Paper presented at the annual meeting of the
National Council on Measurement in Education,
Chicago, April 23-25, 2003

The Classification Accuracy of Measurement Decision Theory

Lawrence Rudner
University of Maryland

Item responses on an operational state assessment were calibrated and analyzed using decision theory with the size of the calibration sample being the primary manipulated variable and score-group classification accuracy being the primary assessment goal. Simple decision theory was shown to be highly accurate in terms of placing individuals into the appropriate score categories. If the intent of an assessment is to classify individuals into discrete categories or determine the proportions of examinees within each score category, then decision theory provides an attractive alternative to classical or modern measurement theory.

Classical measurement theory and item response theory are concerned primarily with rank ordering examinees across an ability continuum. But one is often interested in just classifying examinees into one of a finite number of discrete categories, such as pass / fail or below-basic / basic / proficient / excellent. Sometimes, one is interested in an even more basic outcome, such as just the proportion of all examinees whose ability fall within each discrete category. Classification is a simpler outcome than rank ordering and a simpler model should suffice. This paper presents and evaluates the use of simple decision theory as a tool for classifying examinees based on their item response patterns using actual data from a state-wide examination.

The primary research question in this study is whether decision theory can produce results that are comparable to that of the more complicated IRT classification procedures. An evaluation of the model is presented by examining the classification accuracy of tests scored using decision theory and examining the relationship between accuracy and the number of pilot test examinees used to calibrate decision theory item parameters.

The literature on the use of decision theory to analyze item responses is fairly scant. There was a slight surge of interest in the 1970s (e.g., Hambleton and Novick, 1973; Swaminathan, Swaminathan, Hambleton & Algina, 1975, Huynh, 1976; van der Linden and Mellenbergh, 1978) as a tool to score criterion-referenced tests. Macready and Dayton (1992), Welch and Frick (1993) and Vos (1999) used decision theory as the basis for adaptive testing within a latent class framework. Rudner (2002a) recently provided an evaluation of the model using simulated response patterns. Most of the measurement research to date, however, has applied decision theory to test batteries or as a supplement to item response theory and specific latent class models.

Background

An excellent overview of decision theory can be found in Melsa and Cohn (1978). The objective here is to form a best guess as to the mastery state (classification or latent state) of an individual examinee based on the examinee's item responses, *a priori* item information, and *a priori* population classification proportions. While most applications of decision theory are based on continuous random variables, our application uses the items within a test as the independent random variables.

One starts with the following:

- K possible mastery states, that take on values m_k.
- A test composed of N items.
- An item response vector, $\mathbf{z} = [z_1, z_2, ..., z_N]$ for each examinee. There is no restriction on the values of z_i. They can be dichotomous or polytomous, nominal or ordinal.

Based on a piloting the items with a calibration sample, one obtains

- $P(m_k)$ the probability of a randomly selected examinee belonging in mastery state m_k (e.g., an approximation of the portion of examinees below basic, basic, proficient and advanced).
- $P(z_i|m_k)$ the probability of response z_i given the k-th mastery state.

After calibration, the model is applied to a set of new examinees. For each examinee, one first computes the K different probabilities of the response vector that correspond to the each of the K mastery states. Assuming local independence,

$$P(\mathbf{z}|m_k) = \prod_{i=1}^{N} P(z_i|m_k).$$

That is, the probability of the response vector is equal to the product of the conditional probabilities of the item responses. In decision theory, the local independence assumption is also called the “naive Bayes” assumption. One naively assumes the assumption is true and proceeds with the analysis.

An estimate of the examinee’s mastery state is then formed using the priors and observations. By Bayes’ Theorem,

$$P(m_k|\mathbf{z}) = c\ P(\mathbf{z}|m_k)\, P(m_k) \cdot \quad (1)$$

The posterior probability $P(m_k|\mathbf{z})$ that the examinee is of mastery state m_k given his response vector is equal to the product of a normalizing constant (c), the probability of the response vector given $m_{k,}$ and the prior classification probability. Again, for each examinee, there are K probabilities, one for each mastery state.

