Decision Models for Use with Criterion-Referenced Tests

Wim J. van der Linden
Twente University of Technology

The problem of mastery decisions and optimizing cutoff scores on criterion-referenced tests is considered. This problem can be formalized as an (empirical) Bayes problem with decisions rules of a monotone shape. Next, the derivation of optimal cutoff scores for threshold, linear, and normal ogive loss functions is addressed, alternately using such psychometric models as the classical model, the beta-binomial, and the bivariate normal model. One important distinction made is between decisions with an internal and an external criterion. A natural solution to the problem of reliability and validity analysis of mastery decisions is to analyze with a standardization of the Bayes risk (coefficient delta). It is indicated how this analysis proceeds and how, in a number of cases, it leads to coefficients already known from classical test theory. Finally, some new lines of research are suggested along with other aspects of criterion-referenced testing that can be approached from a decision-theoretic point of view.

With criterion-referenced tests, the test items can be conceived as a sample from a domain of tasks covering a well-defined objective or competency, and the concern is ordinarily with the examinee's domain score. A domain score is the proportion of successes to be expected when an examinee is administered the entire domain; formally, it is known as the relative generic true score (Lord & Novick, 1968, sect. 11.2).

In the above conceptualization, test and criterion consist of the same type of tasks, and it can therefore be stated that the test is referenced to an *internal* criterion. Other conceptualizations using an internal criterion can be found in Cox and Graham (1966) and Wright and Stone (1979, chap. 5). The present paper will also consider the use of decision models in applications in which an independently measured or *external* criterion is employed for referencing a test. This implies that test and criterion behavior are of a different type but that an empirical relation has been established that is strong enough to support an interpretation of the former in terms of the latter.

It is important to observe that in criterion-referenced measurement, some notion of true or latent score is always involved, whether an internal or external criterion is used. It is the continuum underlying the test that is intended to be criterion referenced; the observed score is only used for assessing the examinee's position on this continuum and, thereby, his/her criterion behavior. In this paper this true

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 469–492

score variable will be denoted by T and the observed score by X; the formal meaning of T will be determined by the psychometric models that will be adopted to explain the observed test scores, X.

One of the principal uses of criterion-referenced measurement is in the assignment of students to mastery states. Typically, this involves the selection of a cutoff score on the criterion-referenced scale T. Students with true scores exceeding this cutoff score are considered masters; they are deemed to have reached the learning objectives and may proceed with the next unit or task. Students below this cutoff score are the nonmasters; usually, they are provided with extra learning time or remedial teaching.

In order to be able to classify students as masters and nonmasters, the presence of a carefully chosen cutoff score is not sufficient. All that is known about students is their observed test score, X, while the cutoff score is set on the true score scale, T. In many practical situations, it has been common to ignore this difference and to compare observed scores directly with the true cutoff score. This amounts, however, to assuming that the test is free of measurement error, a situation that will hardly be met in educational measurement. When domain sampling is used, this practice is still less realistic. Then, not only measurement error but also sampling error is involved.

A better solution, therefore, is to introduce a separate cutoff score on the test and to use this for assigning examinees to mastery states. Let c denote the cutoff score on the test score scale and d the cutoff score on the true score scale, and suppose that a psychometric model is available relating X to T (test score to true score). A student is a true master if his/her T score exceeds $T = d$ and is a nonmaster otherwise; but mastery is declared if $X \geqslant c$ and nonmastery if $X < c$. The problem is to choose a value of c that is optimal in some sense for a given value of d.

The purpose of this paper is to show that the above problem can be handled effectively within a decision-theoretic framework and to review applications of this framework. Moreover, it will be shown how decision-oriented concepts can be used to evaluate mastery decisions and to solve other criterion-referenced testing problems. Finally, attention is called to the fact that when applied in objectives-based programs, criterion-referenced tests are the endpoint of several alternative instructional routes (treatments) and constitute the "criteria" against which the treatment-assignment procedure is to be evaluated. Decision models can be used to optimize these procedures as well.

Before elaborating on these points, however, it is emphasized that in what follows two different cutoff scores are involved—the true and the observed cutoff scores. Decision theory can *not* be used to set the former; it can be used to set the latter after a solution to the former has been obtained. To those not accustomed to the concepts of measurement and sampling error, this distinction might seem a bit confusing; and decision-theoretic outcomes such as "if you have adopted a true cutoff score of 16 out of 20 items correct, then you must choose a cutoff score on your test equal to 19 items correct" might seem paradoxical. Nevertheless, the use of decision theory is a rational way of coping with unreliable measurements and can, as will be illustrated later, lead to an improvement in assigning examinees to mastery states. Though this has not always been seen (e.g., Glass, 1978), the decision-theoretic approach to criterion-referenced testing is thus no standard-setting technique but a technique to minimize the consequences of measurement and sampling error, which, preferably as a part of the normal routine, ought to follow each time a standard-setting technique is used.

Criterion-Referenced Testing as a Decision-Theoretic Problem

Decision theory combines information about true states and utilities of outcomes into optimal decisions. Since decision theory is especially concerned with information in the form of data with a random error component, it can also be stated that decision theory combines *probability* and *utility* into optimal decisions. In criterion-referenced testing, the former is provided by a psychometric model; for the latter a variety of techniques are available, all of which somehow scale the value of the decision outcomes to the decision-maker.

Some Decision-Theoretic Notions

There are several excellent introductions to decision theory. A game-theoretic and applied context is offered in Luce and Raiffa (1957) and Raiffa and Schlaifer (1961, part 1). Statistical treatments are given in, for example, Degroot (1970), Ferguson (1967), and Lindgren (1976, chap. 8). Yamane (1973, chaps. 10 and 17) has presented a very short and simple exposition. This paper introduces only a few concepts and principles needed for formulating criterion-referenced testing as a decision-theoretic problem, thereby sacrificing some mathematical precision in order to enhance understanding.

A central concept in decision theory is the *state space*. It is the set of all possible states of nature with respect to which actions are to be taken. It will be represented by Ω; and its individual states by the numerical parameter θ or, when Ω is discrete, by θ_i. The set of actions available to the decision-maker is called the *action space*. This will be denoted by A; and the individual actions, by a or a_j. Suppose that the decision-maker is able to evaluate on a numerical scale the consequences of taking action a while the true state of nature is θ. This numerical evaluation is what is technically known as loss. The function mapping points in $\Omega \times A$ to the loss scale L is the *loss function* $l(\theta, a)$. When A is discrete, a convenient notation is $l_j(\theta)$.[1]

If the state of nature were known, the most obvious thing would be to choose actions with minimal loss. In that case, decision theory would be trivial and would hardly contribute anything. The point is, however, that in most instances the true state of nature is unknown and there is only the disposal of fallible information or data. For the purpose of this paper, it will be assumed that information is available in the form of an observed value z of a random variable Z representing the outcome of some experiment or measurement. Z will be considered to be related to the state of nature by a probability model with parameter θ, $\pi(z;\theta)$. The stochastic character of Z causes the making of decision errors and, generally, the choosing of actions that yield outcomes with larger loss than when the true state of nature had been known. Decision theory is concerned with techniques for selecting decision rules that are, nevertheless, as optimal as possible and with the study of their properties.

