Practical Assessment, Research, and Evaluation

Volume 14 *Volume 14, 2009* Article 8

2009

# Scoring and classifying examinees using measurement decision theory

Lawrence M. Rudner



Recommended Citation
Rudner, Lawrence M. (2009) "Scoring and classifying examinees using measurement decision theory," *Practical Assessment, Research, and Evaluation*: Vol. 14 , Article 8.
DOI: https://doi.org/10.7275/vksg-rh07
Available at: https://scholarworks.umass.edu/pare/vol14/iss1/8

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 14, Number 8, April 2009 ISSN 1531-7714

# Scoring and Classifying Examinees Using Measurement Decision Theory

Lawrence M. Rudner, *Graduate Management Admission Council*

This paper describes and evaluates the use of measurement decision theory (MDT) 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. The presented evaluation investigates: (1) the classification accuracy of tests scored using decision theory; (2) the effectiveness of different sequential testing procedures; and (3) the number of items needed to make a classification. A large percentage of examinees can be classified accurately with very few items using decision theory. A Java Applet for self instruction and software for generating, calibrating and scoring MDT data are provided.

In the introduction to their classic textbook, Cronbach and Gleser (1957) argue that the ultimate purpose for testing is to arrive at classification decisions. Many of today's decisions are indeed binary, e.g., whether to hire someone, whether a person has mastered a particular set of skills, whether to certify an individual. Categorical, as opposed to continuous, outcomes are also common, e.g., the percent of students that perform at the basic, proficient, or advanced level in state assessments.

IRT models have been applied to help make classification decisions by laboriously placing individuals on ability scales and then using cut-points to make classifications. IRT models, however, are not always applicable in practical situations. IRT is fairly complex, relies on several fairly restrictive assumptions, requires large calibration samples, and may not make efficient use of questions when the goal is simple classification. Classification 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.

Often credited to Wald (1939, 1947, 1950), perhaps 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. Applied to 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 as concerned with the fit of the data to a theoretical model as is the case with item response theory (IRT) or in most latent class models. Very few pilot test examinees are needed and, with very few items, classification accuracy can exceed that of item response theory. Further, as this article hopes to show, this simpler model can be relatively easy to explain.

Given these features, it is surprising that decision theory has not attracted wider attention within the measurement community. Indeed, much of the computerized classification testing (CCT) literature reviewed by Thompson (2007) and by Parshall, Spray, Kalohn and Davey (2006) relies on IRT. The decision theory model can work well for small sample license and certification examinations, as the routing mechanism for intelligent tutoring systems, for end-of-unit examinations, and for adaptive testing.

Key articles in the mastery testing literature of the 1970s employed decision theory (Hambleton and Novick, 1973; Huynh, 1976; van der Linden and Mellenbergh, 1978) and should be re-examined in light of today's measurement problems. Lewis and Sheehan (1990) and others used decision theory to adaptively select testlets and 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.

This paper presents an overview and the key concepts of the measurement decision theory model and illustrates them using a binary classification (pass/fail) case and a sample three item test. The quality of the model is demonstrated by examining 1) the classification accuracy of tests scored using decision theory, 2) the effectiveness of different sequential testing procedures by comparing classification accuracies against those of different IRT scenarios, and 3) the number of items needed to make a classification.

## BACKGROUND

The objective is to form a best estimate as to the mastery state (classification or latent state) of an individual examinee based on the examinee's item responses, *a priori* item information, and *a priori* population classification proportions. 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.

The first component is the set of 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. The second component is a set of N pre-calibrated items for which the probability of each possible observation, usually right or wrong, given each mastery state is known *a priori*. Individual responses to the set of 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 $z_1, z_2, \ldots z_N$.

The last component is the decision space. One can form any number of D decisions based on the data. Typically, one wants to determine 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 $d_k$.

Calibration starts with the proportion of examinees in the population that are in each of the K categories and the proportion of examinees within 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 judgment. In the absence of information, equal priors can be assumed. The proportions that respond correctly to each item can be derived from a pilot test involving examinees who have already been classified or transformations of existing data. Once these sets of priors are available, the items are administered to new examinees, responses ($z_1, z_2, \ldots z_N$) are 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 prior 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

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

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

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

The posterior probability $P(m_k|\mathbf{z})$ that the examinee is of mastery state $m_k$ given his response vector, **z,** 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,

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

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 local independent 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 equally applicable to polytomous scoring.

