Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 11 Number 6, October 2006 ISSN 1531-7714

Expected Classification Accuracy using the Latent Distribution

Fanmin Guo
Graduate Management Admission Council

Rudner (2001, 2005) proposed a method for evaluating classification accuracy in tests based on item response theory (IRT). In this paper, a latent distribution method is developed. For comparison, both methods are applied to a set of real data from a state test. While the latent distribution method relaxes several of the assumptions needed to apply Rudner's method, both approaches yield extremely comparable results. A simplified approach for applying Rudner's method and a short SPSS routine are presented.

Estimating and reporting classification accuracy is an important practice for licensure and certification examinations. It has become a more common practice recently for large-scale state assessments, as more and more states report students' proficiency levels in addition to standardized scores. Classification accuracy index is an important piece of evidence for the technical quality of the assessment instrument used for these purposes. Rudner (2001, 2005) proposed a method based on item response theory (IRT) for evaluating accuracy for tests used to classify examinees into one of a finite number of score categories. Rudner first developed his approach for tests with dichotomous items and then extended the method to tests with partial credit items.

For simplification of descriptions, the author will change the terms used by Rudner without changing the concept. With Rudner's method, the first step is to map the x cut score(s) on the reporting scale onto the θ scale in order to divide the θ scale into $x + 1$ score category ranges. Then an individual examinee's $\hat{\theta}$ and its standard error of estimation are used to build a distribution of the $\hat{\theta}$ for this examinee. By summing up the density of the distribution within each score category range across all examinees, Rudner is able to calculate the expected proportion of examinees who fall into each of the score category ranges. A classification table comparing the expected and observed proportions in each score category range provides a basis for evaluating the classification accuracy of the test.

Li and Sireci (2005) adapted this method to number right, or raw, score scales. A simulation study to evaluate the properties of Rudner's method (Martineau, in press) found that his classification accuracy index was a useful method for evaluating the classification categories of 15 or more students in each of the categories.

Rudner's method assumes estimation error is normally distributed around each examinee's estimate of θ. Based on Mislevy's (1984) seminal latent distribution paper, an alternate method has been developed in this paper to accomplish the same goal without the basic assumptions. Like Rudner's method, this latent distribution method

provides for calculating the expected number of examinees in each of the score category ranges and compares them with the observed number of examinees in the ranges. Both methods are applied to a set of real data from a state test for comparison.

Point Estimation of θ vs. Latent Distribution

Most IRT-based tests intend to find a point estimation of an examinee's ability on the latent θ scale. Maximum likelihood method is often used for this purpose by calculating the likelihood function with an examinees' response vector and the item parameters. For dichotomous items, the likelihood function is defined as:

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_i} Q_j^{1-u_i}$$

where i is the examinee;

j is the item on the test;

u is a response to item j for examinee i, coded as 1 for a correct answer and 0 for an incorrect answer;

$P_j^{u_i}$ is the probability of a correct answer to the item j at θ; and

$Q_j^{u_i}$ is the probability of an incorrect answer to the item j at θ that can be calculated as 1-$P_j^{u_i}$.

