Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 10 Number 13, August 2005 ISSN 1531-7714

Expected Classification Accuracy

Lawrence M. Rudner
Graduate Management Admission Council

Every time we make a classification based on a test score, we should expect some number of misclassifications. Some examinees whose true ability is within a score range will have observed scores outside of that range. A procedure for providing a classification table of true and expected scores is developed for polytomously scored items under item response theory and applied to state assessment data. A simplified procedure for estimating the table entries is also presented.

Given a test composed of N items used to classify examinees into one of K score groups, what percent of examinees can we expect to be classified properly? A procedure for computing expected classification accuracy for dichotomous items (Rudner, 2001) is extended to polytomous items and applied in this paper. A simple procedure for estimating accuracy based on data appearing in technical reports is also presented.

By definition, for any given true score, θ , the corresponding observed score, $\hat{\theta}$, is expected to be normally distributed, with a mean of θ and a standard deviation of se(θ). The probability of an examinee with a given true score of θ having an observed score in the interval [*a*, *b*] on the theta scale is then

$$\text{Prob}(a < \hat{\theta} < b \mid \theta) = \phi(\frac{b-\theta}{se(\theta)}) - \phi(\frac{a-\theta}{se(\theta)}) \qquad (1)$$

where $\phi(z)$ is the cumulative normal distribution function. This is the area under the normal curve between *a* and *b* with mean θ and standard deviation se(θ).

Multiplying equation (1) by the expected proportion of examinees whose true score is θ yields the expected proportion of examinees whose true score is θ expected to be in interval [*a*,*b*]. Summing or integrating over all examinees in interval [*c*,*d*] yields the expected proportion of all examinees that have a true score in [*c*,*d*] and an observed score in [*a*,*b*]:

$$\sum_{\theta=c}^{d} P(a < \hat{\theta} < b \mid \theta) f(\theta)$$

where *f*(θ) is the expected proportion of examinees whose true score is θ . If we assume θ is N(μ,σ) then *f*(θ) is the standard normal density function $\varphi(z)$.

Setting [*a*,*b*] and [*c*,*d*] to correspond to the true score intervals defined by the cut scores yields the elements of a classification table showing the expected proportion of all examinees with observed and true scores in each cell. The individual elements of the classification table are:

$$\sum_{\theta=c}^{d} P(a < \hat{\theta} < b \mid \theta) f(\theta) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{se(\theta)}\right) - \phi\left(\frac{a-\theta}{se(\theta)}\right) \right) \varphi(\frac{\theta-\mu}{\sigma}) \qquad (2)$$

where $f(\theta)$ is the expected proportion of examinees whose true score is θ. One computes equation (2) for each cell to obtain the complete K x K classification table. Overall accuracy, then, is the sum of the diagonal entries.

Example

Table 1 contains the item parameters for the 10-item 2001 Maryland State Performance Assessment Program Grade 8 Reading Test, Form A. The test was calibrated and scored by CTB-McGraw Hill using the generalized partial credit model (Muraki, 1992). The K=5 score intervals are [(375,489), (490,529), (530,579), (580,619), (620,650)]. Mean and standard deviation were originally set to be 500 and 50.

Table 1: Generalized Partial Credit Model
Item Parameters for a 10-Item Reading Test

Item No	a	b1	b2	b3
1	0.040	20.103	18.650	
2	0.040	21.231	19.442	
3	0.037	19.573	18.674	
4	0.044	22.838	21.573	22.206
5	0.043	22.941	21.357	
6	0.042	21.926	18.325	
7	0.051	26.644	23.579	
8	0.049	25.684	23.270	
9	0.052	27.247	25.523	
10	0.037	20.104	19.191	

Because the standard error at theta is the reciprocal of the square root of the Test Information Function at theta, the Test Information Function is the sum of the Item Information Function, and under the generalized partial credit model (Donoghue, 1994), the Item Information Function is

$$I_i(\theta) = a_i^2 \left[\sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_i} k P_{ik}(\theta) \right)^2 \right]$$

where a_i is the item discrimination index, the standard error for a given value of theta is:

$$se(\theta) = 1 \Big/ \sqrt{\sum_{i=1}^{n} a_i^2 \left[\sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_i} k P_{ik}(\theta) \right)^2 \right]} \qquad (3)$$

