranges of true scores. Rows in the table represent true scores, θ. At each θ level, $\sigma(x|\theta_i)$ is presented. Columns in the table represent observed scale scores. Four proficiency categories were created with cut points arbitrarily set to represent relatively high standards at the top two categories. Bold numbers in the tables indicate probabilities for observed scores being in the same *score interval* as the true score. As expected, the highest probabilities in any row typically occurs when the observed scores matches θ. On the other hand, there is marked departure from that expectation in the extremes where σ(x|θ) is large and the observed distribution is truncated. In addition, none of the probabilities are particularly large. Note that these scale score probabilities do sum to 1.00 across each θ row.

Table 2
*Probability of Different Observed Scores for Given True Scores: Prob($x_j|\theta_i$)**

True Score		Observed Score x_j														
		Category 1						Category 2				Category 3			Category 4	
θ	σ(x|θ)	-4.0	-3.5	-3.0	-2.6	-2.1	-1.6	-1.1	-0.6	-0.1	0.3	0.8	1.3	1.8	2.3	2.8
-4.0	5.9	**0.52**	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.14
-3.5	4.5	0.48	**0.04**	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.09
-3.0	3.4	0.42	0.06	**0.06**	0.06	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.05
-2.6	2.5	0.32	0.07	0.08	**0.08**	0.08	0.07	0.06	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.02
-2.1	1.8	0.17	0.08	0.09	0.10	**0.11**	0.10	0.09	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01
-1.6	1.2	0.04	0.05	0.08	0.11	0.14	**0.15**	0.14	0.11	0.08	0.05	0.02	0.01	0.00	0.00	0.00
-1.1	0.9	0.00	0.00	0.02	0.06	0.12	0.19	**0.22**	0.19	0.12	0.06	0.02	0.00	0.00	0.00	0.00
-0.6	0.6	0.00	0.00	0.00	0.00	0.02	0.09	0.23	**0.31**	0.23	0.09	0.02	0.00	0.00	0.00	0.00
-0.1	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.24	**0.40**	0.24	0.05	0.00	0.00	0.00	0.00
0.3	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.24	**0.43**	0.24	0.04	0.00	0.00	0.00
0.8	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.24	**0.40**	0.24	0.05	0.00	0.00
1.3	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.24	**0.34**	0.24	0.08	0.01
1.8	0.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.10	0.22	**0.28**	0.22	0.14
2.3	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.12	0.20	**0.23**	0.38
2.8	1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.12	0.17	**0.60**

*i = rows of the table, and j = columns.

The usual classification accuracy function (Livingston & Lewis, 1995) can be derived from this table by summing the probabilities within the target (and nontarget) categories for each true score. Figure 1 plots values for this accuracy function.

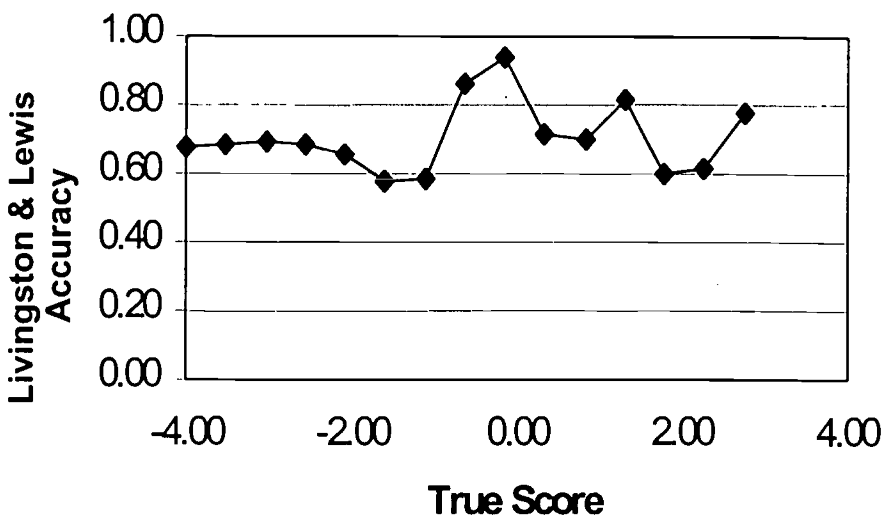

Figure 1. Classification accuracy as a function of true scores.

