GRADUATE MANAGEMENT ADMISSION COUNCIL

Computer Adaptive Testing for Small Scale Programs and Instructional Systems

Lawrence M. Rudner & Fanmin Guo

GMAC® Research Reports • RR-11-01 • January 1, 2011

Abstract

This study investigates measurement decision theory (MDT) as an underlying model for computer adaptive testing when the goal is to classify examinees into one of a finite number of groups. The first analysis compares MDT with a popular item response theory model and finds little difference in terms of the percentage of correct classifications. The second analysis examines the number of examinees needed to calibrate MDT item parameters and finds accurate classifications even with calibration sample sizes as small as 100 examinees.

Computerized adaptive testing (CAT) offers numerous advantages to both the test sponsor and the test taker. Test questions can be selected based on interim ability estimates and answers to previous test questions. The result is a tailored test able to maintain desired precision levels with fewer questions.

Much of the research on adaptive testing has centered on the use of *item response theory* (IRT) as the underlying model. While very attractive and widely used, IRT is fairly complex, relies on several restrictive assumptions, and typically needs large numbers of examinees to obtain satisfactory item parameter estimates.

This paper presents an adaptive testing procedure using *measurement decision theory* (MDT) and compares this procedure with an IRT procedure in terms of classification accuracy. Accuracy is evaluated as a function of model, pool size, and calibration size.

Overview

In the introduction to their classic textbook, Cronbach and Gleser (1957) argue that the ultimate purpose for testing is to arrive at qualitative classification decisions. When the goal is to classify examinees into one of a finite number of categories, such as certify/deny certification or advanced/proficient/basic/below basic, IRT is traditionally the model of choice. IRT provides a point estimate on a continuum as the indicator of ability of each examinee, and classification is finessed by identifying the category that encompasses the point estimate. In terms of adaptive mastery testing, the field has not advanced much beyond the early studies of Kingsbury and Weiss (1979, 1981, and 1983). For a good overview of the literature on computerized classification testing, see Thompson (2007).

Decision theory provides an alternative underlying model. Key articles in the mastery testing literature of the 1970s employed decision theory (Hambleton & Novick, 1973; Huynh, 1976; van der Linden & Mellenbergh, 1977). Lewis and Sheehan (1990), Kingsbury and Weiss (1983), Reckase (1983), and Spray and Reckase (1996) have used decision theory to adaptively select items and testlets and determine when to stop testing. Notable articles by Macready and Dayton (1992), Vos (1999), Welch and Frick (1993), and Rudner (2002, 2009) illustrate the less prevalent item-level application of decision theory examined in this paper.

In this overview, measurement decision theory is described along with the investigated approach to

adaptively administer test questions. A more detailed description can be found in Rudner (2009).

Measurement Decision Theory

The objective of measurement decision theory is to form a best estimate of the true classification of an individual examinee based on the examinee's item responses, a priori item information, and possibly a priori population classification proportions. Thus, the decision theory model has four components: 1) possible mastery states for an examinee, 2) calibrated items, 3) an individual's response pattern, and 4) decisions that may be formed about the examinee.

There are K possible classifications that take on values m_k. In the case of pass/fail testing, there are two possible classifications and K = 2. The second component is a set of items for which the probability of each possible observation, usually right or wrong, is given for each classification. Just as with IRT CAT, these conditional probabilities of right or wrong answers are computed a priori. These conditional population probabilities can be determined a variety of ways, including prior testing, transformations of existing scores, existing classifications, and possibly judgment. One approach to computing the conditional probabilities is to compute the proportions of examinees that respond correctly from a small pilot test involving examinees that have already been classified.

The responses to a set of N items form the third component. Each item is considered to be a discrete random variable stochastically related to the classification levels and realized by observed values z_N. Each examinee has a response vector, **z**, composed of z_1, z_2... z_N. Only dichotomously scored items are considered in this paper.

The last component is the decision space. One can form any number of D decisions based on the data. Typically, one wants to make a classification decision and there will be D = K decisions. With adaptive testing, a decision to continue testing can be added and thus there would be D = K + 1 decisions.