Using Table 1 and equations (2) and (3) yields the classifications shown in Table 2. The sum of the diagonals in Table 2, 81.4%, is the expected accuracy. In this case, the testing agency recognized that the accuracy of individual scores would not be sufficiently high and only reported aggregated test scores. An examination of the marginals reveals that the expected proportion of examinees in each category never differs from the true score category by more than .5%. Thus this 10-item test appears sufficient for reporting aggregated scores.

Table 2: Expected Classification Table - Percent of Examinees in Each Score Category

		Expected Score Category					
		(375-489) 0	(490-529) 1	(530-579) 2	(580-619) 3	(620-650) 4	
True score category	0	33.4	4.9	0.0	0.0	0.0	38.3
	1	4.7	33.3	3.8	0.0	0	41.8
	2	0.0	3.5	14.2	1.1	0	18.8
	3	0.0	0.0	0.4	0.5	0.2	1.1
	4	0.0	0.0	0.0	0	0	0.0
		38.1	41.7	18.4	1.6	0.2	100

A Simplified Estimate

The test contractor routinely provides the standard errors at the different cut scores. By making the convenient assumption that the standard errors are linear within a range, one can compute an estimated se(θ)' as a linear extrapolation of the values provided by the contractor:

$$se(\theta)' = \frac{\theta - a}{b - a}\left(se_b - se_a\right) + se_a \qquad (4)$$

where *a* and *b* are adjacent cut scores on the theta scale and se_a and se_b are the corresponding standard errors.

This allows one to make a simple estimate of accuracy using equation (4) rather than the more complex equation (3). For the test in our example, the standard errors at scaled scores of 375, 490, 530, 580, 620, and 650 are 60, 11, 12, 26, 57, and 106 respectively. Using these values, along with Table 1 and equations (3) and (4), yields the following estimated expected truth table.

The sum of the diagonals is 78.0% which is fairly close to the first estimate. The advantage of this approach is that one does not need to compute the probabilities of selecting option k for each item.

DISCUSSION

The accuracy of a test is usually gauged by summing across all possible scores, e.g., root mean square and goodness-of-fit. Yet if one is only interested in estimating the proportion of students mastering a content area or the proportions of students in a discrete category, then categorical analysis might be more appropriate. This paper presented a categorical approach. In this case, a 10-item test provided accurate classifications 81% of the time. The estimated and actual proportions of examinees in each score interval never differed by more than .5%.

Table 3: Estimated Expected Classification Table - Percent of Examinees in Each Score Category

		Expected Score Category					
		(375-489) 0	(490-529) 1	(530-579) 2	(580-619) 3	(620-750) 4	
True Score Category	0	31	7.1	0.1	0.0	0	38.2
	1	4.9	32.9	4.1	0.0	0.0	41.9
	2	0.0	3.9	13.6	1.3	0.0	18.8
	3	0.0	0	0.4	0.5	0.2	1.1
	4	0.0	0.0	0	0	0	0.0
		35.9	43.9	18.2	1.8	0.2	100

One can use this approach to estimate accuracy with different numbers of items. If we halve the standard errors, possibly by using much better items or more likely by lengthening the test length, accuracy would increase to 91%. That value might be adequate for many purposes. Reducing the standard errors four-fold would yield an expected classification accuracy of 95% - an improvement that may or may not be worth the additional cost.

REFERENCES

Donoghue, J.R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 31(4), 295-311.

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.

Rudner, L.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.

Note

Based on a paper originally presented at the annual meeting of the National Council ion Measurement in Education, San Diego, CA April 2004.

Citation

Rudner, Lawrence M. (2005). Expected Classification Accuracy. *Practical Assessment Research & Evaluation*, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13

Author

Lawrence Rudner is the Vice President for Research and Development at the Graduate Management Admission Council.