Three key concepts from decision theory applied in this paper are briefly 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 individual's response pattern, and
3. sequential decisions - alternative procedures for determining whether to continue testing.

Melsa and Cohn (1978) present an excellent overview of decision theory. That manuscript was the inspiration for this research and is well worth reading.

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 prior test data, the 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

Upon administering a set of pre-calibrated items, one can compute $P(\mathbf{z}|m_k)$, the probability of the response vector given each of K possible classifications, and $P(m_k|\mathbf{z})$, the posterior classification probabilities that consider the prior classification probabilities. The task then is to classify the examinee in one of the K mastery states.

From (2), the probabilities of the vector $\mathbf{z}$= [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 $P(\mathbf{z}|m_2)=.09$, or normalized $P(\mathbf{z}|m_1)= .68$ and $P(\mathbf{z}|m_2)=.32$.

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 $\lambda$.

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

The value of $\lambda$ reflects the selected approaches and judgments 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 $\lambda = 1.0$. 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 relatively close (.19 and .09), 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, $\alpha$ and $P(d_2|m_2)$ is the power of the test, $\beta$. 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 formula (2) is employed with

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

With the example, $\lambda$=.25 and the decision is $d_2$ - non-master.

### *Maximum a posteriori (MAP) decision criterion*

The maximum likelihood decision criterion made use of only the probabilities of the response vector. The minimum probability of error criterion added in the use of the prior classification probabilities $P(m_1)$ and $P(m_2)$. The maximum likelihood *a posteriori* decision criterion also uses both probabilities of the response vector, $P(\mathbf{z}|m_k)$ and the prior classification probabilities $P(m_k)$.

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

By this criterion, one selects the category with the largest value from equation (3). 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 (1), $P(m_k|\mathbf{z})=c\,P(\mathbf{z}|m_k)\,P(m_k)$, MAP is equivalent to the minimum probability of error decision criterion. MAP is also equivalent to the maximum-likelihood decision criterion when the prior probabilities are equal.

### *Bayes risk criterion*

A significant advantage of the decision theory framework is that one can incorporate decision costs into the analysis. By this criterion, costs are assigned to each correct and incorrect decision so the total average costs can be minimized. 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

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

and the criterion can be stated as

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

By this criterion, one selects the category with the smallest value from equation (3). 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 this approach is identical to *MAP*.

## 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 the measurement literature, 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

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

is the sum of the probability of responding correctly if the examinee is a master plus the probability if a non-master. Applying (5), the current probability of correctly responding $P(z_2=1)=.53$ 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 to sequential testing defines the optimal item to be administered next as the one with the lowest expected cost. 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. Equation (4) provided the decision cost as a function of the classification probabilities. If $c_{11}=c_{22}=0$ then

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

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 are, by definition, the posterior probabilities given in (1). Thus, with $c_{12}=c_{21}=1$, the Bayes cost after administering the first question is B=1*(1-.33)*.33 + 1*(1-.66)*.66 = .44.[1]

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


DOI: https://doi.org/10.7275/vksg-rh07

vectors. The commonly used measure of information from information theory, Shannon (1948) entropy, is applicable here (see Cover and Thomas, 1991):

$$H(S) = \sum_{k=1}^{K} - p_k \log_2 p_k \qquad (7)$$

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. Since the goal is to have a peaked distribution of $P(m_k)$, one wants the lowest possible value of *H(S)*. 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 to an incorrect response:

$$H(S_i) = p(z_i{=}1)\ H(S_i|z_i{=}1) + p(z_i = 0)\ H(S_i|z_i = 0) \qquad (8)$$

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 (7).
3. Weight the conditional entropies by their probabilities using (8).

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 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, 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.

## 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 decision. 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 $\alpha$. Suppose $\alpha = .05$ 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 $\alpha$, the decision is to continue testing.

A variant of Neyman-Pearson is the *fixed error rate criterion* - establish two thresholds, $\alpha_1$ and $\alpha_2$, and continue testing until $P(d_2|m_1) < \alpha_1$ and $P(d_1|m_2) < \alpha_2$. 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

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

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

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

where the $P(m_j)$'s are the normalized posterior probabilities, $\alpha$ is the acceptable error rate, and $1-\beta$ is the desired power. If the condition is not meet for any category k, then testing continues. 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

The model is evaluated by addressing the following research questions:

1. Does decision theory result in accurately classified examinees?
2. Are the different sequential testing procedures using decision theory as effective as item selection based on maximum information using item response theory?
3. How many items need to be administered to make accurate classifications?