Table 3 presents Bayes' estimates for $P(\theta_i|x_j)$ with the θ probabilities summing to 1.00 for each possible observed scale value (i.e, each column). Shaded regions show areas of congruence between observed *score categories* and potential true *score categories*. Bold numbers indicate the probability that the unknown true score is in the same *score interval* as the observed score. Summing the $P(\theta_i|x_j)$ for the shaded area within each column (i.e. applying Equation 5) provides our estimate of observed score classification accuracy. These are presented in the last row of the table. Again, these accuracy values indicate probabilities that a student's unknown true score is in the same classification category as their known observed score. For example, a student with a observed score in the –0.6 score interval and classified at proficiency level 2 has only a 37% chance that his/her true score is in that same score interval, but has a 93% chance that his/her true score is also in proficiency level 2. Note that for students with a number of the observed score values (e.g., -2.1, 2.3), the chances that their true scores are in an adjacent category are greater than the chances are that their true scores are in the congruent category.

Table 3
Bayes' Estimates of the $P(\theta_i|ObS_j)$ for all combinations of θ_i and ObS_j and cumulative errors

True Score θ_i	Observed Scale Score, ObS_j														
	Category 1						Category 2				Category 3			Category 4	
	-4.0	-3.5	-3.0	-2.6	-2.1	-1.6	-1.1	-0.6	-0.1	0.3	0.8	1.3	1.8	2.3	2.8
-4.0	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3.5	**0.02**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3.0	**0.09**	**0.02**	**0.01**	**0.01**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
-2.6	**0.25**	**0.10**	**0.06**	**0.04**	**0.02**	**0.01**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
-2.1	**0.40**	**0.32**	**0.23**	**0.15**	**0.09**	**0.05**	0.03	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01
-1.6	**0.23**	**0.47**	**0.47**	**0.40**	**0.29**	**0.18**	0.10	0.05	0.02	0.01	0.01	0.01	0.01	0.00	0.00
-1.1	0.01	0.09	0.21	0.37	0.47	0.42	**0.29**	**0.15**	**0.07**	**0.03**	0.01	0.01	0.00	0.00	0.00
-0.6	0.00	0.00	0.00	0.02	0.12	0.31	**0.45**	**0.37**	**0.20**	**0.08**	0.02	0.00	0.00	0.00	0.00
-0.1	0.00	0.00	0.00	0.00	0.00	0.02	**0.13**	**0.35**	**0.41**	**0.24**	0.07	0.01	0.00	0.00	0.00
0.3	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.06**	**0.24**	**0.41**	0.31	0.09	0.01	0.00	0.00
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.18	**0.39**	0.40	0.17	0.03	0.00
1.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	**0.14**	**0.34**	0.46	0.31	0.07
1.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	**0.26**	0.42	0.34
2.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.07	**0.17**	**0.35**
2.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	**0.22**
Observed Score Accuracy	0.99	0.91	0.78	0.60	0.41	0.25	0.87	0.93	0.92	0.76	0.56	0.85	0.89	0.22	0.57

Figure 2 illustrates the observed score classification accuracy for each score level from the last row of Table 3. The effects of the cut points are clear from the dips in the plot, and again it is clear that for some scores that odds are less than 50-50 of the true score being in the same classification as the observed score.