This paper uses the following notation:

Priors

$P(m_k)$—the probability of a randomly selected examinee having mastery state m_k. In the absence of information and when classification priors are deemed inappropriate, uniform priors can be used.

$P(z_i \mid m_k)$—the probability of response z_i given the k-th mastery state

Observations

z—an individual's response vector z_1, z_2,..., z_N, where $z_i \in (0, 1)$

An estimate of an examinee's true classification is formed using the priors and observations. By Bayes Theorem,

$$P(m_k \mid \mathbf{z}) = c\, P(\mathbf{z} \mid m_k) P(m_k) \qquad (1)$$

The posterior probability $P(m_k \mid \mathbf{z})$ 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. For each examinee, there are K probabilities, one for each mastery state. The normalizing constant in formula (1),

$$c = \frac{1}{\sum_{k=1}^{K} P(\mathbf{z}|m_k)\, P(m_k)}$$

assures that the sum of the posterior probabilities equals 1.0.

Assuming local independence,

$$P(\mathbf{z}|m_k) = \prod_{i=1}^{N} P(z_i|m_k) \qquad (2)$$

That is, the conditional 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. We will naively assume the assumption is true and proceed with our analysis.

The decision is usually made to classify the individual into the most likely group. If nonuniform priors are

used, the approach is called the maximum a posteriori (MAP) decision criterion. If the priors are uniform, this is simply the maximum likelihood decision criterion. MAP itself is a subset of a third more general approach called the Bayes risk criterion, which incorporates weights for false negative and false positives.

Measurement Decision Theory Adaptive Testing Using Information Gain

Rudner (2002, 2009) describes three approaches to adaptive item selection using decision theory—information gain, minimum expected cost, and maximum discrimination. In those analyses, there was little difference in the effectiveness of information gain and minimum expected cost, with both performing better than maximum discrimination. In this paper, only information gain was applied.

The concept of *information gain* builds on the widely used measure of information from information theory, Shannon entropy (Shannon, 1948; see Cover & Thomas, 1991):

$$H(S) = \sum_{k=1}^{K} -p_k \log_2 p_k \qquad (3)$$

where p_k is the proportion of S belonging to class *k*. Base 2 is used for the logarithm because the formula was originally developed for binary data. Entropy can be viewed as a measure of the uniformity of a distribution and has a maximum value when $p_k = 1/K$ for all k. Since the goal is to have a peaked distribution of $P(m_k)$, one wants the lowest possible value of H(S). Thus, one should next select the item that has the greatest expected reduction in entropy, i.e. $H(S_0) - H(S_i)$, where $H(S_0)$ is the current entropy and $H(S_i)$ is the expected entropy after administering item *i*. This expected entropy is the sum of the weighted conditional entropies of the classification probabilities that correspond to a correct and an incorrect response:

$$H(S_i) = p(z_i = 1)\, H(S_i \mid z_i = 1) + p(z_i = 0)\, H(S_i \mid z_i = 0) \qquad (4)$$

This can be computed using the following steps:

1. Compute the normalized posterior classification probabilities that result from a correct and an incorrect response to item i using formula 1.
2. Compute the conditional entropies (conditional on a right response and conditional on an incorrect response) using formula 3.
3. Weight the conditional entropies by their probabilities using formula 4.

A variant of this approach is relative entropy, which is also called the Kullback-Leibler (1951) information measure and information divergence. Chang and Ying (1996), Eggen (1999), and Lin and Spray (2000) have favorably evaluated K–L information as an adaptive testing item selection strategy.

The reader should note that after administering the most informative items, the expected entropy for all the remaining items could be greater than H(S) and result in a loss of information. That is, the classification probabilities would be expected to become less peaked. One may want to stop administering items when there are no items left in the pool that are expected to result in information gain.

Sample Computations

To illustrate the model and the information gain approach, assume three questions have been calibrated against an existing group of masters and nonmasters with the item parameters shown in Table 1. Further assume that 68 percent of the calibration sample was masters.

