8/22/01

Measurement Decision Theory

Lawrence M. Rudner
University of Maryland, College Park

This paper describes and evaluates a decision theory measurement model that can be used to classify examinees based on their item response patterns. The model has a simple framework that starts with the conditional probabilities of examinees in each category or mastery state responding correctly to each item. An overview of measurement decision theory and its key concepts are presented and illustrated using a binary classification (pass/fail) test and a sample three-item test. The research presents an evaluation of the model by examining the: (1) classification accuracy of tests scored using measurement decision theory; (2) differential sequential testing procedures by comparing classification accuracy against that of the best case item response theory scenario; (3) the number of items needed to make a classification; and (4) the number of examinees needed to calibrate measurement decision theory item parameters satisfactorily. The research shows that a large percentage of examinees can be classified accurately with very few items and that surprisingly few examinees are needed for calibration.

Classical measurement theory and item response theory are concerned primarily with rank ordering examinees across an ability continuum. Those models are concerned, for example, with differentiating examinees at the 90th and 92nd percentiles. But one is often interested in classifying examinees into one of a finite number of discrete categories, such as pass/fail or proficient/basic/below-basic. This is a simpler outcome and a simpler measurement model should suffice. This paper presents and evaluates the use of decision theory as a tool for classifying examinees based on their item response patterns.

Measurement decision theory requires only one key assumption - that the items are independent. Thus, the tested domain does not need to be unidimensional, examinee ability does not need to be normally distributed, and one doesn't need to be concerned with the fit of the data to a theoretical model as in item response theory (IRT) or in most latent class models. The model is attractive as the routing mechanism for intelligent tutoring systems, for end-of-unit examinations, for adaptive

testing, and as a means of quickly obtaining the classification proportions on other examinations. Very few pilot test examinees are needed and, with very few items, classification accuracy can exceed that of item response theory. Given these attractive features, it is surprising that the model has not attracted wider attention within the measurement community.

Developed by Wald (1947), first applied to measurement by Cronbach and Gleser (1957), and now widely used in engineering, agriculture, and computing, decision theory provides a simple model for the analysis of categorical data. Isolated elements of decision theory have appeared sporadically in the measurement literature. 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) and should be re-examined in light of today's measurement problems. Lewis and Sheehan (1990) and others used decision theory to adaptively select items. Kingsbury and Weiss (1983), Reckase (1983), and Spray and Reckase (1996) have used decision theory to determine when to stop testing. Most of the research to date has applied decision theory to testlets or test batteries or as a supplement to item response theory and specific latent class models. Notable articles by Macready and Dayton (1992), Vos (1997), and Welch and Frick (1993) illustrate the less prevalent item-level application of decision theory examined in this paper.

As background to the presented research, an overview and key concepts of the measurement decision theory model are presented and illustrated using a binary classification (pass/fail) case and a sample three item test. The research section presents an evaluation of the model by examining the 1) classification accuracy of tests scored using measurement decision theory, 2) different sequential testing procedures by comparing classification accuracy against that of the best case IRT scenario, 3) the number of items needed to make a classification, and 4) the number of examinees needed to satisfactorily calibrate measurement decision theory item parameters.

Background

Overview

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 will be 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

$P(z_n|m_k)$ - the probability of response z_n 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,

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

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 (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 probability of the response vector is equal to the product of the conditional probabilities of the item responses. In this paper, each response is either right (1) or wrong (0) and $P(z_1=0|m_k) = 1- P(z_1=1|m_k)$.

Three key concepts from decision theory are discussed next:

1. decision rules - alternative procedures for classifying examinees based on their response patterns,
2. sequential testing - alternative procedures for adaptively selecting items based on an individuals response pattern, and
3. sequential decisions - alternative procedures for determining whether to continue testing.

The model is illustrated here 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. Further, also based on

pilot test data, the prior classification probabilities are $P(m_1)=0.2$ and $P(m_2)=1-P(m_1) = 0.8$.In the example, the examinee's response vector is [1,1,0].

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

Decision rules

The task is to make a best guess as to an examinee's classification (master, non-master) based on the data in Table 1 and the examinee's response vector. From (2), the probabilities of the vector **z**= [1,1,0] 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 $P(\mathbf{z}|m_2)=.09$.