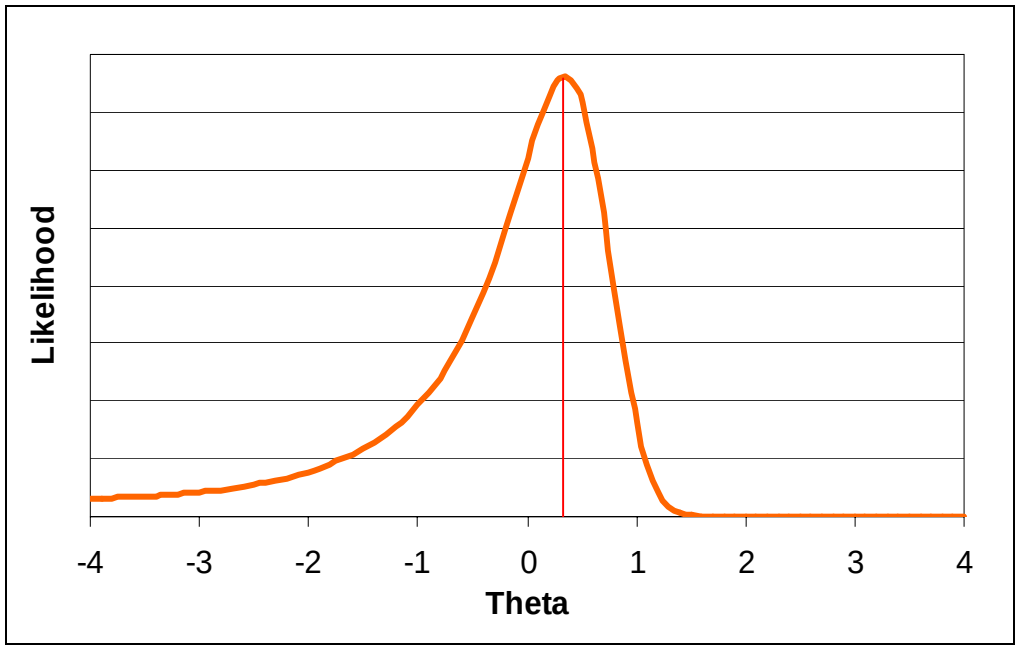
The θ value that maximizes the likelihood function is the estimated $\hat{\theta}$ for this examinee. An estimated standard error of estimation (SEE) of this $\hat{\theta}$ can be calculated, based on the test information at the $\hat{\theta}$. This SEE estimate is the standard deviation of the $\hat{\theta}$ distribution. The $\hat{\theta}$ converges to θ as the number of examinees and number of items increase.

Figure 1 is a plot of an examinees' likelihood function from a 35-item test; the $\hat{\theta}$ for the examinee is 0.33. Note that the likelihood function is not symmetrical around $\hat{\theta}$ and that the lower tail approaches 0 at a much slower rate.

Mislevy (1984) discussed the estimation of latent distributions when point estimation of θ was poorly estimated as a result of too few items. He proposed the method for estimating the population parameters from the individual latent distributions, not the point estimates of θ. Using Bayesian estimation, an examinees' latent distribution was defined as the posterior distribution, which was the likelihood function weighted by a prior distribution. For our purposes, a likelihood function suffices because the item parameters are "known" and the examinees' scored responses are known. It is a simple "conditional maximum likelihood estimation of ability" described by Hambleton and Swaminathan (1985. p. 81). If a "non-informative" prior distribution is set, the posterior distribution is equal to the likelihood function.

A likelihood function of an examinee can be interpreted as the likelihood of an examinee's ability at each θ point given his/her responses and the characteristics of the items. The function is another representation of an examinee's ability as a distribution across the θ scale as plotted in Figure 1 and can be used for the purpose of calculating the expected proportion of examinees in each of the score category ranges. This is similar to Mislevy's situation, where the population distribution in each of the score category ranges is the interest. It serves as an alternative method to the method proposed by Rudner (2001, 2005), who essentially built an examinee's distribution of $\hat{\theta}$ with the point estimate of θ and its standard error of estimation.

Figure 1: A Likelihood Function and Estimated $\theta = 0.33$



Latent Distribution Method

The latent distribution method involves seven steps. Each is defined and described in this section.

1. Map the cut scores onto the θ scale. After setting the standards for a test, x cut scores will be chosen to divide the reporting scale into r score category ranges ($r = x + 1$). Convert these cut scores to their corresponding θ values and add the θ values corresponding to the lowest and highest possible scores on the test. There are M θ values ($M = x + 2$). If m denotes these θ values, then $m = 1, 2, \ldots, M$. Here θ_1 = the θ value corresponding to the lowest possible score and θ_M corresponds to the highest possible score on the test.