A nonrandomized *decision rule* is a prescription specifying for each possible value z of Z what action has to be taken. In mathematical terms, it is a mapping from Z to the action space A: $A = \delta(Z)$. It should be noted that this definition of A as a function of the random variable Z implies not only that the actions are taken at random but also that loss is random: $l(\theta,\delta(Z))$. To solve the problem of selecting optimal decision rules out of the large (possibly even infinite) collection of mappings to be defined from Z to A, the risk function or expected loss is defined as

$$R(\theta,\delta) \equiv E[l(\theta,\delta(Z))], \qquad [1]$$

where the expectation is taken using $\pi(z;\theta)$, i.e., for a given value of θ across Z. The importance of Equation 1 lies in the fact that it shows the loss that can be expected when decision rule δ is used and nature is in state θ.

[1] In this paper only the loss terminology will be used, and henceforth terms such as utility, opportunity loss, and regret will be ignored. These terms are used sometimes with and sometimes without a fundamental difference in meaning. (For a coherent set of definitions, see Lindgren, 1976, chap. 8.)

There are two usual criteria for optimizing decision rules: the minimax criterion and the Bayes risk. In the absence of knowledge concerning which state of nature is true, the *minimax principle* assumes that it is best to prepare for the worst and to establish the maximum risk for each possible decision rule:

$$\max_{\theta} R(\theta,\delta). \qquad [2]$$

Once these are obtained, the best decision rule is the rule minimizing Equation 2. This is the rule δ' obeying

$$\max_{\theta} R(\theta,\delta') = \min_{\delta} \max_{\theta} R(\theta,\delta). \qquad [3]$$

Unlike minimax rules, the *Bayes principle* supposes that some prior knowledge about the state of nature is available and that a probability distribution (the "prior") can be chosen representing this knowledge. Now Θ is a (continuous) random variable, and its distribution will be denoted by the probability density function (*p.d.f.*) $\omega(\theta)$. The *Bayes risk* is defined as the expected value of $R(\Theta,\delta)$ using $\omega(\theta)$

$$B(\delta) \equiv \int R(\theta,\delta)\omega(\theta)d\theta. \qquad [4]$$

A Bayes rule with respect to $\omega(\theta)$ is a rule minimizing this Bayes risk:

$$B(\delta'') = \min_{\delta} B(\delta). \qquad [5]$$

A minimax rule can be conceived as a rule that minimizes Equation 4 as well, but under the restriction that $\omega(\theta)$ is a least favorable element of the class of priors (e.g., Ferguson, 1967, sect. 1.6). Bayes rules are, in general, less pessimistic; they can be based on any prior representing the available knowledge about nature. In this paper, only nonrandomized Bayes rules will be considered further. The possibility of randomized rules will be ignored, since these are expected to lead to acceptability problems when applied in educational settings, and it can be established that for each randomized rule there exists a nonrandomized Bayes rule that is at least as good (Ferguson, 1967, p. 43).

Monotone Rules

Defining a Bayes rule does not imply that its form is known and the actual minimization involved in Equation 5 may be laborious, especially when the class of all possible rules is large. It would therefore be helpful if this class of rules can be reduced beforehand to some, hopefully small, subclass of rules among which the rule being sought is to be found. It is here that the notion of an essentially complete class of decision rules comes in handy. An *essentially complete class,* for a given decision problem, is defined as a class containing rules that are as good as (and for some states of nature possibly even better than) the rules outside this class (see Ferguson, 1967, p. 55).

Even if it were allowed that attention be restricted to some essentially complete subclass of rules, it would also be helpful if all rules in this subclass had a known form so that analytic means could be used to select an optimal rule. This may save the work of first computing Equation 4 for all rules and then choosing a rule with the minimum value. An important class of rules suited for analytical manipulation is that of monotone rules. For a two-action problem, and in view of the application to criterion-referenced testing, attention will be restricted to this type of problem: A (nonrandomized) decision rule has a *monotone* form if there is a value z^* of Z so that action a_0 is taken ($\delta(z) = a_0$) whenever $Z < z^*$, and a_1 otherwise ($\delta(z) = a_1$). (A discussion of monotone multiple-decision problems is to be found in Ferguson, 1967, sect. 6.1, and Lindgren, 1976, sect. 8.3.5.).

There is a theorem in decision theory stating that the class of monotone decision rules is essentially complete if two conditions are met. The first is that the distribution of Z, given $\Theta = \theta$, $\pi(z|\theta)$, has a monotone likelihood ratio; the second, that the loss function is monotone (Ferguson, 1967, sect. 6.1; Karlin & Rubin, 1956). Fortunately, there is no need to bother about the condition of monotone likelihood ratio: This condition is fulfilled for the wide class of distributions known as the exponential family, and the distributions generally used in criterion-referenced testing belong to this family. In a two-action problem, the loss function is monotone if there exists one point $\Theta = \theta_0$ for which $l_0(\theta)$ and $l_1(\theta)$ possess an intersection.

Monotone decision problems entail a special form of the Bayes risk. From Equations 1 and 4, recalling that $\delta(z) = a_0$ for $Z < z^*$ and $\delta(z) = a_1$ for $Z > z^*$, and taking Z to be discrete, it follows that

$$B(\delta) = \int \sum^{z^*-1} l_0(\theta)\pi(z|\theta)\omega(\theta)d\theta + \int \sum_{z=z^*} l_1(\theta)\pi(z|\theta)\omega(\theta)d\theta. \qquad [6]$$

This form shows that the Bayes risk may be interpreted as the expected loss of the decision procedure with respect to the bivariate distribution of (Z,Θ). The Bayes risk can also be written as

$$B(\delta) = \sum^{z^*-1} [\int l_0(\theta)\pi(\theta|z)d\theta]\omega(z) + \sum_{z=z^*} [\int l_1(\theta)\pi(\theta|z)d\theta]\omega(z), \qquad [7]$$

where $\omega(z)$ and $\pi(\theta|z)$ are now the probability (density) function (*p.*(*d.*)*f.*) of Z and Θ given $Z = z$, respectively, and it is assumed that the interchange of integration and summation is allowed. The importance of Equation 7 lies in the fact that the bracketed factor in both terms is the conditional expected loss, given $Z = z$. It is also called the *posterior expected loss*, because it can be viewed as the expected loss once an observation $Z = z$ has been made.