A sufficient statistic for decision making is the likelihood ratio

$$L(\mathbf{z}) = \frac{p(\mathbf{z}|m_2)}{p(\mathbf{z}|m_1)}$$

which for the example is $L(\mathbf{z})= .09/.19 = .47$. This is a sufficient statistic because all decision rules can be viewed as a test comparing $L(\mathbf{z})$ against a criterion value ë.

$$\begin{cases} d_2 & if\ L(\mathbf{z}) > \lambda \\ d_1 & if\ L(\mathbf{z}) < \lambda \end{cases} \qquad (3)$$

The value of ë reflects the selected approaches and judgements concerning the relative importance of different types of classification error.

Maximum-likelihood decision criterion

This is the simplest decision approach and is based solely on the conditional probabilities of the response vectors given each of the mastery states, i.e. $P(\mathbf{z}|m_1)$ and $P(\mathbf{z}|m_2)$. The concept is to select the mastery state that is the most likely cause of the response vector and can be stated as :

Given a set of item responses **z**, *make decision* d_k *if it is most likely that* m_k *generated* **z**.

Based on this criterion, one would classify the examinee as a master - the most likely classification. Using likelihood ratio testing, the decision rule is formula (3) with ë = 1.0. This criterion ignores the prior information about the proportions of masters and non-masters in the population. Equivalently, it assumes the population priors are equal. With the example, few examinees are masters, $P(m_k)$=.20. Considering that the conditional probabilities of the response vectors are fairly close, this classification rule may not result in a good decision.

Minimum probability of error decision criterion

In the binary decision case, two types of errors are possible - decide d_1 when m_2 is true or decide d_2 when m_1 is true. If one thinks of m_1 as the null hypothesis, then in terms of statistical theory, the probability of deciding a person is a master, d_1 when indeed that person is a non-master m_2, is the familiar level of significance, á and $P(d_2|m_2)$ is the power of the test, â. When both types of errors are equally costly, it may be desirous to maximize accuracy or minimize the total probability of error, *Pe.* This criterion can be stated as:

Given a set of item responses **z**, *select the decision regions which minimize the total probability of error.*

This criterion is sometimes referred to as the *ideal observer criterion*. In the binary case, *Pe = $P(d_2|m_1) + P(d_1|m_2)$* and the likelihood ratio test in (2) is employed with

$$\lambda = \frac{P(m_1)}{P(m_2)}$$

With the example, ë=.25 and the decision is d_2 - non-master.

Maximum a posteriori (MAP) decision criterion

The maximum likelihood decision criterion made use of just the probabilities of the response vector. The minimum probability of error criterion also made use of the prior classification probabilities *$P(m_1)$* and *$P(m_2)$*. MAP is another approach that uses the available information:

Given a set of item responses **z**, *decide d_k if m_k is the most likely mastery state.*

In other words,

$$\begin{cases} d_2 & if\ P(m_2|\mathbf{z}) / P(m_1|\mathbf{z}) > 1 \\ d_1 & if\ P(m_2|\mathbf{z}) / P(m_1|\mathbf{z}) < 1 \end{cases}$$

Since from equation (2), $P(m_k|\mathbf{z})=c\ P(\mathbf{z}|m_k)\ P(m_k)$, MAP is equivalent to the minimum probability of error decision criterion.

Bayes Risk Criterion

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 and then minimize the total average costs. 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

$$B=(c_{11}\,P(d_1|m_1) + c_{21}\,P(d_2|m_1))\,P(m_1) + (c_{12}\,P(d_1|m_2) + c_{22}\,P(d_2|m_2))\,P(m_2) \qquad (4)$$

and the criterion can be stated as

> *Given a set of item responses* ***z*** *and the costs associated with each decision, select* d_k *to minimize the total expected cost.*

For two mastery states, the total expected cost can be minimized using the likelihood ratio test in (2) with

$$\lambda = \frac{(c_{21} - c_{11})\,P(m_1)}{(c_{12} - c_{22})\,P(m_2)} \qquad (5)$$

This is also called the *minimum loss criterion* and the *optimal decision criterion*. If costs $c_{11}=c_{22}=0$ and $c_{12}=c_{21}=1$, then B is identical to *Pe* and this approach is identical to *minimum probability of error* and to *MAP*. With $c_{11}=c_{22}=0$ and $c_{21}=2$, $c_{12}=1$, and the sample data, ë=.50 and the decision is d_2 - non-master.