One common rule for classifying an examinee based on these K probabilities is to select the category with the maximum *a posteriori* probability (MAP). An alternate approach is Bayes Risk Criterion, also called the Minimum Loss Criterion and the Optimal Decision Criterion. Costs are assigned to each correct and incorrect decision and then one makes the decision that minimizes the total average cost.

Data

The primary analysis in this study used the scored item responses made by 18,453 students who took the 10 items in the Reading, Grade 8, Booklet A of the 2001 Maryland State Performance Assessment Program (MSPAP). The test is composed of performance items scored using a 3- or 4- point scoring rubric. Individual ability scores computed by the state using the Generalized Partial Credit Model (Muraki, 1992) were first mapped to scaled scores and then to a 5-point performance scale using pre-defined cut scores. With the small number of items per subject per booklet, the state only reported aggregated scores in the form of proportions of students scoring within each of five categories.

The data were trimmed to include only the 15,386 students with scored responses to all 10 items. Omits and absences were excluded.

The analysis on the primary data set was partially replicated using scored item responses in nine other subject areas assessment as part of the MSPAP. Again, large (> 12,000) numbers of trimmed item response sets were available along with the state-assigned score category.

Analysis

Two types of samples were drawn from the trimmed data set. The first, a sample of 1,000 students, was randomly selected to form the trial-sample used to evaluate the model. That sample was used for all the analyses.

To calibrate the system, random samples of different sizes were randomly drawn from the remaining 14,386 students and used to compute estimates of P(m_k) and P(z_i|m_k). Equation (1) was then applied to each student in the trial sample and MAP was used to classify the examinees. The score classification based on decision theory and the score classification used by the state were recorded. The experiment was repeated 100 times for each calibration group size.

The first analysis examined the examinee-level accuracy of decision theory scoring. The proportion of the 1000 trial-sample students whose predicted score category matched the state-assigned score category was examined as a function of the size of the calibration sample.

The second analysis examined the group-level accuracy of decision theory scoring. The marginals of the table of state-assigned and predicted scores were recorded and compared. As a summary measure, the weighted average of the absolute value of the difference between the predicted and actual marginals was computed for each calibration group size:

$$d = \sum_{k=1}^{5} w_k \left| \hat{P}(m_k) - P(m_k) \right|$$

where P(m_k) are the actual proportions of examinees in category k, $\hat{P}(m_k)$ are the proportions predicted using decision theory, w_k=P(m_k), and 3w_k=1.0. This index averages the observed minus predicted error and weights it by the proportion of examinees affected.

Results

Accuracy at the individual level was defined as agreement between the classifications predicted by the model and the actual classifications assigned by the state using item response theory. Table 1 and Figure 1 show the mean and standard deviation of accuracy over the 100 iterations as a function of calibration sample size. With sample sizes of 1,000 and more, accuracy was exceptionally high. Approximately 95% of the classifications predicted by decision theory matched the categories assigned by the state using item response theory. Individual level accuracy is also quite respectable with as few as 100 examinees in the calibration sample.

As the calibration sample size increases, the variances of the accuracy estimates decrease. The variances are small relative to accuracy, implying that the different random samples used for calibration would yield comparable results.

Table 1: Accuracy as a function of calibration sample size

	Accuracy	
Calibration sample size	Mean	Standard Deviation
25	0.714	0.043
50	0.794	0.039
100	0.860	0.023
200	0.901	0.013
500	0.934	0.009
1000	0.947	0.007
1500	0.952	0.007
2000	0.956	0.006

Based on 1,000 examinees in the trial group and 100 iterations for each calibration sample size.