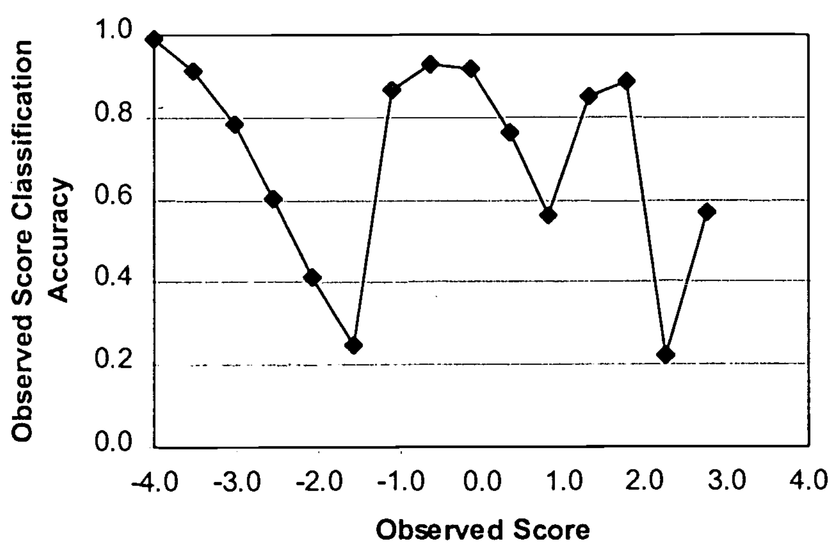

Figure 2. Probability that true achievement is in the same proficiency level assigned from observed test performance.

Table 4 shows true score classification probabilities for each observed score with congruent classifications in bold.

Table 4
Probability of true score being in any category given an observed score

Possible Category for θ	Observed Scale Score														
	Category 1						Category 2				Category 3			Category 4	
	-4.0	-3.5	-3.0	-2.6	-2.1	-1.6	-1.1	-0.6	-0.1	0.3	0.8	1.3	1.8	2.3	2.8
Category 1	**0.99**	**0.91**	**0.78**	**0.60**	**0.41**	**0.25**	0.13	0.07	0.03	0.02	0.02	0.02	0.02	0.02	0.03
Category 2	0.01	0.09	0.22	0.40	0.59	0.75	**0.87**	**0.93**	**0.92**	**0.76**	0.42	0.11	0.01	0.00	0.00
Category 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.22	**0.56**	**0.85**	**0.89**	0.76	0.40
Category 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.09	**0.22**	**0.57**
Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: Bold numbers indicate probabilities that true score is in the same classification as the assigned score

Equation 6 was then applied to the data in Table 3 to obtain expected proportions of students accurately classified. Tables 5 and 6 present the results in two steps. In Table 5, we present proportions by assigned category. In Table 6, we present the proportions across all students. Marginal values in Table 6 show the proportion of students actually assigned to each level, and the proportions of students expected to fall in each category. Note that the proportion of students in the extreme categories is overestimated when observed scores are used in place of true scores (.11 versus .08 and .04 versus in the lowest category and .02 in the highest category).

The sum of the bold values in Table 6 indicate total system accuracy. In this illustration, approximately 76% of the students would be expected to have true classifications equivalent to their observed classifications. In looking at classification accuracy for different observed scores, the most striking result is that students classified in the extreme categories (1 and 4) are more likely to have true scores in adjacent categories.

Table 5
Expected Proportions of Students within Each Assigned Score Categories who Would be Expected to have True Scores in Each Category

Possible Category for True Score	Category for Assigned Score			
	Category 1	Category 2	Category 3	Category 4
Category 1	**0.46**	0.05	0.02	0.02
Category 2	0.54	**0.86**	0.26	0.00
Category 3	0.00	0.08	**0.70**	0.64
Category 4	0.00	0.00	0.03	**0.34**
Total	1.00	1.00	1.00	1.00

Note: Bold numbers indicate the proportion of students within each category whose true score would be expected to fall in the same category as their observed score.

Table 6
Expected Proportions of Students Across All Categories who Would be Expected to have True Scores in Each Category

Possible Category for True Score	Category for Assigned Score				Expected true proportion of students in each category
	Category 1	Category 2	Category 3	Category 4	
Category 1	**0.05**	0.03	0.00	0.00	0.08
Category 2	0.06	**0.51**	0.07	0.00	0.64
Category 3	0.00	0.05	**0.19**	0.02	0.26
Category 4	0.00	0.00	0.01	**0.01**	0.02
Total assigned in each category	0.11	0.59	0.27	0.04	1.00

Note: Bold numbers indicate the proportion of all students whose true score would be expected to fall in the same category as their observed score.

