Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 14, March, 2001 ISSN=1531-7714

Computing the Expected Proportions of Misclassified Examinees

Lawrence M. Rudner
LMP Associates & the Maryland Assessment Research Center for Education Success

Unless a test is perfectly reliable and thus contains no error, every time we make a classification based on a test score, we should expect some number of misclassifications. We can expect that some examinees whose true ability is above the cut score will be incorrectly classified as non-masters (false negatives) and that some number of low-ability examinees will be incorrectly classified as masters (false positives).

This paper provides and illustrates a method to compute the expected number of misclassifications. This information can help policy makers decide whether the risks are sufficiently small or whether the costs for improvements are justified. One particularly useful application of this procedure is to estimate the probability of a true master consistently failing multiple test administrations. Another is to examine the impact of cut score adjustments. Web-based software to apply this method is available at http://pareonline.net/misclass/class.asp. PC based software is available from the author.

Approach

I will develop the procedure using three-parameter item response theory and two state classifications (mastery and non-mastery). There are classical test theory analogs and the logic can easily be extended to more categories. I start with a test that maps individual scores (θ_i's) onto a continuous scale (a θ scale) and a cut score on that scale (θ_c) used to classify examinees into one of two discrete categories. Examinees whose scores are above the cut score will be classified as masters; those below as non-masters.

We make a distinction between the categories examinees should be placed in based on their true scores and the categories they are placed in based on observed score. The goal of this paper is to create, and then analyze, a two by two classification table, such as Table 1, indicating the expected proportions of correct and incorrect classifications.

Table 1: Classification table	
classified master true master	classified non-master true master
classified master true non-master	classified non-master true non-master

In table 1, the upper left and lower right quadrants represent correct classifications; the other two quadrants represent incorrect classifications.

The expected proportion of all examinee classified as a masters that are true non-masters is:

$$\mathrm{P(cm,n)} = \sum_{\theta_i < \theta_c} P(\hat{\theta} > \theta_c | \theta_i) f(\theta_i) / n \tag{1}$$

It should be noted that the phrases *true masters* and *true non-masters* are statistical terms meaning that the true ability is above or below the arbitrarily set cut score. In (1), $P(\hat{\theta} > \theta_c \mid \theta_i)$ is the probability of having an observed score, $\hat{\theta}$, above the cut score given a true score equal to θ_i, $f(\theta_i)$ is the expected number of people whose true score is θ_i and n is the total number of examinees. Thus $P(\hat{\theta} > \theta_c \mid \theta_i) f(\theta_i)$ is the expected number of people whose true score is θ_i that will be classified as masters, i.e., will have an observed score greater than the cut score. We sum this value over all examinees whose true score is less than the cut score and divide by n to obtain the probability of being misclassified as a master (false positive).

Similarly, the expected proportion of false negative is:

$$\mathrm{P(cn,m)} = \sum_{\theta_i > \theta_c} P(\hat{\theta} < \theta_c | \theta_i) f(\theta_i) / n \tag{2}$$

Referring to (1), the probability of having an observed score above the cut score given θ_i, $P(\hat{\theta} > \theta_c \mid \theta_i)$, is the area under the normal curve and to the right of

$$z = \frac{\theta_c - \theta_i}{se(\theta_i)} \tag{3}$$

This is illustrated in Figure 1. The taller bell curve represents the distribution of ability in the entire population and the cut score is set at θ_c = -.2. The smaller curve represents the expected distribution of observed values of theta for examinees with a true value of θ_i= -.5. Examines with a true score of θ_i=-.5 are non-masters and should be classified that way. However, the observed scores will vary around θ_i=-.5. The shaded area to the right of the cut score represents the proportion of examinees whose true score of -.5 that can be expected to be misclassified as masters. Figure 1 is for just one value of theta. To determine P(cm,n), one would have curves for each value of theta less than θ_c.