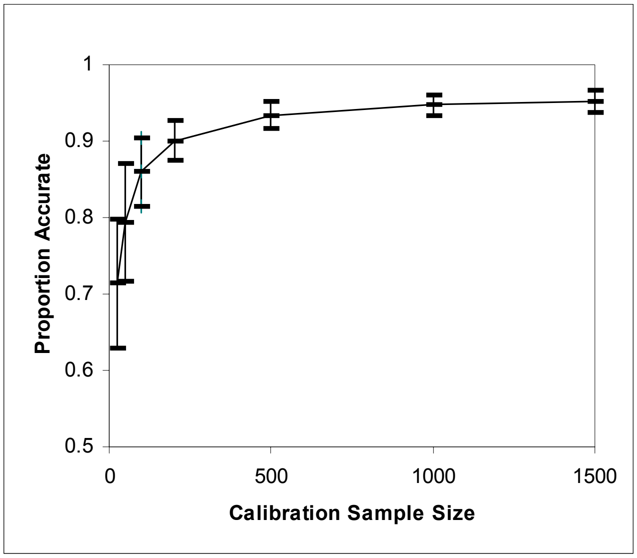

Figure 1: Proportion of trial examines accurately classified as a function of calibration sample size.

Group assignment accuracy was defined as comparable values of predicted and actual proportions of examinees in each score group. As shown in Table 2, the actual proportions of trial sample examinees in each of the five score groups are .192, .554, .224, .018, and .012. The predicted values are quite close. The difference between predicted and actual proportions is 1.2% or less when the calibration group sizes are 100 examinees or more. If one just looks at the proportions of students deemed satisfactory by the state, i.e., scoring 3 and above, then the results are even more impressive. The actual percent of satisfactory students in our trial sample is 25.4%. With 200 examinees in the calibration sample, the predicted percent of examinees with satisfactory scores is 25.3%.

Another way to assess the correspondence between actual and predicted proportions is to examine the weighted average of the absolute values of the corresponding differences, denoted *d* in Table 2. The weighted averages were extremely low. The proportions of effected examinees is not meaningful.

The above analysis was partially replicated using responses to 9 other tests within the MSPAP. In

Table 2: Predicted and actual proportions of examinees in each score group for Reading Grade 8 test booklet.

		Score Group						
Predicted Proportions	Calibration sample size	1 (low)	2	3	4	5 (high)	*d*	Max dif
	25	0.264	0.514	0.204	0.012	0.006	0.0006	0.072
	50	0.216	0.556	0.211	0.008	0.010	0.0001	0.024
	100	0.196	0.564	0.227	0.007	0.006	0.0002	0.011
	200	0.186	0.561	0.240	0.006	0.007	0.0002	0.016
	500	0.185	0.560	0.238	0.005	0.011	0.0001	0.014
	1000	0.181	0.559	0.242	0.005	0.013	0.0001	0.018
	1500	0.180	0.560	0.242	0.005	0.013	0.0001	0.018
	2000	0.180	0.560	0.242	0.005	0.012	0.0001	0.018
Proportions based on state assignments		0.192	0.554	0.224	0.018	0.012		

Based on 1,000 examinees in the trial group and 100 iterations.
d is the weighted average of the absolute values of the difference between predicted and state-assigned proportions.
Max Dif is the largest absolute value of the difference between predicted and state-assigned proportions.

each case, 1000 scored examinees were randomly selected from the set of all examinees with complete item responses and used as the trial sample. Separate random samples of 1,000 examinees were drawn and used to obtain 1) the proportions of examinees within each score group, $P(m_k)$ and 2) the proportions of examinees with each item response in each score group, $P(z_i|m_k)$. The calibrated a priori probabilities were applied to the responses of the trial group examinees. Accuracy and weighted marginal differences were computed. Again, the process was iterated 1000 times using the same trial response set.

The results of the replication are shown in Table 3. Accuracy is again extremely high, maximum marginal differences are quite small, and the weighted average marginal differences are quite small.

Table 3: Accuracy and differences in predicted and actual proportions of examinees in each score group for different assessment instruments.