Discussion and Further Exploration

The primary conclusion of this paper is that classification accuracy functions based on observed scores (Figure 2 in the example) look quite different from accuracy functions based on true scores (Figure 1). The pattern of results is initially surprising in that for some of the observed scores, the most likely true score is in an adjacent classification category. Clearly as numerous states and districts consider and implement high-stakes tests for students, this result is of concern. This finding led us to a further exploration of how observed scores are placed on the true score scale and/or whether using the same cut-points for true and observed scores is the best approach.

Because of error, observed score variance is greater than true score variance. We assumed a true score standard deviation of 1 in the IRT estimation, and the standard deviation of the observed scores (estimated thetas) was 1.15, consistent with a reliability estimate of .75. So, a person with a true score one standard deviation above the mean would have a true score of 1.0, while a person with an observed score one standard deviation above the mean would be at 1.15. One alternative to the procedure for assigning scale scores used above, would be to divide all of the observed scores by 1.15 so that the observed score variance was the same as the true score variance. The result,

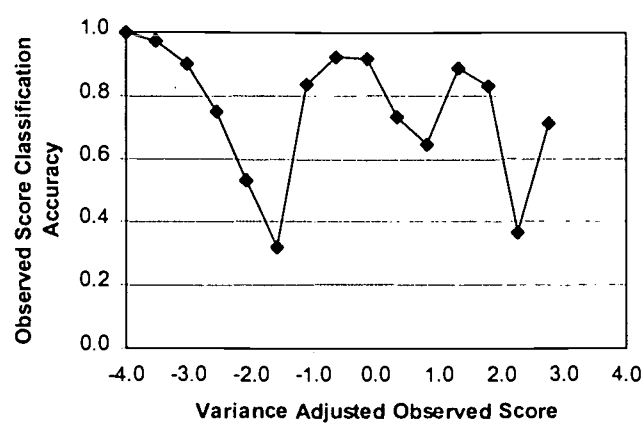

Figure 3. Classification accuracy of variance adjusted observed scores.

shown in Figure 3, reduces but does not eliminate the lower than 50-50 odds of (modified) observed score classification accuracy near the cut points.

Another solution is to use the results in Table 2 to adjust cut scores. It is important to recognize that, depending on standard setting method, cut scores may be set in the observed score metric or on the true score metric. For example, the Bookmark approach which orders items on IRT parameters appears to set standards in the true score metric, while an approach like Contrasting Groups uses the observed score metric. For illustration, we adjusted boundaries on the observed score scale so that for each observed score the most likely true score is always included in the congruent proficiency category. In other words, in Table 2, the observed score boundary between Categories 1 and 2 was moved to bisect the observed scores of –2.6 and –2.1. As a result, the observed score –2.1 is now is Performance Category 2 which is congruent with its the most probable true score (-1.1 at 47%). Likewise, the boundary between Categories 3 and 4 was moved one column to the right. The results of this adjustment are presented in Figure 4 where we see that the lowest accuracy value is now above 50%. With these boundary adjustment, all students are most likely to have a true score in the category congruent with their observed scores than in another category.

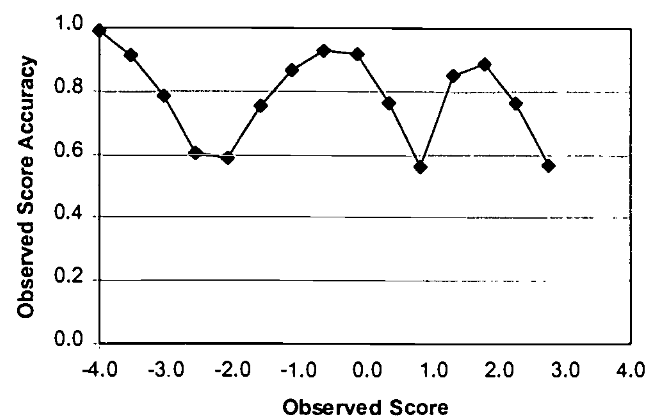

Figure 4. Accuracy of assigned classification after cut points adjusted.