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

### *Data Generation*

These questions are addressed using simulated responses based on IRT parameters for items from the 1999 Colorado State Assessment Program (CSAP) fifth-grade mathematics test (Colorado State Department of Education, 2000) and the 1996 National Assessment of Educational Progress (NAEP) State Eighth Grade Mathematics Assessment (Allen, Carlson, and Zelenak, 2000). Birnbaum's (1968) three parameter model was used. Key statistics for these tests are given in Table 3.

**Table 3: Descriptive statistics for simulated tests.**

| | Simulated test | |
|---|---|---|
| | **CSAP** | **State NAEP** |
| **No of items in pool** | 54 | 139 |
| **Mean a** | .78 | .94 |
| **Mean b** | -1.25 | .04 |
| **Mean c** | .18 | .12 |
| **Mastery states** | 2 | 4 |
| **Cut score(s)** | -.23 | -.23, .97, 1.65 |
| **For a N(0,1) sample** | | |
| **Proportions in each mastery state** | .41, .59 | .41, .42, .12, .05 |
| **Reliability** | .83 | .95 |
| **Chance level** | .52 | .36 |

Reliability here was computed as the square root of 1 minus the squared standard error where the standard error was weighted by the distribution of a N(0,1) sample. The chance level is $\sum P(m_k)^2$, the probability of a correct classification given the cut scores for an examinee randomly selected from a normal distribution.

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

https://scholarworks.umass.edu/pare/vol14/iss1/8
DOI: https://doi.org/10.7275/vksg-rh07

small proportion of examinees for the Advanced level and the use of four mastery state classifications provide a good test for 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. Because classification categories are not reported for CSAP, the mastery/non-mastery cut score used in the study was arbitrarily selected to correspond to the 41th percentile.

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 according to the corresponding cut score interval. Probabilities of a correct response were computed using Birnbaum's (1968) three-parameter IRT model and then probabilistically converted to observable dichotomous scores.

Thus, for each simulated examinee, there is a corresponding true score ($\theta$), corresponding latent state ($m_k$), and a response vector ($\mathbf{z}$). The proportions of examinees in each latent state are, by definition, the prior classification probabilities, $P(m_k)$. The latent states and the response vectors were used to compute the conditional prior probabilities of each response $z_i$ given each mastery state $m_k$, $P(z_i|m_k)$. The specific design of each simulation is discussed along with the results in the next section.

### *Data Recovery*

For decision theory approaches, maximum a posterior (MAP) probabilities were used to determine the observed examinee classifications. For the IRT approaches, theta-hats were estimated using the Newton-Raphson iteration procedure outlined in Lord (1980). Examinees were then classified into the category corresponding to the theta interval containing the estimated theta.

The reader should note that 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 favorable scenarios for both decision theory and IRT. The examinees in the calibration sample are classified without error, thus providing accurate priors for applying decision theory. The data also fit the IRT model perfectly.

Because the data are generated using an IRT model with a continuous theta scale, decision theory with a finite number of discrete categories presents a mis-specified model for recovering the data. From an IRT perspective, the probability of a correct response increases within each slice of the theta scale and theta increases within each slice as well. As a result, the response patterns are more alike within each slice and local independence is clearly violated. This might present a problem if one were to use IRT to directly recover the latent classes.

While the data were generated using a continuous theta scale, this analysis takes a decision theory perspective. The underlying distributions within each category are not of interest. Examines within the same latent class are treated as if they have the same ability. The probabilities of a correct response are considered to be the same for all members of the same class. Thus, while this analysis invokes the "naive Bayes" local independence assumption, within-class local independence is not an issue.

### *Analysis*

Classification accuracy using a simple decision theory model is compared to accuracy using a more complicated item response theory model. Accuracy was defined here as the proportion of correct state classifications. In order to compare results with different numbers of categories, in this case 2 for CSAP and 4 for NAEP, accuracies were converted to Proportion Reduction in Error (PRE):

$$PRE = \frac{(\%\text{ accurate classification} - \%\text{ accurate by chance})}{(100\% - \%\text{ accurate by chance})}$$

PRE is 0.0 when the rule in question is useless and 1.0 when the rule is perfect.

## SIMULATIONS AND 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 by randomly selecting items from the CSAP and NAEP datasets. For each test length, 1,000 examinees from a normal N(0,1) distribution and 1,000 examinees from a uniform U(-2.5,2.5) distribution along with their item responses were simulated. Each condition was then replicated 100 times.