			Accuracy			
Subject	Grade	N items	Mean	Standard Deviation	*d*	Max dif
Mathematics	3	21	0.963	0.008	0.0001	0.015
Social Studies	3	21	0.878	0.015	0.0016	0.039
Science	3	18	0.892	0.013	0.0019	0.053
Reading	5	12	0.924	0.008	0.0007	0.054
Mathematics	5	30	0.929	0.007	0.0009	0.026
Social Studies	5	20	0.854	0.014	0.0008	0.038
Science	5	16	0.884	0.012	0.0012	0.030
Social Studies	8	19	0.911	0.009	0.0016	0.041
Science	8	21	0.959	0.005	0.0001	0.010

Based on 1,000 examinees in the trial group and 100 iterations.

Discussion

Using actual item response data, this study showed that simple decision theory can yield exceptionally high classification accuracies. With a calibration sample of 1,000 randomly selected examinees from the Grade 8 Reading assessment, some 95% of the assessed individuals were placed in the same score category assigned using the complex machinery of item response theory. With as few as 100 calibration group examinees, individual accuracies of 86% were obtained and the predicted overall proportions of examinees in each score group differed from the corresponding actual proportions by less than 1.2%. Comparable results were obtained on nine other MSPAP tests using a calibration size of 1000 examinees.

One of the major limitations of the data analyzed was the relatively small numbers of students in the top two score groups, 1.8% and 1.2% respectively. One would suspect that the probabilities of each item score given membership in one of those two groups would be poorly assessed, even with large calibration group samples. With 500 or more examinees in the calibration group, high- ability examinees tended to be placed in the highest score group, rather than group 4. We suspect the results would have been even more impressive had these score groups been combined or merged with group 3.

In another study looking at the same Reading Grade 8 data set, Rudner(2002b) noted that IRT can be expected to provide the correct individual true score classification some 81% of the time. So how could decision theory provide accuracies well above 81%? The difference lies in the different definitions of accuracy. The IRT study compared scores predicted using IRT to true scores. This study compares scores predicted by decision theory to scores assigned using IRT. In other words, this study demonstrates that decision theory can capture those assigned classifications. Accuracy in terms of true score was not examined.

The simplicity and feasibility makes decision theory attractive for many applications. Tests embedded in computer instruction, for example, could be based on relatively small calibration samples. Short tests could be embedded in larger tests for a relatively small number of examinees to provide data for comparing assigned proportions. Subsets of items could be clustered and analyzed to provide skill level scores.

While this study provides support for the use of decision theory at the item level, it also raises a large number of questions for further research. This study used a handful of information-rich, polytomously scored items. Will decision theory work as well for multiple- choice items? How few items are needed to determine group proportions? How can we best select items with the goal of determining group classifications? How many are needed to obtain adequate individual classification accuracies? What is gained by more items? How many missing responses can be tolerated? Would adding a cost structure provide advantages?

Notes

An interactive tutorial on measurement decision theory is available at http://ericae.net/mdt/. Windows-based software for applying decision theory to test items is being developed and will be available at that web site.

References

Hambleton, R. and Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159-170.

Huynh, H. (1976). Statistical considerations for mastery scores. *Psychometrika*, 41, 65-79.

Macready, G. and Dayton, C. M. (1992). The application of latent class models in adaptive testing. *Psychometrika*, 57(1), 71-88.

Melsa, J.L and Cohn, D.L. (1978). *Decision and Estimation Theory.* New York: McGraw-Hill Book Company.

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

Rudner, L.M. (2002a). *Measurement Decision Theory*. Final report to the Office of Educational Research and Improvement, U.S. Department of Education.

Rudner, L.M. (2000b). Expected Classification Accuracy. Paper under review.

Swaminathan, H., Swaminathan, R. K., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 12, 87-98.

van der Linden, W. J. and Mellenbergh, G.J. (1978). Coefficients for tests from a decision-theoretic point of view. *Applied Psychological Measurement,* 2, 119-134.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics*, 24(3), 271-92.

Welch, R.E. & Frick, T. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research & Development,* 41(3), 47-62.