2. Calculate likelihood. Calculate the likelihood of an examinee i in score category range r using his or her scored responses, μ_1 to μ_n, and the item parameters, $a_1, b_1, c_1, \ldots, a_n, b_n, c_n$, of a 3-parameter logistic IRT model, for example.

$$L_{ri} = \int_{m}^{m+1} L(u_1, u_2, \ldots, u_n \mid \theta) d\theta$$

For the convenience of computations, the continuous θ scale can be made into discrete categories based on θ points with equal distances. The likelihood can be calculated as follows:

$$L_{ri} = \sum_{\theta=m}^{m+1} L(u_1, u_2, \ldots, u_n \mid \theta)$$

3. Normalize the likelihood. Normalize the likelihood so that the sum of the likelihood for each examinee equals to 1. This is necessary because the distributions must be truncated at the lowest and highest obtainable scores, not at $-\infty$ and ∞, in order to do the calculations. After normalization, the sum of the likelihood across all examinees will be equal or very close to the total number of examinees. This, in turn, will simplify the interpretation of the results.

$$NormL_{ri} = \frac{L_{ri}}{\int_{1}^{M} L(u_1, u_2, \ldots, u_n \mid \theta) d\theta}$$

Similarly, the sum of the likelihood in the denominator can be calculated as the sum across the discrete θ values from θ_1 to θ_M.

4. Compute the observed number of examinees in a score category range *s*. The observed number of examinees in a score category range *s* is the number of examinees whose maximum likelihood point estimates, $\hat{\theta}$, fall in that category range.

$$O_s = N_{(m \le \hat{\theta} < m+1)}$$

5. Compute the expected number of examinees in each cell $N_{(s,r)}$ of the classification table. For the examinees who fall in the observed score category range *s*, their numbers in each of the expected score category range *r* can be calculated as follows:

$$N_{sr} = \sum_{i \in s} NormL_{ri}$$

6. Assemble the classification table. Table 1 is an example. Please note that decimal places will be encountered in the cells as a result of the proportional redistribution of some individual examinees into more than one expected score category.

Table 1: Example of a Classification Table

		Expected N in score category range		
		r_1	…	r_M
Observed N in score category range	s_1	$N_{(1,1)}$	$N_{(1,\ldots)}$	$N_{(1,M)}$
	…	$N_{(\ldots,1)}$	$N_{(\ldots,\ldots)}$	$N_{(\ldots,M)}$
	s_M	$N_{(M,1)}$	$N_{(M,\ldots)}$	$N_{(M,M)}$

7. Calculate the accuracy index. The accuracy index can be calculated in the same way proposed by Rudner. It is simply the percentage of the sum of the diagonal divided by the total number of examinees.

$$\frac{\sum_{s=r=1}^{M} N_{(s,r)}}{\sum_{s=1}^{M}\sum_{r=1}^{M} N_{sr}}$$

The accuracy index indicates the percentage of examinees who are correctly classified. The higher the index, the more accurate the test is in classifying examinees into categories.

Because the likelihood functions are computed using item parameters, an important assumption here is that the item parameters are "known." That is, the item parameter estimates are reasonably close to their true values. For best results, be sure the item parameters are calibrated with a large number of examinees whose abilities cover the entire range of the θ scale where scores are reported. For the latent distribution method, some of the assumptions required for Rudner's method, such as the normality of the SEE estimates and the reasonable approximation of $\hat{\theta}$ to θ, are no longer needed.

An Alternate Approach to Applying Rudner's Method

Instead of computing the proportion of examinees in each of the expected score categories, the number of examinees was calculated using the same basic steps as those for the latent distribution method, but with a few minor changes. This alternate calculation yields the same results and is easier to apply. The following changes were applied:

1. The lowest and highest possible scores in Step 1 of the latent distribution method were replaced by the negative and positive infinities.

