Paper presented at the annual meeting of the
American Educational Research Association
New Orleans, LA April 1-5, 2002

An Examination of Decision-Theory Adaptive Testing Procedures

Lawrence M. Rudner
University of Maryland, College Park

Rather than make a classification decision (pass/fail, below basic/basic/proficient/advanced) for an individual after administering a fixed number of items, it is possible to sequentially select items to maximize information, update the estimated classification probabilities and then evaluate whether there is enough information to terminate testing. In measurement this is frequently called adaptive or tailored testing. In statistics, this is called sequential testing.

Much of the research on adaptive testing has centered on the use of Item Response Theory as the underlying model. While very attractive and widely used, IRT is fairly complex and relies on several fairly restrictive assumptions. Decision theory provides an alternative underlying model for sequential testing and was probably first applied as an adaptive procedure for educational testing by Ferguson (1969). Key articles in the mastery testing literature of the 1970s employed decision theory (Hambleton and Novick, 1973; Huynh, 1976; van der Linden and Mellenbergh, 1977). Lewis and Sheehan (1990), Kingsbury and Weiss (1983), Reckase (1983), and Spray and Reckase (1996) have used decision theory adaptively select items and testlets to determine when to stop testing. Notable articles by Macready and Dayton (1992), Vos (1999), and Welch and Frick (1993) illustrate the less prevalent item-level application of decision theory examined in this paper. Decision theory is very attractive due to it's simplicity, wide acceptance in many fields, lack of assumptions, robustness, and computational ease.

This paper presents adaptive testing procedures using measurement decision theory and then compares those procedures with IRT in terms of classification accuracy using two sets of simulated item response data.

Overview

Decision Theory

The objective is to form a best guess as to the mastery state (classification) of an individual examinee based on the examinee's item responses, *a priori* item information, and *a priori* population classification proportions. Thus, the 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 mastery states, that take on values m_k. In the case of pass/fail testing, there are two possible states and K=2. One usually knows, *a priori*, the approximate proportions for the population of all examinees in each mastery state.

The second component is a set of items for which the probability of each possible observation, usually right or wrong, given each mastery state is also known *a priori*,

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 mastery states 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 guess the mastery state and there will be D=K decisions. With adaptive or sequential testing, a

decision to continue testing will be added and thus there will be D=K+1 decisions. Each decision will be denoted d_k.

Testing starts with the proportion of examinees in the population that are in each of the K categories and the proportion of examinees with each category that respond correctly. The population proportions can be determined a variety of ways, including from prior testing, transformations of existing scores, existing classifications, and judgement. In the absence of information equal priors can be assumed. The proportions that respond correctly can be derived from a small pilot test involving examinees that have already been classified or transformations of existing data. Once these sets of priors are available, the items are administered, responses (z_1, z_2, ... z_N) observed, and then a classification decision, d_k, is made based on the responses to those items.

In this paper, pilot test proportions are treated as probabilities and the following notation is used:

Priors

$P(m_k)$ - the probability of a randomly selected examinee having a mastery state m_k

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

Observations

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

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

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

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. 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,

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

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

In this paper, each response is either right (1) or wrong (0) and $P(z_1=0|m_k) = 1- P(z_1=1|m_k)$. The model is illustrated with an examination of two possible mastery states m_1 and m_2 and two possible decisions d_1 and d_2 which are the correct decisions for m_1 and m_2, respectively. The examples use a three-item test with the item statistics shown in Table 1 with prior classification probabilities of $P(m_1)=0.2$ and $P(m_2)=1-P(m_1) = 0.8$.

Table 1: Conditional probabilities of a correct response, $P(z_i=1|m_k)$

	Item 1	Item 2	Item 3
Masters (m_1)	.6	.8	.6
Non-masters (m_2)	.3	.6	.5

Suppose an examinee responded to all three items and had a response vector **z**= [1,1,0]. The probability of **z** if the examinee is a master is .6*.8*.4 = .19, and .09 if he is a non-master. That is, $P(\mathbf{z}|m_1)$=.19 and $P(\mathbf{z}|m_2)$=.09. If we multiply by the priors and normalize, then $P(m_1|\mathbf{z})$=(.2*.19)/ (.2*.19+.8*.09) = .35 and $P(m_2|\mathbf{z})$=.65. Thus we would make the decision to classify this examinee as a non-master. Throughout this paper, we select the category with the maximum a posteriori probability as the most likely category.