Sequential testing

Rather than make a classification decision for an individual after administering a fixed number of items, it is possible to sequentially select items to maximize information, update the estimated mastery state 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.

At each step, 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. To illustrate decision theory sequential testing, again consider the situation for which there are two possible mastery states m_1 and m_2 and use the item statistics in Table 1. Assume the examinee responded correctly to the first item and the task is to select which of the two remaining items to administer next.

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

$$P(z_i = 1) = P(z_i = 1|m_1)P(m_1) + P(z_i = 1|m_2)P(m_2) \qquad (5)$$

Applying (5), 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.

Minimum expected cost

This approach defines the optimal item to be administered next as the item with in the lowest expected cost. Equation (4) provides the decision cost as a function of the classification probabilities. If $c_{11}=c_{22}=0$ then

$$B=c_{21}\,P(d_2|m_1)\,P(m_1) + c_{12}\,P(d_1|m_2)\,P(m_2) \qquad (6)$$

In the binary decision case, the probability of making a wrong decision is one minus the probability of making a right decision and 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 (1980), Macready and Dayton (1992), Vos (1997), 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 (6).
2. Assume the examinee will respond incorrectly. Compute the posterior probabilities using (1) and then costs using (6).
3. Multiply the cost from step 1 by the probability of a correct response to the item

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

4. Multiply the cost from step 2 by the probability of an in correct 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 responses 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:

$$H(S) = \sum_{k=1}^{K} - p_k \log_2 p_k \qquad (5)$$

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.

$$H(S_0) - H(S_i) \tag{6}$$

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

$$H(S_i) = p(z_i=1)\ H(S_i|z_i=1) + p(z_i=0)\ H(S_i|z_i=0) \tag{7}$$

This can be computed using the following steps:

1. Compute the normalized posterior classification probabilities that result from a correct and to 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 (5).
3. Weight the conditional entropies by their probabilities using (7).

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 $H(S)=.91$. 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) 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.

Maximum Discrimination

Item response theory adaptive testing is most effective when the next item to be administered is the one with the most information at the cut score, the examinee's ability level, or the current estimate of the examinees ability (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

This paper has discussed procedures for making a classification decision and procedures for selecting the next items to be administered sequentially. This section presents procedures for deciding when one has enough information to hazard a classification guess. One could make this determination after each response.

Perhaps the simplest rule is the *Neyman-Pearson decision criteria* - continue testing until the probability of a false negative, $P(d_2|m_1)$, is less than a preselected value á. Suppose á= .05 was selected. After the first item, the probability of being a non-master is $P(m_1|\mathbf{z}) = .66$. If the examinee is declared a non-master, then the current probability of this being a false negative is (1-.33). Because this is more than á, the decision is to continue testing.

A variant of Neyman-Pearson is the *fixed error rate criterion* - establish two thresholds, á$_1$ and á$_2$, and continue testing until P(d_2|m_1) < á$_1$ and P(d_1|m_2) < á$_2$. Another variant is the *cost threshold criteria.* Under that approach, costs are assigned to each correct and incorrect decision and to the decision to take another observation. Testing continues until the cost threshold is reached. A variant on that approach is to change the cost structure as the number of administered items increases.

Wald's (1947) sequential probability ratio test (SPRT, pronounced spurt) is clearly the most well-known sequential decision rule. SPRT for K multiple categories can be summarized as

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

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

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

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)

Methodology

This research addresses the following questions:

1. Does measurement decision theory result in accurately classified examinees?
2. Are the different sequential testing procedures using decision theory as effective as maximum information item selection using item response theory?
3. How many items need to be administered to make accurate classifications? and
4. How many examinees are needed to satisfactorily calibrate measurement decision theory item parameters?

These questions are addressed using two sets of simulated data. In each case, predicted mastery states are compared against known, simulated true mastery states of examinees.