2. The likelihood function in Step 2 of the latent distribution method was replaced by a normal distribution with a mean of the $\hat{\theta}$ and a standard deviation of the standard error of estimation (SEE) for each examinee and can be computed as $\phi_{ri} = \phi(m+1, \hat{\theta}, \text{SEE}) - \phi(m, \hat{\theta}, \text{SEE})$, the density between m and $m+1$.

3. Step 3 of the latent distribution method is not needed because the area under a normal curve is 1.

4. $\sum_{i \in s} NormL_{ri}$ in Step 5 of the latent distribution method was replaced with $\sum_{i \in s} \phi_{ri}$.

The alternative computations for Rudner's method need only several lines of codes in a statistical analysis program like SPSS or SAS. An example with SPSS codes is included in the Appendix.

Comparison with an Example

Both Rudner's and the latent distribution methods were applied to real test data for comparison. The data used is from a state test with 32 items and reported scores ranging from 275 to 575. The reliability is 0.87 (Cronbach's alpha). Three proficient categories are reported for students: Basic (275 to 410), Proficient (411 to 446), and Advanced (447 to 575). Table 2 presents the reported scores and their corresponding θ values used in calculating the number of examinees falling into each score category range for both methods. In Table 2, m_1 is the θ corresponding to the lowest possible score for the latent distribution method and $-\infty$ for Rudner's method; m_2 is the first cut score; m_3 is the second cut score; and m_M is the θ corresponding to the highest achievable score for the latent distribution method and ∞ for Rudner's method.

Table 2: θ and Reported Scores Used in the Calculations

	m_1	m_2	m_3	m_M
Reported Score	275	410.5	446.5	575
θ for Latent Distribution Method	–3	0.1755	1.1475	3
θ for Rudner's Method	$-\infty$	0.1755	1.1475	∞

Item parameters used for calculating examinees' likelihood are the same ones from which the point estimates of $\hat{\theta}$ and the standard errors of estimation were derived.

Table 3 presents the observed and expected number of examinees and their percentages using Rudner's method, and Table 4 exhibits the same information calculated with the latent distribution method. The estimated classification accuracy indices are .858 for Rudner's method and .870 for the latent distribution method.

Table 3: Classification Table Using Rudner's Method							
		Expected					
		Basic		**Proficient**		**Advanced**	
Observed	Basic	2690	44.1%	282	4.6%	1	0.0%
	Proficient	236	3.9%	1961	32.2%	190	3.1%
	Advanced	1	0.0%	155	2.5%	582	9.5%
Sum		2927	48.0%	2398	39.3%	773	12.7%

Table 4: Classification Table Using the Latent Distribution Method							
		Expected					
		Basic		**Proficient**		**Advanced**	
Observed	Basic	2747	45.0%	226	3.7%	0	0.0%
	Proficient	225	3.7%	1951	32.0%	212	3.5%
	Advanced	0	0.0%	129	2.1%	608	10.0%
Sum		2927	48.7%	2306	37.8%	820	13.4%

Table 5 shows the differences between the two methods in the expected number of examinees in each of the three proficiency categories. The differences are calculated by subtracting Table 3 values (Rudner's method) from Table 4 values (the latent distribution method) for each cell. As shown in the sums, the latent distribution method tends to put more examinees in the Basic and Advanced categories and fewer examinees in the Proficient category than Rudner's method.

Table 5 also shows that, for the low ability examinees who fell into the observed Basic category (about 48% of the total), the latent distribution method puts more of those examinees into the expected Basic category than does Rudner's method. For the examinees in the observed Proficient and Advanced categories (about 52% of the examinees with higher ability), the latent distribution methods puts more of those examinees into the expected Advanced category than Rudner's method does. It seems that the latent distribution method produces an expected ability distribution with fatter tails than that produced by Rudner's method. However, the difference is very small between the two methods, about 1.2%, as indicated by the classification accuracy indices.