Note how Equation 7 suggests a way of minimizing the Bayes risk: As $\omega(z)$ is a nonnegative constant for each value z of Z, the Bayes risk is minimal if for each $Z = z$, an action with smallest posterior expected loss is chosen. (The monotonicity assumed in Equation 7 implies that this is action a_0 up to some value z^*, and a_1 thereafter.) This minimization, using the posterior expected loss for each $Z = z$, is called the *extensive form* of analysis, whereas techniques directly dealing with the Bayes risk are known as the *normal form* of analysis.

Some authors (e.g., Davis, Hickman, & Novick, 1973) have claimed that the extensive form of analysis offers computational advantages and should be preferred over the normal form. Undoubtedly this is true when only one decision is to be made and the "data" are already available in the form of one observed value z of Z. It should be understood that this is the situation always referred to by authors with a subjectivistic interpretation of prior distributions; and the above authors are considered to adhere to this interpretation. When a series of decisions are to be made and several, if not all possible, values of Z are collected, the situation changes somewhat. Then it may be prudent to establish optimal decision *rules*, and, in doing so, not to compute posterior expected losses for each $Z = z$ but to use analytic means. The distinction between a single decision and a series of decisions is thus closely related to the interpretation of the prior distribution, an issue which will be further considered below.

Criterion-Referenced Testing Formalized

Criterion-referenced testing can now be formulated as a decision-theoretic problem. In criterion-referenced testing, the state space consists of two possible states, $\Omega = \{\overline{M}, M\}$, obtained by dichotomizing the criterion-referenced (true-score) scale underlying the test by a cutoff score d. Theoretically, it is attractive to define a true score scale, T, which is independent of test length and ranges from 0 to 1. In order to arrive at simple results, it will, however, be assumed that T runs from 0 to n (number of items in the test) and that the nonmastery and mastery states are defined as $\overline{M} = [0,d)$ and $M = [d,n]$. The action space also consists of two possible states, $A = \{a_0, a_1\}$, where a_0 is the action of granting status $\overline{M}$ to a student, and a_1 status M. For each student a test score is available. Although, in principle, a great variety of statistics defined on the vector of item responses can be chosen as a test score, only the number-correct score, X, will be considered in this paper. Thus, in criterion-referenced testing problems, the observed score X is an interpretation of Z from the previous sections; and the true score T, of Θ. It will further be assumed that a psychometric model provides the probability function relating the observed values of X to a given value $T = \tau$. A new notation, $f(x|\tau)$, will be used for this probability density to indicate that it is an interpretation of the probability model relating data to true state in the general decision model from the previous sections. The same will be done for the *p.(d.)f.*'s to be defined below. Finally, linking up with common practice in criterion-referenced testing, the decision rule is taken to be monotone,

$$\delta(X) = \begin{cases} a_0 & \text{if } X < c \\ a_1 & \text{if } X \geq c. \end{cases} \qquad [8]$$

The optimal decision rule δ^*—or, equivalently, the optimal cutoff score c^*—to be identified is the value of c minimizing the expected loss with respect to (X,T),

$$B(c) = \int_0^n \sum_{x=0}^{c-1} l_0(\tau)f(x|\tau)g(\tau)d\tau + \int_0^n \sum_{x=c}^{n} l_1(\tau)f(x|\tau)g(\tau)d\tau \qquad [9]$$

$$= \sum_{x=0}^{c-1} \int_0^n [l_0(\tau)p(\tau|x)d\tau]h(x) + \sum_{x=c}^{n} \int_0^n [l_1(\tau)p(\tau|x)d\tau]h(x),$$

where $g(\tau)$, $h(x)$, and $p(\tau|x)$ are the *p.(d.)f.* of T, X, and T given $X = x$, respectively, and $l_j(\tau)$, $j = 0, 1$, is the loss function (compare Equations 6 and 7).

The interpretation of the prior distribution $g(\tau)$ in this paper will not be the personal or subjective probability interpretation that $g(\tau)$ represents the decision-maker's belief in the true score value τ of one given person, as is the usual interpretation in Bayesian decision theory. The empirical Bayes approach introduced by Robbins (1956, 1964) will instead be adopted. In this approach it is assumed that the same decision occurs repeatedly without a change in $\pi(z|\theta)$ and $\omega(\theta)$. Each time a value of Z is observed, thus generating a sequence $(Z_1, Z_2, \ldots, Z_N)$ that can be used for estimating $\omega(\theta)$. For the criterion-referenced testing problem this means that $g(\tau)$ is interpreted as the true score distribution for some population of students. The sequence $(X_1, X_2, \ldots, X_N)$ is the vector of test scores of a sample of N students from this population observed prior to the moment that the optimal cutoff score on the test is established; it is used for estimating the parameters of the true score distribution, $g(\tau)$, which has a form specified by the psychometric model. Once the optimal cutoff score, c^*, has been estimated, it can be employed for making mastery decisions, not only for the first N students but also for every following student from this population. It seems natural to use subsequent observations of X for improving the estimates of $g(\tau)$ and c^* as well. (For fully Bayesian approaches relevant to criterion-referenced testing, refer to Hambleton, Hutten, & Swaminathan, 1976; Hambleton & Novick, 1973; Lewis, Wang, & Novick, 1975; Novick, Lewis, & Jackson, 1973; Swaminathan, Hambleton, & Algina, 1975.)

Probability Models Used in Criterion-Referenced Testing

In this section, some probability models used in criterion-referenced testing will be reviewed. These models are the classical test model, the beta-binomial model, and the bivariate normal model. For the first and the last model, the case of an internal and of a directly measured external criterion will be considered. The beta-binomial model seems most useful with an internal criterion, e.g., when sampling from an item domain may be assumed. For notational convenience, "T" will be used as a generic symbol for the true score underlying the test; the formal meaning of T is, however, different for each model.

Classical Test Model

In the classical test model the true score for a fixed person is defined as the expectation of his/her observed score across replications, and the error of measurement is the deviation of his/her observed score from this expectation. Usually, the model is formulated not for a fixed person but for a population of persons. In that case the true score, T, and the error of measurement are considered random variables, being random across, respectively, persons and replications. (For an introduction to the classical test model, see Lord & Novick, 1968.)

A result from the classical model needed in what follows is the linear regression of T on X. From classical test theory it can be shown that when the regression function of T on x may be assumed to be linear, it is equal to

$$E(T|x) = \rho_{XX'}x + (1 - \rho_{XX'})\mu_X, \qquad [10]$$

μ_X and $\varrho_{XX'}$ being the expected value and reliability coefficient of X (Lord & Novick, 1968, p. 65). Equation 10 is known as Kelley's regression line.

Having an external and directly measured criterion, say Ξ, concern will not be with the regression of T but of Ξ on X. Then, under the assumption of a linear regression function,

$$E(\Xi|x) = \mu_\Xi + (x - \mu_X)\rho_{X\Xi}\sigma_\Xi/\sigma_X, \qquad [11]$$

where μ and σ represent expected values and standard deviations and $\varrho_{X\Xi}$ is the correlation coefficient between X *and* Ξ.