Table 1: Conditional Probabilities of a Correct Response, $P(z_i=1 \mid m_k)$

	Item 1	Item 2	Item 3
Masters (m_1)	.8	.9	.6
Nonmasters (m_2)	.3	.4	.2

If an individual has a response vector **z** = [1,1,0] and the prior probability of being a master is .68, then, by formula 2, the probability of their being a master is (.8 * .9 * (1 – .6)) * .68 = .196 and the probability of their being a nonmaster is (.3 * .4 * (1 – .2)) * .32 = .031. This individual is most likely a master and could be classified as such. Rather than use the odd probabilities of .196 and .031, we can normalize the probabilities. The probability of being a master given **z**, the group priors, and the item parameters is .196 /

(.196 + .031) = .86 and the probability of being a nonmaster is (1 – .86) = .031/(.196 + .031) = .14.

To illustrate the information gain approach to adaptive testing, again assume the conditional probabilities in Table 1, and that the prior probability of being a master is .68. Further assume the individual responded correctly to question 1. The task then, is to determine whether question 2 or 3 is likely to provide more information and hence should be administered next. Calculations are shown in Table 2.

if both questions 1 and 2 are right is $-.93 * \log_2 (.93) + -.07 * \log_2(.07) = .38$. If the response to question 2 is incorrect then the normalized probabilities of being a master and nonmaster are .49 and .51, respectively and the entropy would be .99. Considering the probability of responding correctly to question 2 is .825, the expected entropy for question 2 given a correct response to question 1 is .825 * .38 + .175 * .99 = .48 and the information gain is .61 – .48 =.13. For question 3, the expected entropy is .55 and the

Table 2: Computation of Expected Information Gain for Items 2 and 3

	Response (z_i)	Posterior classification probabilities	Conditional entropy	$P(z_i)$	$H(S_i)$	Expected info gain
Item 2	Right	$P(m_1)$=.93	.38	.825		
		$P(m_2)$=.07			.48	.13
	Wrong	$P(m_1)$=.49	.99	.175		
		$P(m_2)$=.51				
Item 3	Right	$P(m_1)$=.94	.31	.54		
		$P(m_2)$=.06			.55	.06
	Wrong	$P(m_1)$=.73	.83	.46		
		$P(m_2)$=.27				

Based on the response to the first question, the probability of this examinee being a master is .8 * .68/(.8 * .68 + .3 * .32) = .85 and the conditional probability of being a nonmaster is .15. Given the .85 and .15, the probability of responding correctly to the unadministered question 2 is .85 * .9 + .15 * .4 = .825 and the probability of responding correctly to question 3 based on the available information is .54.

Applying formula (3) to the above for question 2, the current entropy is –.85 * log2(.85) + –.15 * log2 (.15) = .61. The item selection task, then, is to identify which question, if administered next, is expected to reduce entropy the furthest.

If the examinee does indeed respond correctly to question 2 then the nonnormalized probability of being a master and nonmaster are (.8 * .9) * .68 = .49 and (.3 * .4) *.32 = .04, respectively. Normalized, these probabilities are .93 and .07, respectively. The entropy

expected information gain, .61 – .55 = .06, is less than that of question 2. Thus, question 2 is expected to provide more information and would be selected next.

Method

The first objective of this research was to evaluate the classification accuracies of a decision-theory adaptive testing approach relative to the classification accuracy of an IRT adaptive testing. This was addressed using simulated datasets based on actual item parameters and actual thetas. In order to replicate a realistic scenario, content and exposure controls were added to both item selection algorithms.

Examinees were simulated by randomly drawing 5,000 ability estimates from a collection of quantitative scores from the Graduate Management Admission Test® (GMAT®). These theta estimates from the GMAT exam were treated as true scores, and each

examinee was classified into one of four groups, corresponding to the 0 to 40th, 40th to 60th, 60th to 80th, and 80th to 100th percentiles, based on these theta scores. These classifications were treated as the true group classification for the examinees.

IRT item parameters were based on samples of items from an operational GMAT pool. For this part of the study, the a priori conditional group probabilities of a correct response for measurement decision theory were computed as the area of the IRT curve within each score band weighted by the proportion of examinees at each theta level within the interval.