Examinees (simulees) 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. When the probability was greater than the random draw, the simulee was coded as responding correctly to the item. When the probability was less, the examinee was coded as responding incorrectly. Thus, as with real testing, individual simulees sometimes responding incorrectly to items they should have been able to answer correctly.

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).

For each test, a calibration sample of 1000 examines and separate trial data sets 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.

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 ..9 7 1.65
Mastery states	2	4

The simulated state-NAEP draws from a large number of items and a very reliable test. The cut scores correspond 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 40^{th} percentile.

The accuracy of classifications using measurement decision theory relative to classifications using item response theory and the accuracy of sequential testing models relative to IRT computer adaptive testing were examined using these datasets. 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 IRT baseline, on the other hand, was 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

Classification Accuracy

A key question is whether use of the model will result in accurate classification decisions. Accuracy was evaluated under varying test lengths, datasets, and underlying distributions. Test

lengths were varied from 3 items to the size of the item pool. For each test length, 100 different tests were generated by randomly selecting items from the CSAP and NAEP datasets. For each test, 1,000 examinees and their item responses were simulated.

The results for select test sizes with the CSAP are shown in Table 4 and all CSAP values are plotted in Figure 1. There is virtually no difference between the accuracies of decision theory scoring and IRT scoring with either the uniform or normal underlying ability distributions. With the NAEP items, four classification categories, and normal examinee distributions, decision theory was consistently more accurate than IRT scoring (see Figure 2). With uniform distributions, IRT has a slight advantage until 25 items when the curves converge.

Table 4: Classification accuracy of simulated examinations using MAP decision theory and IRT scoring by item bank, test size and underlying ability distribution.

	uniform		normal	
size	map	irt	map	irt
CSAP items, 2 categories				
5	.850	.842	.762	.752
10	.900	.892	.810	.804
15	.924	.914	.839	.834
20	.936	.926	.857	.853
25	.945	.936	.869	.865
30	.951	.942	.879	.877
State-NAEP items, 4 categories				
5	0.513	0.623	0.61	0.539
10	0.638	0.694	0.68	0.635
15	0.705	0.742	0.72	0.682
20	0.745	0.766	0.755	0.724
25	0.773	0.787	0.774	0.75
30	0.8	0.802	0.791	0.772
35	0.823	0.818	0.805	0.79
40	0.838	0.827	0.813	0.799

Figure 1: Accuracy of decision theory (MAP), and IRT scoring as a function of test length and ability distribution for simulated tests based on CSAP.

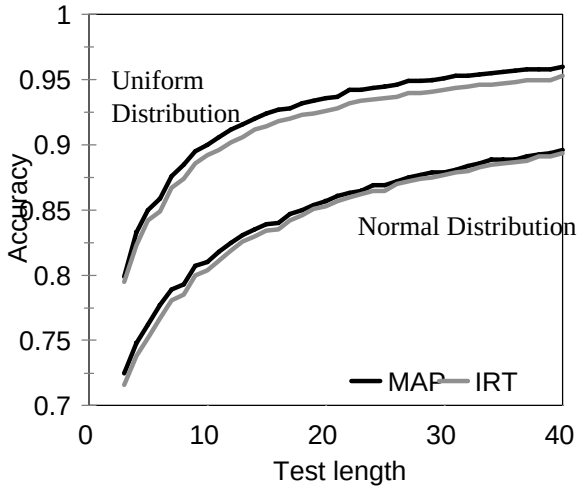
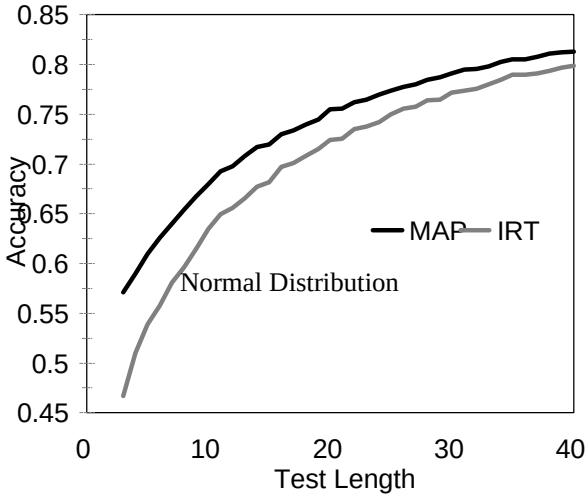


Figure 2: Accuracy of decision theory (MAP), and IRT scoring as a function of test length for simulated tests based on state-NAEP.