It should be noted that apart from finite variances, the classical model does not involve distributional assumptions.

Beta-Binomial Model

When the process of a fixed student answering test items can be viewed as a sequence of Bernoulli trials—that is, trials (1) that have two possible outcomes, success and failure; (2) that have a probability of success constant for all trials; and (3) that are stochastically independent—the beta-binomial model seems to be a natural choice. The number of successes in a Bernoulli process follows the binomial distribution; hence, the conditional probability function of X given τ is

$$f(x|\tau) = \binom{n}{x}\tau^{x}(1-\tau)^{n-x}. \qquad [12]$$

Because of its flexible form and the ease with which it can be combined with Equation 12, the incomplete beta function ratio is often chosen as the *p.d.f.* of T,

$$\begin{aligned} g(\tau) &= B^{-1}(v,w-n+1)\tau^{v-1}(1-\tau)^{w-n} \qquad [13] \\ &\equiv \frac{\partial}{\partial\tau} I_{\tau}(v,w-n+1), \end{aligned}$$

where $B(v,w-n+1) \equiv \int_0^1 \tau^{v-1}(1-\tau)^{w-n}d\tau$ is the complete beta function (e.g., Johnson & Kotz, 1970, chap. 24), $v > 0$, and $w > n-1$.

It remains to indicate why the parameter τ in Equation 12 can be interpreted as a true score value. A possible explanation is that domain sampling is assumed and that τ is interpreted as the (expected) proportion a person should answer correctly when the entire domain is administered. Another explanation is that no item sampling is assumed but, as is usual in latent trait theory, that the item responses are considered the outcome of a stochastic process and that the probability of success is equal to τ for all items.

From Equation 12 and Equation 13, it follows that

$$h(x) = \binom{n}{x}B^{-1}(v,w-n+1)B(v+x,w-x+1), \qquad [14]$$

which is known as the negative hypergeometric distribution but also as the beta-binomial or the Pólya distribution. Keats and Lord (1962) have shown that simple moment estimators for v and w can be derived that are based on μ_X and the KR-21 reliability coefficient, and they have suggested that the fit of the test data to the beta-binomial model be checked by estimating Equation 14 and comparing it with the empirical observed score distribution. Proceeding in this way, they found satisfactory results for a wide range of differently skewed test score distributions. Model tests like the Keats-Lord test are only valid when the empirical distribution has been obtained independently of the data set used to estimate Equation 14.

For future reference, note that the cumulative distribution function of T given $X = x$ is known as

$$\begin{aligned} P(d|x) &= \int_0^d p(\tau|x)d\tau \qquad [15] \\ &= I_d(v+x,w-x+1). \end{aligned}$$

Since, for integer values of *v* and *w*,

$$I_d(v+x, w-x+1) = \sum_{\gamma=v+x}^{v+w} f(\gamma|d), \qquad [16]$$

where *f*(.|*d*) is the binomial probability function (Johnson & Kotz, 1969, sect. 3.8), Equation 15 can be obtained via a cumulative binomial table. Normal approximations are also available (Johnson & Kotz, 1970, sect. 24.6).

The conditions imposed on item difficulty by the beta-binomial test model are different for a deterministic and a stochastic conception of item responses. For both conceptions equal item difficulty is required, but this can be avoided in the case of the deterministic conception by giving separate samples to each person (van der Linden, 1979).

Bivariate Normal Model

Analogous to practice in (predictive) validity studies, the bivariate normal model sometimes seems a suitable approximation when relating test scores to an external criterion. Occasionally, this model is also used as the limiting form of the beta-binomial model after a variance-stabilizing transformation to the binomial parameter has been applied.

The model simply says that the distribution of (X, Ξ) follows the bivariate normal. Assuming that the X and Ξ scores are in their standardized form, this yields for the cumulative distribution function (*c.d.f.*) of Ξ, given $X = x$,

$$P(d|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{d} \exp\left[-\frac{(\xi-\rho x)^2}{2(1-\rho^2)}\right] d\xi, \qquad [17]$$

where $\rho \equiv \rho_{x\Xi}$ and ξ is a realization of Ξ. In the case of an internal criterion, the *c.d.f.* of T, given $X = x$, has the same form, but now $\rho \equiv \sqrt{\rho_{XX'}}$ and ξ must be replaced by τ. Note that the choice of the bivariate normal model involves an idealization, since it takes the number-correct score, X, as continuous.

Applications of Decision Theory to Criterion-Referenced Testing

The use of decision theory for optimizing sequences of mastery decisions will be considered in this section. In doing so, threshold, linear, and nomal loss functions will be considered. Although threshold loss functions have received the most attention, there may be many instances in which continuous loss functions, such as the linear and the normal ogive function, are to be preferred.

Threshold Loss

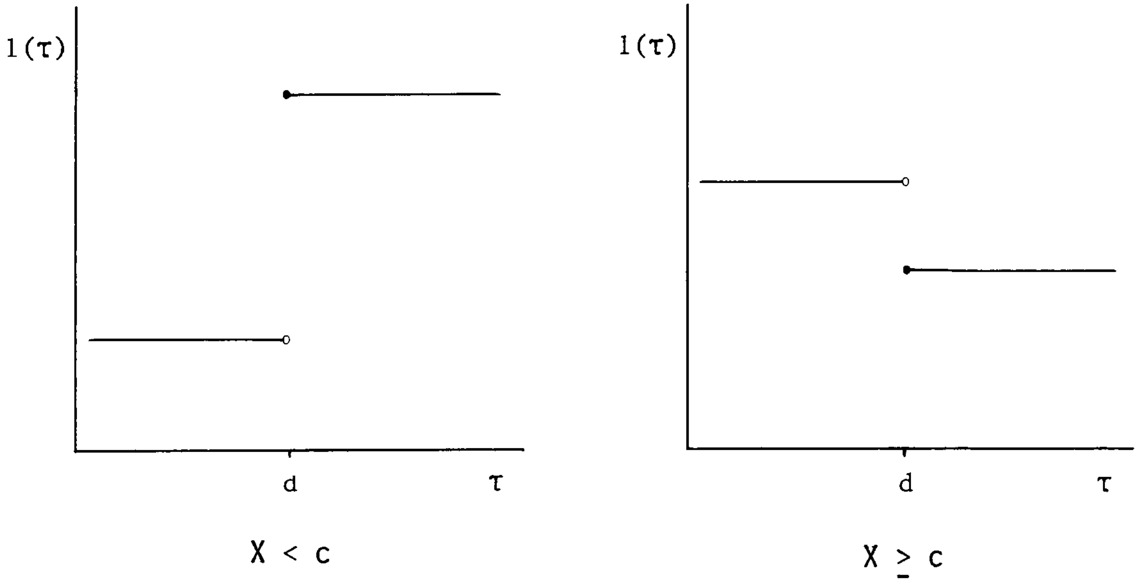
When a threshold loss function is chosen, it is assumed that the "seriousness" of all possible consequences of the decisions can be summarized by four constants, one for each of the four possible outcomes:

$$l(T) = \begin{cases} l_{00} & \text{for } T < d,\ X < c \\ l_{10} & \text{for } T \geq d,\ X < c \\ l_{01} & \text{for } T < d,\ X \geq c \\ l_{11} & \text{for } T \geq d,\ X \geq c. \end{cases} \qquad [18]$$

Threshold loss functions are often represented in fourfold tables; however, the author prefers a graphical form as in Figure 1, with one display for $X < c$ and $X \geqslant c$, showing a discontinuity at $T = d$.