Sequential Testing

To illustrate adaptive testing in this context, assume the examinee responded correctly to the first item in Table 1. The task is to select which of the two remaining items to administer next. After each item is administered, the posterior classification probabilities $p(m_k|\mathbf{z})$ are treated as updated prior probabilities $p(m_k)$ and used to help identify the next item to be administered.

After responding correctly to the first item, the current updated probability of being a master is .6*.2/(.6*.2+.3*.8) = .33 and the probability of being a non-master is .66 from formula (1).

The current probability of responding correctly is

$$P(z_i=1) = P(z_i=1|m_1)P(m_1) + P(z_i=1|m_2)P(m_2) \qquad (2)$$

Applying (2), the current probability of correctly responding to item 2 is $P(z_2=1)$=.8*.33+ .6*.66 = .66 and, for item 3, $P(z_3=1)$=.53. The following are some approaches to identify which of these two items to administer next.

Using the classification probabilities and the probabilities of responding correctly to each of the items remaining after a correct response to item 1, the following three adaptive testing approaches are examined:

Minimum expected cost

A significant advantage of the decision theory framework is that one can incorporate decision costs into the analysis. By this criteria, costs are assigned to each correct and incorrect decision so the total average costs can be minimuized. For example, false negatives may be twice as bad as false positives. If c_{ij} is the cost of deciding d_i when m_j is true, then the expected or average cost B is

$$B=(c_{11}\,P(d_1|m_1) + c_{21}\,P(d_2|m_1))\,P(m_1) + (c_{12}\,P(d_1|m_2) + c_{22}\,P(d_2|m_2))\,P(m_2)$$

If $c_{11}=c_{22}=0$ (no cost for a correct decision) then

$$B=c_{21}\,P(d_2|m_1)\,P(m_1) + c_{12}\,P(d_1|m_2)\,P(m_2) \qquad (2)$$

In the binary decision case, the probabilities of making a wrong decision are one minus the probabilities of making a right decision. The probabilities of making a right decision is, by definition, the posterior probabilities given in (1). Thus, with $c_{12}=c_{21}=1$, the current Bayes cost is B=1*(1-.33)*.33 + 1*(1-.66)*.66 = .44.[1]

Minimum expected cost is often associated with sequential testing and has been applied to measurement problems by Lewis and Sheehan (1990), Macready and Dayton (1992), Vos (1999), and others.

The following steps can be used to compute the expected cost for each item.

1. Assume for the moment that the examinee will respond correctly. Compute the posterior probabilities using (1) and then costs using (2).
2. Assume the examinee will respond incorrectly. Compute the posterior probabilities using (1) and then costs using (2).
3. Multiply the cost from step 1 by the probability of a correct response to the item.
4. Multiply the cost from step 2 by the probability of an incorrect response to the item.
5. Add the values from steps 3 and 4.

Thus, the expected cost is the sum of the costs of each response weighted by the probability of that response. If the examinee responds correctly to item 2, then the posterior probability of being a master will be (.8*.33)/(.8*.33+.6*.66)=.40 and the associated cost will be 1*(1-.40)*.40+1*(1-.60)*.60 =.48. If the examinee responds incorrectly, then the posterior probability of being a master will be (.2*.33)/(.2*.33+.4*.66)=.20 and the associated cost will be 1*(1-.20)*.20+1*(1-.80)*.80 =.32. Since the probability of a correct response from (5) is .66, the expected cost for item 2 is .66*.48+(1-.66)*.32 = .42.

The cost for item 3 is .47 if the response is correct and .41 if incorrect. Thus, the expected cost for item 3 is .53*.47+(1-.53)*.41 = .44. Since item 2 has the lowest expected cost, it would be administered next.