Sequential Testing Procedures

For this analysis, data sets of 10,000 normally distributed N(0,1) examinees and their responses to the CSAP and state-NAEP items were generated. 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.

Under the IRT approach, the items with the maximum information at the examinee's true score were selected without replacement. Thus, the procedure was optimized for IRT.

As shown in Table 5, 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 5: 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

Sequential decisions

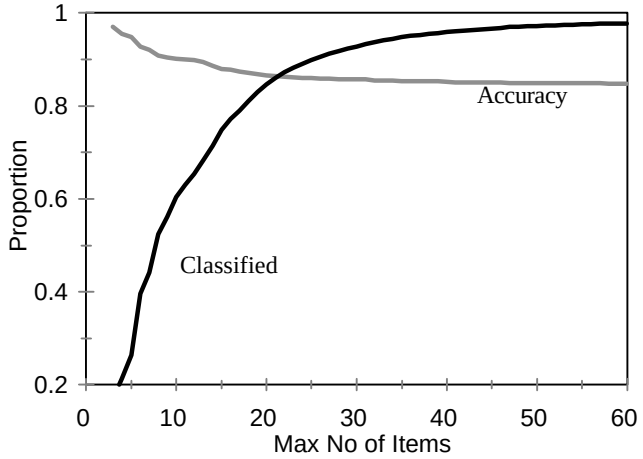
After each item was administered above, Wald's SPRT was applied to determine whether there was enough information to make a decision and terminate testing. Power and error rate where set to á=â= .05. Table 6 shows the proportion of examinees for which a classification decision could be made, the percent 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 those examinees were classified correctly into one of the 4 state-NAEP categories 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 6: 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 3. 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 3: 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).



Calibration

Another key question for any measurement model is the sample size needed to obtain satisfactory priors. With item response theory, the minimum acceptable calibration size is some 1000 examinees, which severely limits applications of the model.

The priors for the measurement decision theory model are the proportions of examinees in the population in each mastery state, $P(m_k)$, and the probabilities of responding correctly (and consequently the probabilities of responding incorrectly) given each mastery state, $P(z_i=1|m_k)$. These priors will usually be determined by piloting items with a calibration sample.

To determine the necessary number of calibration examinees, examinee classification accuracy as a function of calibration sample size and test size was assessed. Samples sizes of [20,30,40,50,60,70,80,90,100,200,300,400,500,600,700,800,900,1000] and test sizes of [5,10,15,20,25,30,35,40,45] were examined using state-NAEP and CSAP items. Under each condition, 100 tests were created by randomly selecting the appropriate number of items from the selected item pool. These tests were then each administered to 1000 simulees and the accuracy of the classification decision using MAP was determined.

Classification accuracy is usually best for tests calibrated on larger samples. In order to place the observed accuracies on a common scale, the accuracy of each sample size condition was divided by the accuracy of the corresponding 1000 calibration examinee condition to form a relative accuracy scale.

Accuracy of the priors is also limited by the size of the smallest cell. For the CSAP, this was always the non-masters (approximately 40% of the calibration sample). For NAEP, this was the Advanced category (approximately 17%). Variations due to cell size were controlled by dividing

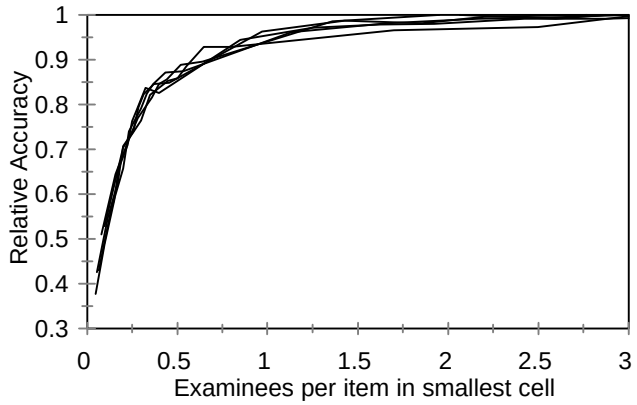
the number of calibration examinees in the smallest cell by the number of items on the simulated test. Thus, relative accuracy as a function of the number of calibration examinees per item in the smallest cell was used to help evaluate the needed calibration sample size.