First, an optimal cutoff score will be derived for unspecified distributions, only assuming a monotone likelihood ratio for the distribution of X given τ with respect to X, and then solutions for different probability models will be discussed.

Figure 1
An Example of a Threshold Loss Function



For a threshold loss function, the Bayes risk given in Equation 9 will take the form

$$B(c) = \sum_{x=0}^{c-1} \int_0^d l_{00}p(\tau|x)h(x)d\tau + \sum_{x=0}^{c-1} \int_d^n l_{10}p(\tau|x)h(x)d\tau \qquad [19]$$

$$+ \sum_{x=c}^{n} \int_0^d l_{01}p(\tau|x)h(x)d\tau + \sum_{x=c}^{n} \int_d^n l_{11}p(\tau|x)h(x)d\tau.$$

For convenience, and without loss of generality, rescale Equation 18 and choose $l_{00} = l_{11} = 0$, assuming positive values for l_{10} and l_{01} to satisfy the condition of monotonicity.

Now

$$B(c) = \sum_{x=0}^{c-1} \int_d^n l_{10} p(\tau|x)h(x)d\tau + \sum_{x=c}^{n} \int_0^d l_{01} p(\tau|x)h(x)d\tau \qquad [20]$$

$$= \sum_{x=0}^{c-1} \left[l_{10}\left[1 - P(d|x)\right]\right] h(x) + \sum_{x=c}^{n} \left[l_{01}P(d|x)\right]h(x),$$

where $P(d|x) = \int_0^{d} p(\tau|x)d\tau$. Thus, $B(c)$ consists of the sum of two expected losses, one for the false positive and the other for the false negative decisions. Adding terms to the first sum, and subtracting these from the second one,

$$B(c) = \sum_{x=0}^{n} \left[l_{10}\left[1 - P(d|x)\right]\right] h(x) + \qquad [21]$$

$$\sum_{x=c}^{n} \left[(l_{10} + l_{01})P(d|x) - l_{10}\right]h(x).$$

The solution now depends only on the second term, for the first term is independent of c. Since $l_{10} + l_{01} > 0$, $h(x) \geqslant 0$, and assuming that $P(d|x)$ is decreasing in x, $B(c)$ is minimal for the smallest value of c for which $(l_{10} + l_{01})P(d|x) - l_{10}$ is negative or, equivalently, for which

$$P(d|x) < \frac{l_{10}}{l_{10} + l_{01}} \qquad [22]$$

holds. (For the sake of completeness, note that this solution may not be unique when $h(x) = 0$ for some adjacent values of x, but this possibility will be further ignored.)

Since $P(d|x)$ is not observable, a psychometric model enabling its estimation is needed. If the beta-binomial model applies, Equation 15 can be used for this purpose. After its parameters have been estimated, a cumulative binomial table can be used to solve Equation 22 via Equation 16.

In case the bivariate normal distribution can be assumed, X is continuous and Equation 22 must be replaced by an equality. The optimal cutoff score on X, c^*, is obtained via Equation 17 as

$$c^* = \frac{d - z\sqrt{1 - \rho^2}}{\rho}, \qquad [23]$$

z being the value found when entering a cumulative normal table with the right-hand side of Equation 22 and, dependent on whether an internal or external criterion is used, $\varrho = \sqrt{\varrho_{XX'}}$ or $\varrho = \varrho_{X\Xi}$.

It is also possible to use Lord's Method 20, which is almost an entirely empirical method, to assess the distribution of (X,T) (Lord, 1969). Once this has been done, $P(d|x)$ can be checked for monotonicity and used to solve Equation 22. When the criterion is external, a sample distribution of (X,Ξ) can be used for this purpose. Although both methods seem to be attractive because they involve few assumptions, they are not recommended unless the sample sizes are large enough to guarantee stable solutions. Moreover, the condition of monotonicity will not be satisfied in many cases, so that extra assumptions, for example, to smooth $P(d|x)$, may be needed after all. Without additional distributional assumptions, no solutions to Equation 22 can be obtained for the classical test model.

Several authors have contributed to the theory of criterion-referenced testing with threshold loss functions. Hambleton and Novick (1973) have shown, with nonempirical Bayesian terminology, that decisions are optimal that grant mastery status to examinees with a likelihood ratio $P(d|x)/[1 - P(d|x)]$ smaller than the loss ratio l_{10}/l_{01}. It is easy to see that for criterion-referenced testing problems in which the decision rule may be considered to be monotone, this leads to the optimal cutoff score derived above. Hambleton and Novick's (1973) procedure reminds us that no absolute losses need be specified but that their ratio is sufficient. This can also be seen from Equation 22, by dividing the right-hand side by l_{01} or l_{10}.

To the author's knowledge, Alf and Dorfman (1967) were the first to use the bivariate normal model with threshold loss and to arrive at solution Equation 23. Their context, however, was aptitude testing with a future criterion measure.

Huynh (1976) has given results for the beta-binomial model corresponding with what has been derived above, but his approach was quite different. Although he introduced an external criterion, he defined his loss function on the true score continuum underlying the test. Only after assuming the relation between true score and the external criterion as a 0–1 function (students have a probability of 0 for achieving some specified level of performance on the external criterion up to some true score values, and equal to 1 thereafter), was he able to arrive at Equation 22.

Mellenbergh, Koppelaar, and van der Linden (1977) conducted a case study in which a threshold loss function and the beta-binomial model were used to optimize the decision rule for a number of criterion-referenced tests. Table 1 gives the estimated optimal cutoff scores $\hat{c}^*$ as well as the cutoff scores c actually used in the classroom for five tests from their study. Two more tests were analyzed, but these did not show a satisfactory fit to the model. The computer program used in this study (Koppelaar, van der Linden, & Mellenbergh, 1977) also produces estimates of the proportions of (mis)classifications to be expected when a cutoff score has been chosen, obtained by integrating and summating the product of Equation 12 and an estimate of Equation 13 over the proper ranges of T and X. Table 1 gives these estimated proportions for the cutoff scores actually used in the classroom.