Information Gain

This entire essay is concerned with the use of prior item and examinee distribution information in decoding response vectors to make a best guess as to the mastery states of the examinees. The commonly used measure of information from information theory (see Cover and Thomas, 1991), Shannon (1948) entropy, is applicable here[2]:

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

where p_k is the proportion of S belonging to class k. Entropy can be viewed as a measure of the uniformness of a distribution and has a maximum value when $p_k = 1/K$ for all k. The goal is to have a peaked distribution of $P(m_k)$ and to 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*, i.e. the sum of the weighted conditional entropies of the classification probabilities that correspond to a correct and to an incorrect response

---

[1] The generalized formula for cost in this context is $B = \sum_{i=1}^{K}\sum_{j=1}^{K} c_{ij} P(m_j|\mathbf{z}) P(m_i|\mathbf{z})$.

[2] Log base 2 because entropy is in terms of bits of information.

$$H(S_i) = p(z_i=1)\ H(S_i|z_i=1) + p(z_i=0)\ H(S_i|z_i=0) \quad (4)$$

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 (1).
2. Compute the conditional entropies (conditional on a right response and conditional on an incorrect response) using (3).
3. Weight the conditional entropies by their probabilities using (4).

Table 2 shows the calculations with the sample data.

Table 2: Computation of expected classification entropies for items 2 and 3.

	Response (z_i)	**Posterior classification probabilities**	**Conditional entropy**	**$P(z_i)$**	**$H(S_i)$**
Item 2	Right	$P(m_1)$=.40	.97	.66	.89
		$P(m_2)$=.60			
	Wrong	$P(m_1)$=.20	.72	.33	
		$P(m_2)$=.80			
Item 3	Right	$P(m_1)$=.38	.96	.53	.92
		$P(m_2)$=.62			
	Wrong	$P(m_1)$=.29	.87	.47	
		$P(m_2)$=.71			

After administering the first item, $P(m_1)$=.33, $P(m_2)$=.66, and $H(S_0)$=.91. Item 2 results in the greatest expected entropy gain and should be administered next.

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), Lin and Spray (2000) have favorably evaluated K-L information as an adaptive testing strategy.

The reader should note that, the expected entropy after administering item 3 would be greater than $H(S_0)$ and result in a loss of information. That is, the classification probabilities are expected to become less peaked should item 3 be administered. As a result, this item shouldn't be considered as a candidate for the next item. One may want to stop administering items when there are no items left in the pool that are expected to result in information gain, although the author does not know of any study that has investigated this logical termination rule.

Maximum Discrimination

When the purpose of the test is to classify examinees, the optimal IRT item selection strategy is to sequence items based on their information at the cut score (Spray and Reckase, 1994). The analog here is to select the item that

best discriminates between the two most likely mastery state classifications. One such index is

$$M_i = \left| \log \frac{p(z_i = 1 | m_k)}{p(z_i = 1 | m_{k+1})} \right|$$

where m_k and m_{k+1} are currently the two most likely mastery states. In the binary case, m_k and m_{k+1} are always m_1 and m_2 and the item order is the same for all examinees. Here, item 2 would be selected as the next item to be administered.

Sequential Decisions

A natural extension to sequential testing using decision theory is to apply Wald's (1947) well-known sequential decision rule, the sequential probability ratio test (SPRT, pronounced spurt). SPRT for K multiple categories can be summarized as

$$d_k \ if \ \frac{P(m_k)}{P(m_{k-1})} > \frac{1-\beta}{\alpha} \quad for\ k = K$$

$$d_k \ if \ \frac{P(m_{k+1})}{P(m_k)} < \frac{\beta}{1-\alpha} \quad for\ k = 1$$

$$d_k \ if \ \frac{P(m_k)}{P(m_{k-1})} > \frac{1-\beta}{\alpha} \ and \ \frac{P(m_{k+1})}{P(m_{k-1})} < \frac{\beta}{1-\alpha} \quad for\ k = 2,3,\ldots K-1$$

where the $P(m_j)$'s are the normalized posterior probabilities, á is the acceptable error rate, and â is the desired power. If the condition is not meet for any category k, then testing continues. In the measurement field, there is a sizeable and impressive body of literature illustrating that SPRT is very effective as a termination rule for IRT-based computer adaptive tests (c.f. Reckase, 1983; Spray and Reckase, 1994, 1996; Lewis and Sheehan, 1990; Sheehan and Lewis, 1992)

Method

The objective of this research is to evaluate the classification accuracies of different decision-theory sequential testing approaches relative to the classification accuracy of IRT adaptive testing. This was addressed using large numbers of simulated datasets.