Table 5: Differences between the Two Methods

		Expected		
		Basic	**Proficient**	**Advanced**
Observed	Basic	57	–56	–1
	Proficient	–11	–10	22
	Advanced	–1	–26	26
Sum		45	–92	47

Conclusions and Discussion

Today, more and more achievement tests are reporting both individuals' scores and performance categories, such as Pass/Failure or Basic/Passing/Advanced. Evaluating the accuracy of the classification of examinees into the categories becomes increasingly important in educational settings. Rudner (2001, 2005) proposed an index for this purpose. An alternative was proposed in this paper using latent distributions.

Both methods were applied to a set of real test data from a state test for comparison. The comparison showed that the latent distribution method tends to put more low ability examinees into the expected low ability categories and more high ability examinees into expected high ability categories than does Rudner's method. However the difference is very small, about 1.2%.

The latent distribution method uses the same strategy as Rudner's method. The classification index is the percentage of agreement between the observed and the expected proportions of examinees in each of the categories under the IRT framework. The latent distribution method differs from Rudner's method in calculating the expected number of examinees in each category with the posterior distributions (the normalized likelihood function) of the examinees. As a result, some assumptions for Rudner's method are no longer needed. Therefore, the latent distribution method might be a more robust method when the estimation of θ is less accurate due to small number of items on a test or low test information at some ability levels. The comparison was made with a reliable test ($\alpha = 0.87$). Further research is needed to see how the conclusion of small differences between the two methods holds when the θ estimation becomes poor.

When the assumptions are met, or even approximated, Rudner's method is a very easy method. Using the procedure and calculations proposed in this paper, it only takes several lines of code in a statistical package to calculate the expected number of examinees in the classification table. While harder to apply, the latent distribution method outlined in this paper has a stronger theoretical foundation. This method is always applicable. The limitation of this method is that it relies on sound parameter estimates as expected classifications are computed at every possible theta point.

References

Hambleton, R. K, & Swaminathan, H. (1985). *Item response theory*. Boston: Nijhoff Publish.

Li, S., & Sireci, S. (2005, April). *Evaluating the accuracy of proficiency classifications using item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Martineau, J. A. (in press). An Expansion and Practical Evaluation of Expected Classification Accuracy. *Applied Psychological Measurement*.

Mislevy, R.J. (1984). Estimating latent distribution. *Psychometrika*, *49*, 359-381.

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

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

Acknowledgements

The views and opinions expressed in this paper are those of the author and do not necessarily reflect the views and opinions of the Graduate Management Admission Council®.

Citation

Guo, F. (2006). Expected Classification Accuracy using the Latent Distribution. *Practical Assessment Research & Evaluation*, 11(6). Available online: http://pareonline.net/getvn.asp?v=11&n=6

Author

Fanmin Guo has a PhD in Research Methodology from the University of Pittsburgh and is the Director of the Psychometric Research at the Graduate Management Admission Council. For questions or comments, please contact Fanmin Guo at fguo at gmac.com.

Appendix

An SPSS Example of Computing Expected Classifications with Rudner's Method.

Data and Data File Layout.

The data file contains one record for each examinee with four variables: Examinee ID, estimated θ (ThtEst), standard error of estimation (ThtSEE), and observed classification (Group: coded as 1=ObsBasic, 2=ObsProficient, and 3=ObsAdvanced).

SPSS codes for calculating the densities in the three categories for each examinee.

```
COMPUTE ExBasic = CDF.NORMAL(-3,ThtEst,ThtSEE).
COMPUTE ExProfi = CDF.NORMAL(1.1475,ThtEst,ThtSEE) -
   CDF.NORMAL(.1755,ThtEst,ThtSEE).
COMPUTE ExAdvan = 1 - CDF.NORMAL(1.1475,ThtEst,ThtSEE).
EXECUTE.
```

SPSS codes for calculating the expected number of examinees in each of the cells in a classification table.

```
SPLIT FILE
   SEPARATE BY Group.
DESCRIPTIVES
   VARIABLES=ExBasic ExProfi ExAdvan
   /STATISTICS=SUM.
EXECUTE.
```