The IRT adaptive testing approach started with a test question of about median difficulty. Item responses were simulated using Birnbaum's (1968) three-parameter IRT model. Using the true theta score, the examinee's probability of a correct response was compared to a random number between 0 and 1 and coded as either responding correctly or incorrectly.

After each question, ability was estimated to the nearest of 13 quadrature points using maximum likelihood estimation. Rather than having wild swings in ability estimates and item difficulties, a maximum step size of ±3 quadrature points was used for the all-right or all-wrong response vectors. The next item to be administered was randomly selected from the five unadministered items with the right content whose difficulties were closest to the interim ability estimate. The final ability estimate was computed as the maximum likelihood point estimate and then compared to the cut scores to determine the final group classification. The random selection provided exposure control; a bin-structure-based item selection algorithm provided content balancing. Accuracy was defined as the proportion of correct classifications.

The MDT adaptive testing approach used information gain as previously outlined. Testing started with a randomly selected item from the first bin with near median difficulty. As with the IRT simulation, responses were based on the three-parameter model using the true theta scores. After each question was administered and scored, the interim group classification likelihoods were computed using uniform priors. The next item to be administered was randomly chosen from the five unadministered items with the highest expected information gain. The item exposure and content balance were implemented in the same way as the IRT CAT approach. The final group classification was the group with the maximum likelihood.

The item bank size was varied, using pools of 200, 300 and 600 questions. Test length was also varied, using lengths of 10, 15, 20, 25, and 30 questions.

Key statistics for this dataset are shown in Table 3.

Table 3: Key Statistics for the IRT and MDT Adaptive Testing Comparison

Mean θ	0.6760
SD θ	1.2381
Cut 1	0.3645
Cut 2	0.9975
Cut 3	1.7241

The second objective was to evaluate classification accuracy as a function of calibration sample size. IRT can require 1,000 or more examinees in order to obtain quality item parameters. Assuming random selection, how many examinees are needed to obtain satisfactory MDT item parameter estimates?

A second analysis was conducted for the MDT approach using calibration sample sizes of 100, 200, 500, 1,000, and 2,000 examinees to obtain the a priori conditional probabilities of a correct response for examinees in each of the four groups. The 100 and 200 sample size conditions represent minimal sample sizes that might be used by smaller testing programs. Because a calibration sample size of 1,000 to 2,000 is needed to make fine distinctions along three parameters, it was felt these samples would yield close to population values for the simpler MDT model.

Examinee samples were again drawn from GMAT thetas and responses were again generated by comparing a random value to the response probabilities using the three-parameter model. Examinees were grouped into their true classification and the proportions of examinees within each group that responded correctly were used as the a priori conditional group probabilities. In order to prevent individual questions from dominating the analysis, minimum and maximum conditional probabilities of .02 and .98 were used. The MDT Calibrate software package (Rudner, 2009) was used. Only one bank of a random sample of 200 questions was used in this second analysis.

Once the item bank was calibrated, the examinations were administered adaptively using the information gain approach. There was no overlap between the calibration samples and the simulees that were administered these MDT CAT examinations.

Results

Adaptive Testing Accuracy

Classification accuracy of IRT and MDT adaptive testing as a function of test length and pool size are shown in Table 4. As expected, as test length increases so does classification accuracy for both the IRT and the MDT approaches. Pool size made little difference. When the task is simply to classify, it appears that large pools are not needed. There were also minimal differences between MDT and IRT. The simpler MDT model was just as accurate as the more complicated IRT model in terms of classifying examinees into one of four groups.