The results for select test sizes with the CSAP and NAEP are shown in Table 4. For CSAP, 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. With uniform distributions, IRT has a slight advantage until the test length reaches 30 items.

### *Sequential Testing Procedures*

For this analysis, two 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 fixed common datasets, items were selected and mastery states were predicted using three sequential testing approaches (minimum cost, information gain, and maximum discrimination) and three IRT approaches.

**Table 4: Proportion Reduction in Error 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 | 0.697 | 0.681 | 0.508 | 0.487 |
| 10 | 0.798 | 0.782 | 0.607 | 0.595 |
| 15 | 0.847 | 0.827 | 0.667 | 0.657 |
| 20 | 0.871 | 0.851 | 0.704 | 0.696 |
| 25 | 0.889 | 0.871 | 0.729 | 0.721 |
| 30 | 0.901 | 0.883 | 0.750 | 0.746 |
| **State-NAEP items, 4 categories** | | | | |
| 5 | 0.293 | 0.453 | 0.387 | 0.275 |
| 10 | 0.475 | 0.556 | 0.497 | 0.426 |
| 15 | 0.572 | 0.625 | 0.560 | 0.500 |
| 20 | 0.630 | 0.660 | 0.615 | 0.566 |
| 25 | 0.670 | 0.691 | 0.645 | 0.607 |
| 30 | 0.710 | 0.713 | 0.671 | 0.642 |
| 35 | 0.743 | 0.736 | 0.693 | 0.670 |
| 40 | 0.765 | 0.749 | 0.706 | 0.684 |

The decision theory approaches are applied as described earlier. For the minimum-cost decision theory approach, the costs of deciding $d_i$ when $m_j$ is true were set symmetrically at $|i-j|$ for all i, j. After the desired number of items were administered, all examines were classified using MAP.

Under the first of the three IRT approaches, the items with the maximum information at the examinee's true score were sequentially selected without replacement. While this is not feasible in real life, it presents a best case scenario when the goal is to obtain accurate estimates along the entire theta scale. Under the second IRT approach, the items with the maximum information at the examinee's currently estimated ability level were sequentially selected without replacement. This is a realistic and practical approach when the goal is to obtain accurate estimates along the entire theta scale. Following the suggestion of Spray and Reckase (1994), the third approach sequentially presented the items with the maximum information at the cut score closest to the examinee's currently estimated ability level. This approach is optimal when the goal is to classify examinees into one of a discrete number of score groups. After the desired number of items were administered, all examines were classified into the score group corresponding to the terminal estimate of theta.

As shown in Table 5, there is not a great deal of variance across the different approaches. The minimum cost and information gain decision theory approaches consistently out-performed the first two IRT approaches, and out-performed the IRT cut score approach when 20 or fewer items were administered. 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 first two IRT approaches are 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.

### *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. Error rates where set to $\alpha=\beta= .05$. Table 6 shows the proportion of examinees for which a classification decision could be made, the percent of those examinees that were correctly classified, PRE, 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.

The proportions classified and the corresponding accuracy as a function of the maximum number of items administered from Table 6 are shown in Figure 1. 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.

## DISCUSSION

The simple measurement model presented in this paper is applicable to 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.

An individual's response pattern is evaluated against the conditional probabilities. One computes the probabilities of the response vector given each mastery level. Using Bayes' theorem, the conditional probabilities can be converted to *a posteriori* probabilities representing the likelihood of each