Examinees were simulated by randomly drawing an ability value from normal N(0,1) and uniform (-2.5, 2.5) distributions and classifying each examinee based on this true score. Item responses were then simulated using Birnbaum's (1968) three parameter IRT model. For each item and examinee, the examinee's probability of a correct response is compared to a random number between 0 and 1 and coded as either responding correctly or incorrectly to the item. The items parameters were based on samples of items from the 1999 Colorado State Assessment Program fifth grade mathematics test (Colorado State Department of Education, 2000) and the 1996 National Assessment of Educational Progress State Eighth Grade Mathematics Assessment (Allen, Carlson, and Zelenak, 2000).

Key statistics for each simulated test are given in Table 3

Table 3: Descriptive statistics for simulated tests

	Simulated test	
	CSAP	**State NAEP**
No of items in item pool	54	139
Mean *a*	.78	.94
Mean *b*	-1.25	.04
Mean *c*	.18	.12
Reliability for N(0,1) sample	.83	.95
Cut score(s)	-.23	-.23. .97 1.65
Mastery states	2	4

For each test, a calibration sample of 1,000 examines and a separate trial data set of 10,000 examinees were generated. The calibration sample was used to compute the measurement decision theory priors - the probabilities of a randomly chosen examinee being in each of the mastery states and the probabilities of a correct response to each item given the mastery state.

The simulated state-NAEP data set drew from a large number of items and a very reliable test. The cut scores corresponded to the IRT theta levels that delineate state-NAEP's Below Basic, Basic, Proficient, and Advanced ability levels. The relatively small cell size for the Advanced level and the use of four mastery state classifications provide a good test for measurement decision theory.

The CSAP is a shorter test of lower reliability and the sample of items has mean difficulty (mean b) well below the mean examinee ability distribution. Classification categories are not reported for CSAP. The mastery/non-mastery cut score used in the study was arbitrarily selected to correspond to the 40th percentile.

Using these common datasets, items were selected and mastery states were predicted using three sequential testing approaches (minimum cost, information gain, and maximum discrimination) and the baseline IRT approach. The costs of deciding d_i when m_j is true were set at |i-j| for all i,j. Under the IRT approach, the items with the maximum information at the examinee's true score were selected without replacement. While this is not feasible in real life, it presents a favorable scenario for item response theory.

Accuracy was defined as the proportion of correct state classifications. To determine the correct state classification, the examinee's true score was compared to the cut scores. To determine the observed classification, maximum *a posterior* (MAP) probabilities were used with the decision theory approaches and thetas estimated using the Newton-Raphson iteration procedure outlined in Baker (2001) were used with the IRT approach.

The reader should note that measurement decision theory approaches do not incorporate any information concerning how the data were generated, or any information concerning the distribution of ability within a category.

The simulation compares the best-case scenario for measurement decision theory against the best case for IRT. The examinees in the calibration sample are classified without error, thus providing extremely accurate priors for applying decision theory. The IRT baseline was also designed to provide a best-case scenario for that model. The data fit the IRT model perfectly. Adaptive IRT testing used the items with the most information at the (usually unknown) true scores to optimally sequence the test items.