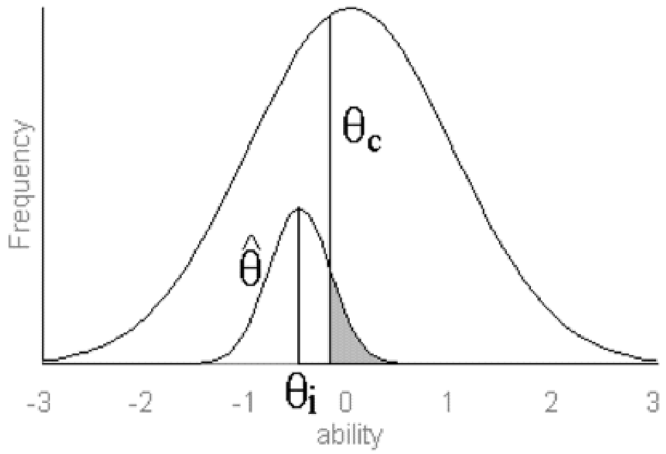

Figure 1: False positives for examinees at one ability level.

In (3), se(θ_i), the standard error of measurement evaluated at a score of θ_i, is

$$se(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{4}$$

where $I(\theta_i)$ is the test information function evaluated at a score of θ_i. Lord(1980, pages 72-74) provides equations for $I(\theta_i)$ using weighted composite scoring, number right scoring, and item response theory scoring. Using IRT, the test information function at θ_i is the sum of the item information functions at θ_i which can be evaluated from the IRT *a, b*, and *c* item parameters.

The expected proportions of examinees whose true score is θ_i, $f(\theta_i)/n$, can be estimated from the pilot sample. Denoting the probability of obtaining a score of θ_i as $P(\theta_i)$, and looking at θ as a continuous rather than a discrete variable, (1) and (2) become:

$$P(\text{cm}, \text{n}) = \int_{\theta_i=-\infty}^{\theta_c} P(\hat{\theta} > \theta_c | \theta_i) P(\theta_i) d\theta_i \tag{5}$$

$$P(\text{cn}, \text{m}) = \int_{\theta_i=\theta_c}^{\infty} P(\hat{\theta} < \theta_c | \theta_i) P(\theta_i) d\theta_i \tag{6}$$

To complete the set of equations needed for our two-by-two classification table, the probability of correctly being classified as a master and the probability of correctly being classified as a non master are

$$P(\text{cm}, \text{m}) = \int_{\theta_i=\theta_c}^{\infty} P(\hat{\theta} > \theta_c | \theta_i) P(\theta_i) d\theta_i \tag{7}$$

$$P(\text{cn}, \text{n}) = \int_{\theta_i=-\infty}^{\theta_c} P(\hat{\theta} < \theta_c | \theta_i) P(\theta_i) d\theta_i \tag{8}$$

If one assumes a normal distribution of mean μ and standard deviation , then $P(\theta_i)$ is the height of the Gaussian curve evaluated at $(\theta_i - \mu)/\sigma$.

Illustration

To illustrate an application of the above formulas, I generated a data set consisting of the item parameters for 50 items. The *a, b,* and *c* parameters were each normally distributed with means of 1.43, 0.00, and 0.20 respectively. Assuming a normally distributed examinee population with mean=0 and sd=1.0, the weighted average standard error is .393 which corresponds to a classical reliability coefficient of .92.

Applying equations (5) through (8) with a passing score of θ_c= -.25, which is about the 40th percentile, yields

Table 2: Expected and true examinee classifications			
		Expected Classification	
		Master	Non-Master
True	Master	53.9	5.0
	Non-Master	6.5	34.6
			Accuracy = 88.5%

This is a highly reliable test that is expected to accurately classify examinees an impressive 88.5% of the time. On closer examination however, 5.0/(5.0+53.9)= 8.5% of the true masters can be expected to be incorrectly classified as non-masters and 6.5/(6.5+34.6)=15.8% of the non-masters can be expected to be incorrectly classified as masters. Whether this is sufficiently accurate is matter of judgment.

The percent of expected misclassifications can be used as the benefit side of a cost benefit analysis. We can examine the potential gain obtained by increasing the quality of the test by simulating the addition of items with peak information at the cut score. If we add 10 items with parameters a=2.0, b=-.25 and c=0, we obtain the following classifications:

Table 3: Expected and true examinee classifications adding 10 optimal items			
		Expected Classification	

		Master	Non-Master
True	Master	55.4	3.6
	Non-Master	4.8	36.3
			Accuracy = 91.7%

Accuracy goes up from 89.1% to 91.7% and the proportion of false negatives goes down from 5.0 to 3.6% of all test takers. Here, the marginal benefits of improving the test may not justify the associated costs. On the other hand, this is a 28% reduction in the number of false positives (5.0-3.6)/5.0. This could be very worthwhile if false negatives are costly.