Table 7 shows the results for 100 random 30 item tests using state-NAEP items. The data under the different test size conditions using state-NAEP items are quite similar and plotted in Figure 4. In Figure 4, the x-axis is truncated at 3 subjects per item. Beyond that value, the curves are flat. The results using CSAP were virtually identical. One can see from Figure 4 that relative accuracy levels off as the number of calibration examinees in the smallest cell approximates a little more than the test size. Thus, a random sample of only 25 to 40 examinees per cell would be needed to calibrate a 25 item test.

Table 7: Accuracy by number of examinees per item in the smallest cell using 100 random 30 items tests formed from state-NAEP items

Sample size (a)	smallest cell size (b)	accuracy ©	relative accuracy (d)	Examinees per item in (b) (e)
20	2	0.38	0.43	0.07
30	4	0.50	0.56	0.13
40	6	0.59	0.66	0.20
50	7	0.66	0.74	0.23
60	9	0.71	0.79	0.30
70	10	0.74	0.82	0.33
80	11	0.75	0.85	0.37
90	13	0.78	0.87	0.43
100	16	0.78	0.87	0.53
200	34	0.86	0.96	1.13
500	77	0.89	0.99	2.57
1000	163	0.89	1.00	5.43

Figure 4: Accuracy of tests formed from state-NAEP items relative to tests calibrated with 1000 examinees as a function of the number of calibration examinees per item in the smallest cell.



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.

The model has a very simple framework - one starts with the conditional probabilities of examinees in each mastery state responding correctly to each item. One can obtain these probabilities from a very small pilot sample. This research demonstrated that a minimum cell size of one examinee per item is a reasonable calibration sample size. The accuracies of tests calibrated

with such a small sample size are extremely close to the accuracies of tests calibrated with hundreds of examinees per cell.

An individual's response patterns is evaluated against these conditional probabilities. One computes the probabilities of the response vector given each mastery level. Using Bayes' theorem, the conditional probabilities can be converted to an *a posteriori* probabilities representing the likelihood of each mastery state. Alternative decision rules were presented. Using the *maximum a posteriori, MAP,* decision rule, this research found that the model was as good as or better than three parameter item response theory in accurately classifying examinees. Accuracy was also identical when making binary decisions using items from the Colorado State Assessment Program. The model was noticeably more accurate than IRT when making classifying examinees into one of four categories using items from state-NAEP. The measurement decision theory model is especially attractive when the IRT assumptions are violated or IRT cannot be applied.

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. Two new approaches where introduced. Information gain, which is based on entropy and comes from information theory, was almost identical to minimum cost. A second, 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.

The research also showed that very few pilot test examinees are needed to calibrate the system. One or two examinees per cell per item result in a test that is as accurate as one calibrated with

hundreds of pilot test examinees per cell. The results were consistent across item pools and test lengths. The essential data from the pilot is the proportions of examinees within each mastery state that respond correctly. One does not truly need *a priori* probabilities of a randomly chosen examinee being in each mastery state. Uniform priors can be expected to increase the number of needed items and not seriously affect accuracy given properly chosen stopping rules.

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.

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.

Note

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.

References

Allen, Nancy L., James E. Carlson, and Christine 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

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, Chuan-Ju; Spray, Judith (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. (1977). The use of probabilitistic models in the assessment of mastery. *Journal of Educational Statistics.* 2(2), 99-120.

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

Spray, Judith A. and Reckase, Mark 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, Judith A. and Reckase, Mark 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.

van der Linden, W. J. and Vos, H. J. (1966) A compensatory approach to optimal selection with mastery scores. *Psychometrika,* 61(1), 155-72.

Vos, Hans 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.

Wood, R. (1976). Adaptive testing: A Bayesian procedure for the efficient measurement of ability. *Programmed Learning and Educational Technology,* 13, 2, 36-48.