Finally, Lindgren (1976, sect. 8.4.4) has shown, in the context of hypothesis testing, that Equation 22 is the solution to the more general problem of testing any two composite hypotheses.

Linear Loss

As can be seen in Figure 1, the threshold loss function shows a "threshold." It can be argued that in many situations this discontinuity at $T = d$ is an unrealistic representation of the loss actually incurred. Moreover, the threshold loss function assumes that for examinees to the left or to the right of d, the loss is constant, no matter how large their distance from d is, and this also seems unrealistic in many cases. In view of this, van der Linden and Mellenbergh (1977) proposed a linear loss function:

$$l(\tau) = \begin{cases} b_0(\tau - d) + a_0 & \text{for } X < c \\ b_1(d - \tau) + a_1 & \text{for } X \geqslant c, \qquad (b_0+b_1) > 0. \end{cases} \qquad [24]$$

Figure 2 displays an example of Equation 24. For the nonmastery decision the loss is increasing in τ, while it is decreasing for the mastery decision; this seems typical of many testing situations. For further interpretation of Equation 24 and its parameters, refer to van der Linden and Mellenbergh (1977).

Substituting Equation 24 in Equation 9, and using $E(T|x) = \int_0^n \tau p(\tau|x)d\tau$ and $\int_0^n p(\tau|x)d\tau = 1$, it follows that

$$\begin{aligned} B(c) &= \sum_{x=0}^{c-1} \int_0^n \left[b_0(\tau - d) + a_0\right] p(\tau|x)h(x)d\tau + \qquad [25] \\ &\quad \sum_{x=c}^{n} \int_0^n \left[b_1(d - \tau) + a_1\right] p(\tau|x)h(x)d\tau \\ &= \sum_{x=0}^{c-1} \left[b_0\left[E(T|x) - d\right] + a_0\right] h(x) - \\ &\quad \sum_{x=c}^{n} \left[b_1\left[E(T|x) - d\right] - a_1\right] h(x) \\ &= \sum_{x=0}^{n} \left[b_0\left[E(T|x) - d\right] + a_0\right] h(x) - \\ &\quad \sum_{x=c}^{n} \left[(b_0 + b_1)\left[E(T|x) - d\right] + (a_0 - a_1)\right] h(x). \end{aligned}$$

The first sum being a constant, Equation 25 is minimal if the second sum is maximal. Since $(b_0 + b_1) > 0$, $h(x) \geqslant 0$, and assuming that $E(T|x)$ is increasing in x, $B(c)$ is minimal if c is put equal to the smallest value of x for which

$$(b_0 + b_1)\left[E(T|x) - d\right] + (a_0 - a_1) \qquad [26]$$

Table 1
Results for Five Tests from the Beta-Binomial Model with Threshold Loss $l_{00}=l_{11}=0$, $l_{01}=l_{10}=1$

Statistic	Test A	 B	 C	 D	 E
No. of Students	127	106	163	147	150
No. of Items	18	20	20	19	20
c	14	16	16	15	16
$\hat{c}*$	15	17	20	17	17
$\hat{p}_{00}$	.410	.526	.914	.607	.418
$\hat{p}_{01}$	.217	.170	.067	.202	.138
$\hat{p}_{10}$	.043	.052	.005	.033	.057
$\hat{p}_{11}$	.330	.252	.015	.158	.387

is positive. (The possibility of nonunique solutions is again ignored.)

Analogous to the previous case, a model is needed to specify the regression function $E(T|x)$. A straightforward procedure is to adopt the classical test model with linear regression, that is, to substitute Kelley's regression line Equation 10 into Equation 26. This yields a value for $c*$ equal to the smallest integer larger than

$$\mu_X + \frac{d - (a_0 - a_1)/(b_0 + b_1) - \mu_X}{\rho_{XX'}} \qquad [27]$$

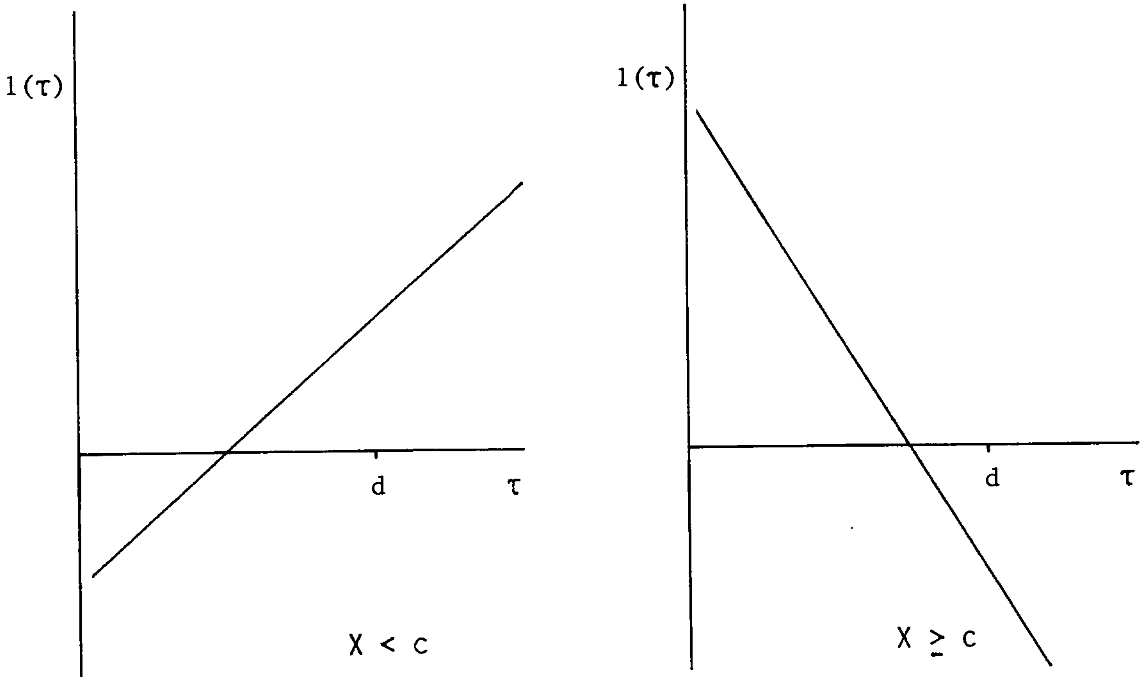
(van der Linden & Mellenbergh, 1977). Kelley's regression line is also implied by the bivariate normal and the beta-binomial model (in the latter case with KR-21 replacing $\varrho_{XX'}$; Lord & Novick, 1968, sect. 23.8), so Equation 27 is the solution for these models as well.