**Table 5: Proportion Reduction in Error for sequential testing methods as a function of maximum test length.**

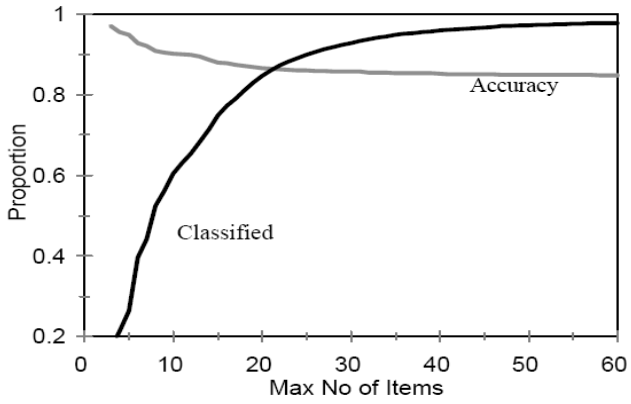
| Max No of items | IRT Approaches | | | Decision Theory Approaches | | |
|---|---|---|---|---|---|---|
| | Max I(θ) | Max I(θ') | Max I(cut) | Max Disc | Min Cost | Info Gain |
| **CSAP items, 2 categories** | | | | | | |
| 5 | 0.607 | 0.564 | 0.661 | 0.564 | 0.661 | 0.661 |
| 10 | 0.702 | 0.679 | 0.706 | 0.690 | 0.715 | 0.717 |
| 15 | 0.729 | 0.733 | 0.748 | 0.727 | 0.752 | 0.750 |
| 20 | 0.756 | 0.760 | 0.775 | 0.779 | 0.770 | 0.764 |
| 25 | 0.772 | 0.783 | 0.787 | 0.779 | 0.787 | 0.789 |
| **State NAEP items, 4 categories** | | | | | | |
| 5 | 0.576 | 0.447 | 0.530 | 0.418 | 0.596 | 0.594 |
| 10 | 0.645 | 0.640 | 0.659 | 0.546 | 0.681 | 0.675 |
| 15 | 0.704 | 0.682 | 0.704 | 0.646 | 0.720 | 0.714 |
| 20 | 0.723 | 0.722 | 0.737 | 0.709 | 0.737 | 0.736 |
| 25 | 0.748 | 0.750 | 0.761 | 0.741 | 0.755 | 0.755 |
| 30 | 0.756 | 0.770 | 0.772 | 0.756 | 0.767 | 0.767 |

**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 | Prop Reduct Error | Mean No of items |
|---|---|---|---|---|
| 5 | 0.260 | 0.948 | 0.892 | 4.6 |
| 10 | 0.604 | 0.902 | 0.797 | 7.4 |
| 15 | 0.749 | 0.880 | 0.752 | 9.1 |
| 20 | 0.847 | 0.865 | 0.721 | 10.2 |
| 25 | 0.899 | 0.860 | 0.710 | 10.8 |
| 30 | 0.928 | 0.857 | 0.704 | 11.3 |
| 40 | 0.960 | 0.852 | 0.694 | 11.8 |
| 50 | 0.972 | 0.849 | 0.688 | 12.2 |
| 100 | 0.988 | 0.847 | 0.684 | 13.0 |

.

**Figure 1**: 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).




mastery state. 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. The model was noticeably more accurate than IRT when classifying examinees into one of four categories. Conceivably, the decision theory model will be 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 were 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.

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 local independence is often ignored in measurement and one might expect only minor violations, its role in 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 classification. 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. This does not appear to be a problem for this measurement application of decision theory.

Measurement Decision Theory is clearly a simple yet powerful and widely applicable model. The advantages of this model are many -- it 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 model is already the basis for a highly visible commercial tool to help test-takers prepare to for the GMAT®. The author can envision a much wider use of the model. It is a natural routing mechanism for intelligent tutoring systems. Under this model, 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? What effect does the location of cut scores have on the ability of decision theory to classify examinees? Can the model be effectively extended to 30 or more categories to provide a rank ordering of examinees? How can one make good use of the fact that the data are ordinal? How can the concept of entropy be employed in the examination of tests? Are there new item analysis procedures that can improve 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?

## References

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

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. and Spray, J. (2000). Effects of item-selection criteria on classification testing with the sequential probability ratio test. ACT Research Report Series.

Lord, Frederick M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

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.

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

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.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2006). *Practical considerations in computer-based testing*. New York: Springer.

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*, 16(1), 65-76.

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).

Thompson, Nathan A. (2007). A Practitioner's Guide for Variable-length Computerized Classification Testing. *Practical Assessment Research & Evaluation,* 12(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1

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

Wald, A. (1939). A New Formula for the Index of Cost of Living,. *Econometrica* 7(4), 319-331.

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

Wald, A. (1950). *Statistical Decision Functions*. 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.

## Footnote

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

## Notes

1. An interactive tutorial is available on-line 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.
2. 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. If you are interested in the source code please contact the author.

## Acknowledgements

The author is grateful for the extremely helpful comments made on an earlier draft by Chan Dayton, George Macready and two anonymous reviewers.

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.

## Citation

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

## Corresponding Author

Lawrence M. Rudner
Graduate Management Admission Council
1600 Tysons Blvd, #1400
McLean, VA 22102 USA

Email: LRudner [at] gmac.com or LMRudner [at] gmail.com