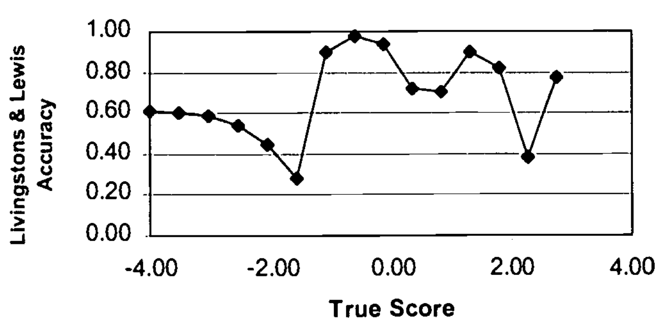

Figure 5. True score accuracy after cut point boundaries are shifted on the observed score scale.

Figure 5 shows the results of from the perspective of Livingston and Lewis's classification accuracy after the cutpoints were adjusted. By adjusting observed score boundaries to improve observed score classification accuracy, we reduce true score classification accuracy near the cutpoints (compare Figures 1 and 4). Another key difference associated with moving the cut-points is in the estimates of the proportion of students in each performance category. Table 7 shows the classification matrix along with the marginal proportion of students in

12

each achievement category after the observed score cutpoints were moved. In improving classification accuracy for individual scores, we are reducing the accuracy with which the overall proportion of students at each level is estimated.

Table 7
Expected Proportions of Students Across All Categories who Would be Expected to have True Scores in Each Category After Observed Score Boundaries are Adjusted

Possible Category for True Score	Category for Assigned Score				Expected true proportion of students in each category
	Category 1	Category 2	Category 3	Category 4	
Category 1	0.026	0.054	0.005	0.000	0.085
Category 2	0.008	0.558	0.070	0.000	0.636
Category 3	0.000	0.048	0.208	0.005	0.261
Category 4	0.000	0.000	0.004	0.007	0.012
Total assigned in each category	0.034	0.661	0.295	0.012	1.002

Note: Bold numbers indicate the proportion of all students whose true score would be expected to fall in the same category as their observed score.

Our overall conclusion is that there is no way, short of a perfectly reliable test, of simultaneously maximizing observed score classification accuracy and the accuracy with which overall population distributions are estimated. Nonetheless, observed score classification accuracy curves do provide information about individual observed scores that is quite useful. Further, these curves provide a way of illustrating the consequences of particular decisions about the scaling and equating of performance category cutscores.

References

Burkett, G. R. (1995). *PARDUX.* Monterey, CA: CTB/MCGraw-Hill.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement (3rd edition).* New York: American Council on Education and Macmillan.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2), 179-197.*

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Thissen, D. (1991). *Multilog™ User's Guide.* Lincolnwood, IL: Scientific Software.

Appendix

Visual depiction of probability computations follows on next two pages.

<table>
<tr><th colspan="13">Matrix 1: Probabilities of Observed Scores Given True Scores (all K x k cells would be filled)</th></tr>
<tr><th rowspan="2">True Score</th><th colspan="11">Observed Score</th><th rowspan="2">Sum</th></tr>
<tr><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>f</th><th>g</th><th>h</th><th>i</th><th>j</th><th>k</th></tr>
<tr><td>A</td><td>p(a|A)</td><td>p(b|A)</td><td>p(c|A)</td><td>p(d|A)</td><td>p(e|A)</td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>B</td><td>p(a|B)</td><td>p(b|B)</td><td>p(c|B)</td><td>p(d|B)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>C</td><td>p(a|C)</td><td>p(b|C)</td><td>p(c|C)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>D</td><td>p(a|D)</td><td>p(b|D)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>E</td><td>p(a|E)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>F</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>G</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>H</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1</td></tr>
<tr><td>I</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(k|I)</td><td>1</td></tr>
<tr><td>J</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(j|J)</td><td>p(k|J)</td><td>1</td></tr>
<tr><td>K</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(i|K)</td><td>p(j|K)</td><td>p(k|K)</td><td>1</td></tr>
</table>

Each true score, A – K, has a standard error of measurement which is used to calculate the probability of each reported score, a – k.