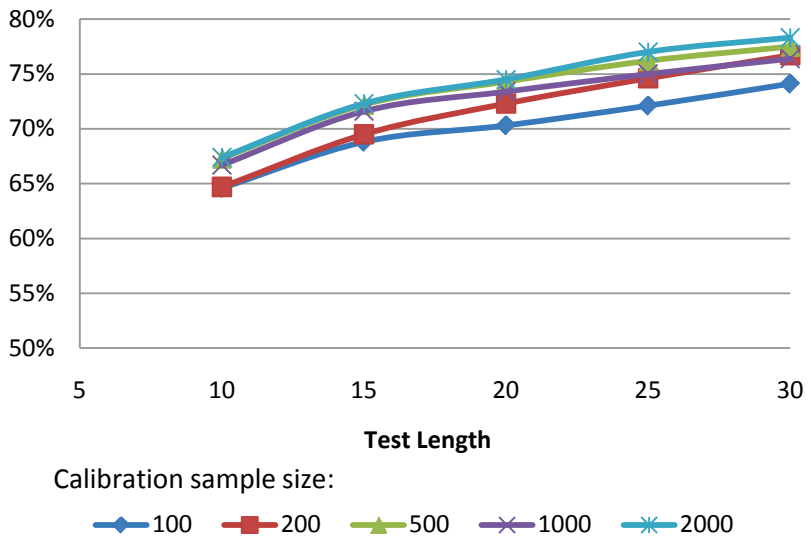
Table 5 and Figure 1 show MDT classification accuracy as a function of test length and calibration sample size. It is readily apparent that for any given test length, there is little difference in accuracy between an item bank calibrated on 100 examinees and one calibrated on 2,000. While accuracy does increase with sample size, the increases are small. Test length is a far more important predictor of accuracy.

Table 4: Classification Accuracies of IRT and MDT Adaptive Testing as a Function of Test Length and Pool Size

Pool size	Model	Test length				
		10	15	20	25	30
600	IRT	.657	.705	.736	.766	.777
	MDT	.709	.746	.775	.792	.804
300	IRT	.642	.700	.731	.760	.783
	MDT	.694	.732	.754	.776	.787
200	IRT	.652	.708	.740	.756	.783
	MDT	.675	.722	.742	.753	.779

Table 5: Proportion of Examinees Classified Correctly as a Function of Test Length and Calibration Sample Size					
Calibration size	**Test length**				
	10	**15**	**20**	**25**	**30**
100	.646	.688	.703	.721	.741
200	.647	.695	.723	.746	.767
500	.673	.722	.743	.762	.775
1,000	.667	.716	.734	.750	.764
2,000	.674	.723	.745	.770	.783

Figure 1. Classification accuracy as a function of test length and calibration sample size



Discussion

This research shows that the simpler measurement decision theory model can be as effective as the popular three-parameter IRT model in terms of correctly classifying examinees into one of four groups. Using real item parameters and real theta scores, the IRT and MDT CAT algorithms provide for extremely similar percentages of correct classifications. A second analysis shows that calibration sample sizes as small as 100 can yield high classification accuracies in MDT CAT.

The authors believe that MDT clearly is a simple yet powerful and widely applicable model. The advantages of this model are many—the model yields accurate mastery state classifications, can incorporate a small item pool, is simple to implement, requires little pre-testing, is applicable to criterion referenced tests, can be used in diagnostic testing, can be adapted to yield classifications on multiple skills, can employ sequential testing and a sequential decision rule, and should be easy to explain to nonstatisticians.

The model can be used as the routing mechanism for intelligent tutoring systems. Items could be piloted with a few examinees to vastly improve end-of-unit examinations. Certification examinations could be created for specialized occupations with a limited number of practitioners available for item calibration. Short tests could be prepared for physicians or teachers to help make tentative decisions. A small collection of items from one test could be embedded in another test to yield meaningful cross-regional information.

While an attractive feature of the model is that it can work with small calibration samples, the use of such small samples can be very problematic. Even if the small sample is representative of the population, the a priori item parameters may not be calibrated adequately. With 100 examinees and four groups, an average cell size of 25 people is used to estimate population parameters. The resultant parameters may not be very stable. Coupling poor item parameters with a very short test can be expected to lead to substantial classification error.