In the event of an external criterion, the regression line shown in Equation 11 is the proper choice, and Equation 27 is to be replaced by

$$\mu_X + \frac{d - (a_0 - a_1)/(b_0 + b_1) - \mu_\Xi}{\rho_{X\Xi}\sigma_\Xi/\sigma_X} . \qquad [28]$$

Just as with the threshold loss model, $E(T|x)$ and $E(\Xi|x)$ may be estimated using Lord's Method 20 or an empirical distribution of (X,Ξ); but, again, these procedures should only be used with large samples, or, better still, to check if assumptions such as Equations 10 and 11 prove to be reasonable.

Figure 2
An Example of a Linear Loss Function



Putting both loss lines in Equation 24 equal to each other, it appears that the T coordinate of the intersection is equal to $d - (a_0 - a_1)/(b_0 + b_1)$. Therefore, the solutions given by Equation 27 can be viewed as the first integer value to the right of the point yielding $d - (a_0 - a_1)/(b_0 + b_1)$ as "prediction" under the linear regression model. This fact, which holds for any regression model being increasing in x, is due to the elegant way the linear expectation operator and loss function combine in Equation 25.

When $a_0 = a_1 = a$, both loss lines intersect at $T = d$ and an interesting case arises. Then, all loss function parameters vanish from Equations 27 and 28, and, e.g., Equation 27 takes the form

$$\mu_X + \frac{d - \mu_X}{\rho_{XX'}} \qquad [29]$$

The practical meaning of this will be considered later in this paper. For an empirical illustration of the use of the linear loss, refer to van der Linden and Mellenbergh (1977).

Normal Ogive Loss

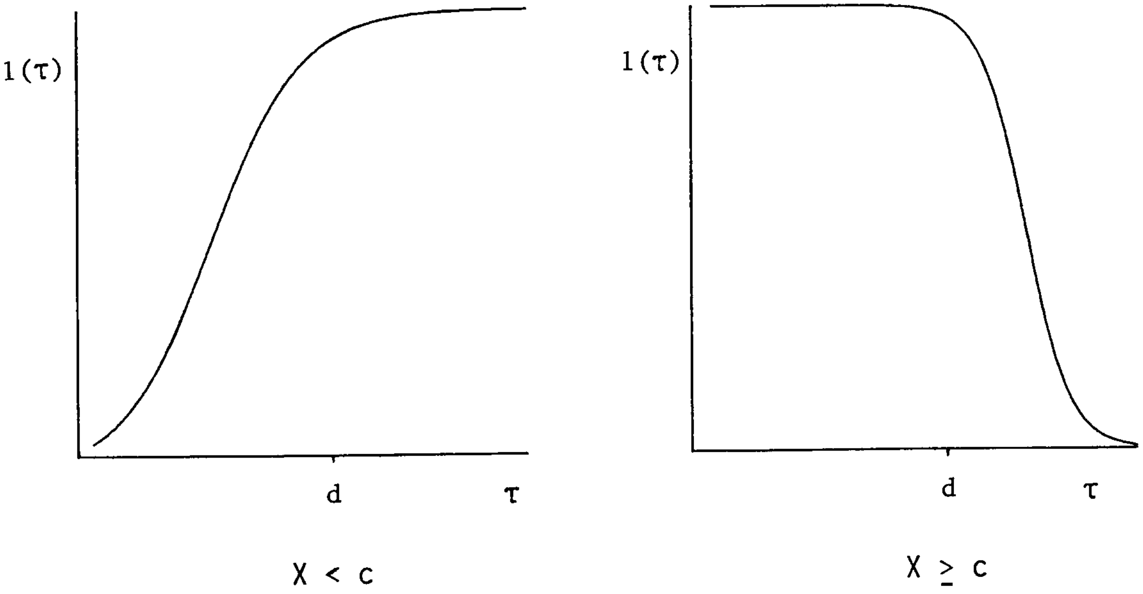
Another way to meet the objections to threshold loss has been proposed by Novick and Lindley (1978). They have recommended the choice of (de)cumulative normal distribution functions, which not only have realistic properties but also can be combined with a normal model for the test data.

For criterion-referenced testing, the natural choice seems to be

$$l(\tau) = \begin{cases} \Phi\left(\dfrac{\tau - \mu_0}{\sigma_0}\right) & \text{for } X < c \\ \\ \Phi\left(\dfrac{\mu_1 - \tau}{\sigma_1}\right) & \text{for } X \geq c, \end{cases} \qquad [30]$$

with Φ denoting the standardized normal distribution function; and μ_j and σ_j $(j = 0, 1)$, their location and scale parameters. Figure 3 depicts an example showing how the loss increases and decreases in τ when deciding for nonmastery and mastery, respectively.

Figure 3
An Example of a Normal Ogive Loss Function



Following Novick and Lindley, suppose that T given x is normally distributed with linear regression function $\alpha + \beta x$, variance σ^2, and homoscedasticity. From Equations 9 and 30, $B(c)$ is equal to

$$\sum_{x=0}^{c-1} \int_{-\infty}^{+\infty} \left[\Phi\left(\frac{\tau - \mu_0}{\sigma_0}\right) d\Phi\left(\frac{\mu - \alpha - \beta x}{\sigma}\right)\right] h(x) +$$

$$\sum_{x=c}^{n} \int_{-\infty}^{+\infty} \left[\Phi\left(\frac{\mu_1 - \tau}{\sigma_1}\right) d\Phi\left(\frac{\mu - \alpha - \beta x}{\sigma}\right)\right] h(x).$$

But, since in general

$$\int_{-\infty}^{+\infty} \Phi\left(\frac{u - t}{v}\right) d\Phi\left(\frac{s - t}{w}\right) = \Phi\left(\frac{u - s}{(v^2 + w^2)^{1/2}}\right)$$

(which follows when realizing that the left-hand side is the probability of the difference between two independent normally distributed random variables with different location and scale), it applies that

$$B(c) = \sum_{x=0}^{c-1} \Phi\left(\frac{\alpha + \beta x - \mu_0}{(\sigma^2 + \sigma_0^2)^{1/2}}\right) h(x) + \sum_{x=c}^{n} \Phi\left(\frac{\mu_1 - \alpha - \beta x}{(\sigma^2 + \sigma_1^2)^{1/2}}\right) h(x).$$

Thus,

$$B(c) = \sum_{x=0}^{n} \Phi\left(\frac{\alpha + \beta x - \mu_0}{(\sigma^2 + \sigma_0^2)^{1/2}}\right) h(x) - \qquad [31]$$

$$\sum_{x=c}^{n} \left[\Phi\left(\frac{\alpha + \beta x - \mu_0}{(\sigma^2 + \sigma_0^2)^{1/2}}\right) - \Phi\left(\frac{\mu_1 - \alpha - \beta x}{(\sigma^2 + \sigma_1^2)^{1/2}}\right)\right] h(x).$$

Now c^* is the smallest value of x for which

$$\Phi\left(\frac{\alpha + \beta x + \mu_0}{(\sigma^2 + \sigma_0^2)^{1/2}}\right) - \Phi\left(\frac{\mu_1 - \alpha - \beta x}{(\sigma^2 + \sigma_1^2)^{1/2}}\right)$$

or

$$\frac{\alpha + \beta x - \mu_0}{(\sigma^2 + \sigma_0^2)^{1/2}} - \frac{\mu_1 - \alpha - \beta x}{(\sigma^2 + \sigma_1^2)^{1/2}}$$

is larger than zero. This is the first integer larger than

$$\frac{\mu^* - \alpha}{\beta}, \qquad [32]$$

with $\mu^* = \Sigma_j w_j \mu_j / \Sigma_j w_j$ and $w_j = (\sigma^2 + \sigma_j^2)^{-1/2}$ (Novick & Lindley, 1978).