Results

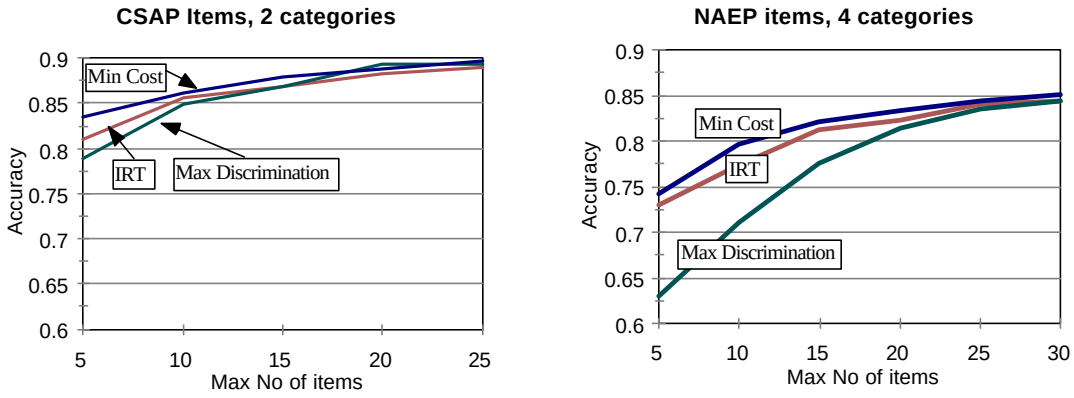
Adaptive Testing Accuracy

The results are shown in Table 4 and graphed in Figure 1. The minimum cost and information gain decision theory approaches consistently out-performed the IRT approach in terms of classification accuracy. The fact that the classification accuracies for these two decision theory methods are almost identical implies that they tend to select the same items. Optimized to make fine distinctions across the ability scale, the IRT approach is less effective if one is interested in making coarser mastery classifications. The simple maximum discrimination approach was not as effective as the others, but was reasonably accurate.

Table 4: Accuracy of sequential testing methods as a function of maximum test length

Max No of items	IRT	Decision Theory Approaches		
		Max Disc	Min Cost	Info Gain
CSAP items, 2 categories				
5	.810	.789	.836	.836
10	.856	.850	.862	.863
15	.869	.868	.880	.879
20	.882	.893	.889	.886
25	.890	.893	.897	.898
State NAEP items, 4 categories				
5	.730	.630	.743	.742
10	.774	.711	.797	.793
15	.812	.775	.822	.818
20	.824	.815	.833	.832
25	.840	.835	.844	.844
30	.845	.845	.852	.852

Figure 1: Accuracy of sequential testing methods as a function of maximum test length



Sequential decisions

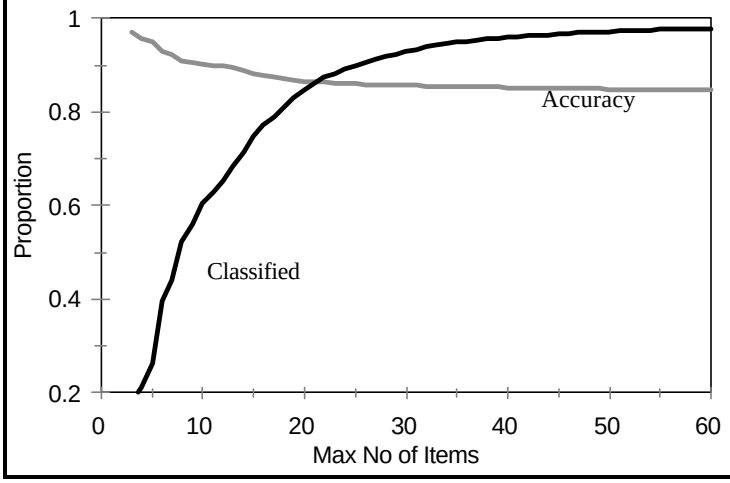
After each item was administered above, Wald's SPRT was applied to determine whether there was enough information to make a decision and terminate testing. Power and error rate where set to á=â= .05. Table 5 shows the proportion of examinees for which a classification decision could be made, the percent of those examinees that were correctly classified, and the mean number of administered items as a function of maximum test length using items from state-NAEP. With an upper limit of only 15 items, for example, some 75% of the examinees were classified into one of the 4 NAEP score categories. A classification decision could not be made for the other 25%. Eighty-eight percent of the classified examinees were classified correctly and they required an average of 9.1 items. SPRT was able to quickly classify examinees at the tails of this data with an underlying normal distribution.

Table 5: Proportion of examinees classified using SPRT, information gain, and state-NAEP items, the accuracy of their classifications, and the mean number of administered items as a function of the maximum number of administered items.