A key question not addressed here is the local independence assumption. We naively assumed that the responses to a given item are unaffected by responses to other items. While the local independence is often ignored in measurement and one might expect only minor violations, its role in measurement decision theory is not fully understood. The topic has been investigated in the text classification literature. Despite very noticeable and very serious violations, naive Bayes classifiers perform quite well. Domingos and Pazzani (1997) show that strong attribute dependencies may inflate the classification probabilities while having little effect on the resultant classifications. They argue that naive Bayes classifiers have broad applicability.

One can view MDT as a crude IRT model that uses a step function rather than a continuous function. Indeed MDT parameters in the first investigation used IRT probabilities weighted by frequency. Investigating the likelihood for a limited number of classifications is the same mathematical concept as examining the likelihood of each value on the theta continuum. Viewed this way, one must question whether the use of observed proportions as the MDT a priori item parameters is as good as using an underlying population model such as in IRT. There is a logical disconnect in that regard.

Another issue with the MDT model is that, at present, there is no accepted way to assess reliability within the model. There is statistical literature on consistency and asymptotic normality of maximum likelihood estimates that might prove helpful.

This paper only examined fixed length tests. Thompson (2007) discusses variable length computerized classification testing where the goal is to reduce classification error while using as few questions as possible. Testing continues until the probability of new questions changing the most likely classification category is low. Wald's (1947) sequential probability ratio testing (SPRT), perhaps the foremost application of decision theory, has been used by numerous researchers (e.g., Lin & Spray, 2000; Parshall, Spray, Kalohn, & Davey, 2006; Reckase, 1983) to define termination rules. An interesting line of research would be to apply SPRT to a MDT adapted test. Rudner (2002) reports one such study.

The research questions are numerous. How can the model be extended to multiple rather than dichotomous item response categories? How can bias be detected? How effective are alternative adaptive testing and sequential decision rules? Can the model be effectively extended to 30 or more categories and provide a rank ordering of examinees? How can we make good use of the fact that the data is ordinal? Are there new item analysis procedures that can improve measurement decision theory tests? How can the model be best applied to criterion-referenced tests assessing multiple skills, each with a few items? Why are minimum cost and information gain so similar? How can different cost structures be effectively employed? How can items from one test be used in another? How does one equate such tests?

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC Research and Development Department at research@gmac.com.

Notes

An earlier version of this paper was presented as a keynote address at the meeting of the International Association on Computer Adaptive Testing in Arnhem, Netherlands, June 2010.

The views and opinions expressed in this paper are those of the authors. No endorsement by any organization should be inferred or implied.

An interactive tutorial is available online at http://pareonline.net/sup/mdt/. The tutorial allows you to vary the results of the a priori parameters, the examinee's response pattern, and the cost structure. Various rules for classifying an examinee and sequencing items are then presented along with the underlying calculations.

Software for generating, calibrating, and scoring measurement decision theory data is available at http://pareonline.net/sup/mdt/MDTToolsSetup.exe Updated April 2010, this is version .895. No support is provided.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory*. New York, NY: Wiley.

Cronbach, L.J., & Gleser, G.C. (1957). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Domingos P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130. Retrieved from http://citeseer.nj.nec.com/48.html.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249–61.

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

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

Kingsbury, G.G., & Weiss, D.J. (1979). *An adaptive testing strategy for mastery decisions*. (Research Report 79–05). Minneapolis: University of Minnesota, Psychometric Methods Laboratory.

Kingsbury, G. G., & Weiss, D. J. (1981). *A validity comparison of adaptive and conventional strategies for mastery testing*. (Research Report 81–3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York,NY: Academic Press.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(2), 367–86.

Lin, C.-J., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. (Research Report 2000-8). Iowa City, IA: ACT, Inc.

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

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Mediated and User-Adapted Interaction, 5*, 253–282.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–255). New York, NY: Academic Press.

Rudner, L.M. (2002, April). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Rudner, L.M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation,* 14(8). Retrieved from http://pareonline.net/getvn.asp?v=14&n=8.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656. Retrieved from http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, *16*, 65–76.

Spray, J. A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*(4), 405–14.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, 12(1). Retrieved from http://pareonline.net/getvn.asp?v=12&n=1.

van der Linden, W. J., & 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.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

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