Apply Bayes theorem and estimate of true ability distribution to transform probabilities

<table>
<tr><th colspan="12">Matrix 2: Probabilities of True Scores Given Observed Scores (all K x k cells would be filled)</th></tr>
<tr><th rowspan="2">True Score</th><th colspan="11">Observed Score</th></tr>
<tr><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>f</th><th>g</th><th>h</th><th>i</th><th>j</th><th>k</th></tr>
<tr><td>A</td><td>p(A|a)</td><td>p(A|b)</td><td>p(A|c)</td><td>p(A|d)</td><td>p(A|e)</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>B</td><td>p(B|a)</td><td>p(B|b)</td><td>p(B|c)</td><td>p(B|d)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>C</td><td>p(C|a)</td><td>p(C|b)</td><td>p(C|c)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>D</td><td>p(D|a)</td><td>p(D|b)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>E</td><td>p(E|a)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>F</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>G</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>H</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(H|k)</td></tr>
<tr><td>I</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(I|i)</td><td>p(I|j)</td><td>p(I|k)</td></tr>
<tr><td>J</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(J|i)</td><td>p(J|j)</td><td>p(J|k)</td></tr>
<tr><td>K</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>p(K|h)</td><td>p(K|i)</td><td>p(K|j)</td><td>p(K|k)</td></tr>
<tr><td>Sum</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
</table>

Figure A. Construction of matrix with $P(x_j|\theta_i)$ and transformation into matrix for $P(\theta_i|x_j)$.

Partition cells by the proficiency level cut points and sum the conditional probabilities.

Matrix 2: Probabilities of True Scores Given Observed Scores, *Partitioned by Achievement Levels*											
True Score	Observed Score										
	Category 1			Category 2				Category 3		Category 4	
	a	b	c	d	e	f	g	h	i	j	k
A											
B	$\sum_{i=A}^{C} P(\theta_i \mid x_a)$			Shaded areas indicate misclassification probabilities							
C											
D											
E	$\sum_{i=D}^{G} P(\theta_i \mid x_a)$			The unshaded areas give probabilities of accurate classifications							
F											
G											
H											
I											
J											
K											
Sum	1	1	1	1	1	1	1	1	1	1	1

Matrix 2: Probabilities of True Scores Given Observed Scores, *Partitioned by True and Observed Achievement Levels*				
Potential True Score	Observed Score Category			
	1	2	3	4
Category 1	$\sum_{j=a}^{c} [\sum_{i=A}^{C} (P(\theta_i \mid x_j) * \frac{Freq_j}{\text{Total of All Students}})]$	$\sum_{j=d}^{g} [\sum_{i=A}^{C} (P(\theta_i \mid x_j) * \frac{Freq_j}{\text{Total of All Students}})]$		
Category 2	Shaded = Proportion incorrectly classified	Unshaded = Proportion correctly classified		
Category 3				
Category 4				

Figure B. Transition to accuracy of classification for each assigned score level and overall proportion of correct classifications.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Establishing the reliability of student proficiency classifications: The accuracy of observed classifications	
Author(s): R. Gene Hoffman & Lauress L. Wise	
Corporate Source: Human Resources Research Organization	Publication Date: April 25, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: [signature]	Printed Name/Position/Title: R. Gene Hoffman, Center Manager	
Organization/Address: HumRRO 295 W Lincoln Trail Blvd Radcliff, KY 40160	Telephone: 270 351-6088	FAX: 270-351-3620
	E-Mail Address: ghoffman@ne.infi.net	Date: 5/20/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)