Max No of items	Proportion Classified	Accuracy	Mean # of items
5	0.260	0.948	4.6
10	0.604	0.902	7.4
15	0.749	0.880	9.1
20	0.847	0.865	10.2
25	0.899	0.860	10.8
30	0.928	0.857	11.3
40	0.960	0.852	11.8
50	0.972	0.849	12.2
100	0.988	0.847	13.0

The proportions classified and the corresponding accuracy as a function of the maximum number of items administered are shown in Figure 2. The proportion classified curve begins to level off after about a test size limit of 30 items. Accuracy is fairly uniform after a test size limit of about 10 or 15 items.

Figure 2: Proportion of examinees classified and the accuracy of those classifications as a function of the maximum number of administered items (state-NAEP items, four latent states, sequential testing using information gain, sequential decisions using SPRT).



Discussion

In their introduction, Cronbach and Gleser (1957) argue that the ultimate purpose for testing is to arrive at qualitative classification decisions. Today's decisions are often binary, e.g., whether to hire someone, whether a person has mastered a particular set of skills, whether to promote an individual. Multi-state conditions are common in state assessments, e.g., the percent of students that perform at the basic, proficient, or advanced level. The simple measurement model presented in this paper is applicable to these and other situations where one is interested in categorical information.

This research examined three ways to adaptively, or sequentially, administer items using the model. The traditional decision theory sequential testing approach, minimum cost, was notably better than the best-case possibility for item response theory. Information gain, which is based on entropy and comes from information theory, was almost identical to minimum cost. A simpler approach using the item that best discriminates between the two most likely classifications also fared better than IRT, but not as well as information gain or minimum cost. The research also showed that with Wald's SPRT, large percentages of examinees can be accurately classified with very few items. With only 25 sequentially selected items, for example, some 90% of the simulated state-NAEP examinees were classified with 86% accuracy.

This is clearly 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 non-statisticians.

It is the author's hope that this research will capture the imagination of the research and applied measurement communities. The author can envision wider use of the model as the routing mechanism for intelligent tutoring systems. Items could be piloted with a few number of 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 teachers to help make tentative placement and advancement decisions. A small collection of items from a one test, say state-NAEP, could be embedded in another test, say a state assessment, to yield meaningful cross-regional information.

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 in addition to advantages in terms of simplicity, learning speed, classification speed, storage space and incrementality.

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? How can the concept of entropy be employed in the examination of tests? 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 number of 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? The author is currently investigating the applicability of the model to computer scoring of essays. In that research, essay features from a large pilot are treated as items and holistic scores as the mastery states.

Notes

1. 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.

2. An interactive tutorial is available on-line at http://ericae.net/mdt/. The tutorial allows you to vary the results of the pilot, 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.

References

Allen, N.L., J.E. Carlson, and C. A. Zelenak (2000). *The NAEP 1996 technical report*. Washington, DC: National Center for Educational Statistics. Available online: http://nces.ed.gov/nationsreportcard/pubs/main1996/1999452.asp

Baker, F. (2001). *The Basics of item response theory*. Second edition. College Park: MD: ERIC Clearinghouse on Assessment and Evaluation.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Colorado State Department of Education (2000). Colorado Student Assessment Program (CSAP), technical report, grade 5 mathematics. Available online: http://www.cde.state.co.us/cdeassess/download/pdf/as_csaptech5math99.pdf

Cover, T.M. and J.A. Thomas, *Elements of information theory*. New York: Wiley, 1991.

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

Domingos P. and M. Pazzani (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103--130. Available online: 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.

Ferguson, R.L. (1969). The development, implementation, and evaluation of a computer assisted branched test for individually prescribed instruction. Doctoral dissertation. University of Pittsburgh, Pittsburgh, PA.

Hambleton, R. and 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. (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: Academic Press.

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

Lewis, C. and 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. ACT Research Report Series.

Macready, G. and 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: Academic Press.

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

Sheehan, K. and Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, v16 n1 p65-76 Mar 1992

Spray, J. A. and 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.

Spray, J. A. and Reckase, M. D. (1994). The selection of test items for decision making with a computer adaptive test. Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).

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.

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.

h:\lr\mastery\aera2c.wpd