A common approach to minimizing the number of false negatives is to render due process by providing repeated opportunities to demonstrate mastery. If we make the convenient assumption that test scores from different administrations are independent, then the probability of a true master being misclassified after three attempts is the product of the probabilities or, in this case, $.085^3$=.0006. With three opportunities to pass, only a very small fraction of true masters will be misclassified as non-masters. Of course, the probability that a non-master will pass in three tries increase from .158 to .158+.158*(1-.158)+.158*(1-.158(1-.158))= .428.

One common, albeit often misguided, approach to reducing the number of false negatives is to set the operational cut score to a value lower than that recommended by a standard setting panel. This can be modeled by maintaining the original cut score in equations (5) through (8), but integrating to and from the adjusted cut score rather than the original cut score. Using an operational cut score of -.40 rather than -.25 yields the following results:

Table 4: Expected and true examinee classifications using an adjusted operational cut score			
		Expected Classification	
		Master ($\theta_{c'}$>-.40)	Non-Master ($\theta_{c'}$<-.40)
True	Master (θ_c>-.25)	56.2	2.7
	Non-Master (θ_c< -.25)	9.8	31.3
			Accuracy = 87.5%

Now the true master is half as likely to be misclassified, 2.7/(2.7+56.2)= 4.6%. However, the non-master now has a 9.8/(9.8+31.3)=23.8% chance of being classified as a master. If the original standard were meaningful then setting a lower operational cut score is a poor alternative. If the rationale for lowering the operational test score is to recognize the error associated with assessment, then the approach is misguided. Error is assumed to be normally distributed. An individual's score is as likely to be above his or her true score as it is to be below. It would, however, be appropriate to make an adjustment in order to recognize the error associated with the standard setting practice. This could be viewed as simply implementing the judgment of a higher authority.

Concluding remarks

This paper consistently talked about true masters and true non-masters. One must recognize that the classifications always involve judgment (Dwyer, 1996) and that, despite the use of quantitative techniques, cut scores are always arbitrary (Glass, 1978). We cannot say that a person has mastered Algebra just because his or her true or observed score is above some cut point. Algebra, or almost any domain, represents a collection of skills and hence is not truly unidimensional. Because we are talking about a multidimensional set, it is illogical to talk about mastery as if it were a unidimensional set. The only true masters are those who get everything right on the content sampled from the larger domain.

Nevertheless, we recognize the need to establish cut scores; mastery in this paper refers to people who score above some established cut score. When that mastery - nonmastery decision affects real people, then the expected impact of that decision should be examined. This paper provides a way to estimate the number of false positives and false negatives using 1) the standard error, which could be the standard error of measurement or an IRT standard error, and 2) the expected examinee ability distribution, which could be estimated from a pilot sample or based on a distribution assumption, such as a normality assumption. It is our hope that this tool will lead to better, more informed, decision making.

Notes:

Internet-based software to apply this technique is available at http://pareonline.net/misclass/class.asp. Comparable

QuickBasic source code and a Windows executable file are available from the author.

This research was sponsored with funds from the National Institute for Student Achievement, Curriculum and Assessment, U.S. Department of Education, grant award R305T010130. The views and opinions expressed in this paper are those of the author and do not necessarily reflect those of the funding agency.

References and additional reading

Cizek, Gregory J. (1996). An NCME Instructional module on setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20-31.

Dwyer, Carol Anne (1996). Cut scores and testing: statistics, judgment, truth, and error. *Psychological Assessment*, 8(4), 360-62.

Geisinger, Kurt F. (1991).Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice*, 10(2), 17-22.

Glass, Gene V (1978).Standards and criteria.*Journal of Educational Measurement*, 15(4), 237-61.

Lord, Frederick M. (1980). *Applications of item response theory*. New Jersey: Lawrence Erlbaum and Associates.

Lawrence M. Rudner is the Director of the ERIC Clearinghouse on Assessment and Evaluation and the Assistant Director of the Maryland Assessment Research Center for Education Success, Department of Measurement Statistics and Evaluation, 1129 Shriver Laboratory, University of Maryland College Park, MD 20742.

Descriptors: Classification; *Cutting Scores; *Error of Measurement; Pass Fail Grading; * Scoring; * Statistical Analysis

Citation: Rudner, Lawrence M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14). Available online: http://PAREonline.net/getvn.asp?v=7&n=14.