For an internal criterion, the linear regression function is Kelley's line, so that α and β follow from Equation 10, whereas for an external criterion they follow from Equation 11.

In Equation 30, μ_j governs the location of the normal ogive on the true score scale; and σ_j, its sensitivity to changes in true score value in the neighborhood of the location. Maximum sensitivity is attained for $\sigma_j \rightarrow \infty$. If also $\mu_j = d$, then Equation 30 approaches the threshold loss function with $l_{00} = l_{11} = 0$ and $l_{01} = l_{10} = 1$. No approach to threshold loss with different values for l_{01} and l_{10} is possible, unless different transformations on both parts of Equation 30 are applied.

Result Equation 32 can be related to the results derived under linear loss. Substituting Kelley's regression line in Equation 32, that is, putting $\alpha = (1 - \varrho_{XX'})\mu_x$ and $\beta = \varrho_{XX'}$, it appears that Equation 32 is equal to

$$\mu_X + \frac{\mu^* - \mu_X}{\rho_{XX'}} . \qquad [33]$$

Thus, μ^* and the intersection of both linear loss lines $d - (a_0 - a_1)/(b_0 + b_1)$ play an identical part in Equations 27 and 32 and 33. It is also interesting to note that when $\mu_0 = \mu_1$, the weights w_j cancel out and $\mu^* = \mu_0 = \mu_1$. When $\mu_0 = \mu_1 = d$, which seems to be a natural choice in criterion-referenced testing, solution Equations 32 and 33 reduce to the ordinary regression solution in Equation 29.

The Regression from the Mean Effect

In the previous section, the optimal cutoff scores were obtained by using the regression of T on x the other way around. In Equation 29, for example, take the true cutoff score d, go against the regression of T on x, and next choose the first integer value to the right. In Equations 27 and 32, T values different from d are chosen.

Ignoring the discrete character of c^* at this point, it follows from Equation 29 that $(d - \mu_T) = (c^* - \mu_X)\varrho_{XX'}$. Since according to classical test theory $\mu_X = \mu_T$, and in practice $\varrho_{XX'} < 1$, it follows that c^* is always further away from $\mu_X = \mu_T$ than d. For a fixed value of d this implies that the average performance of a population of students, $\mu_X = \mu_T$, and c^* are related negatively: The higher the average performance, the lower the optimal cutoff score. Hard-working populations are rewarded by low cutoff scores, while less hard-working populations will just be penalized and will be confronted with high cutoff scores. This is the opposite of what happens when norm-referenced standards are used. They vary up and down with the performances of the examinees. The behavior of c^*, which at first glance may seem counterintuitive, thus has to do with the fact that the presence of $\varrho_{XX'}$ in the denominator of Equation 29 "attenuates" the difference $d - \mu_T$ and causes what may be called a regression from the mean effect.

A comparable effect has been noted for the case of threshold loss by Mellenbergh, Koppelaar, and van der Linden (1977). The Bayes risk in Equation 22 is a linear combination of two joint probabilities—Prob $\{T \geqslant d, X < c\}$ and Prob $\{T < d, X \geqslant c\}$—with l_{10} and l_{01} as weights. If the performances of students go up, for example, then the former approaches Prob $\{X < c\}$ and the latter zero. Thus, the Bayes risk will tend to be minimal for a low value of c.

Internal and External Optimality of Mastery Decisions

In the preceding sections the problem was how to find cutoff scores on the test that are optimal for a given true cutoff score, loss function, and psychometric model. A different aspect of the criterion-referenced testing problem has been addressed by Mellenbergh and van der Linden (1979) and van der Linden and Mellenbergh (1978). Their problem was not to optimize mastery decisions but to assess how optimal they were once some cutoff score had been chosen.

As a basis for deriving an index for the optimality of the decision procedure, they chose the Bayes risk. This is negatively related to the quality of the decision procedure—the lower the Bayes risk, the better the procedure—and may take values outside the standard interval [0,1], which is the usual interval for coefficients for tests. Therefore, a rescaling was suggested:

$$\delta = 1 - (B - B_c)/(B_n - B_c) = (B_n - B)/(B_n - B_c). \quad [34]$$

In this coefficient δ, B is the Bayes risk as defined in Equation 4, and B_n and B_c are two reference points. B_c is the Bayes risk when complete information about the true scores is available; and B_n, when no information is available. The former (hypothetical) situation was formalized as a functional relation between X and T mapping the test values $0, 1, \ldots, n$ into the true score space. This function was left unspecified; it was only considered to be increasing in x. The situation of no information was represented by the assumption that X and T were distributed independently with *p.(d.)f.* $k(x,\tau)$ given by

$$k(x,\tau) = h(x)g(\tau). \quad [35]$$

For threshold loss function Equation 18, with $l_{00} = l_{11} = 0$ and $l_{01} = l_{10} = l$, the Bayes risk given in Equation 19 reduces to

$$B = (p_{01} + p_{10})l, \quad [36]$$

where p_{01} and p_{10} are the probabilities of a false positive and negative decision, respectively. From Equation 35, it can be seen that

$$B_n = (p_{0.}p_{.1} + p_{1.}p_{.0})l, \quad [37]$$

with $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$ $(i,j = 0, 1)$. Following van der Linden and Mellenbergh (1978),

$$B_c = \begin{cases} (p_{1.} - p_{.1})l & \text{for } p_{1.} \geqslant p_{.1} \\ (p_{.1} - p_{1.})l & \text{for } p_{1.} < p_{.1} \end{cases} \quad [38]$$

Substituting Equations 36, 37, and 38 into Equation 34, it appears that δ is equal to the well-known coefficient H of Loevinger:

$$\delta = \begin{cases} (p_{11} - p_{1.}p_{.1})/p_{.1}p_{0.} & \text{for } p_{1.} \geq p_{.1} \\ \\ (p_{11} - p_{1.}p_{.1})/p_{1.}p_{.0} & \text{for } p_{1.} < p_{.1} \end{cases} \qquad [39]$$

In the case of an internal criterion and test scores fitting the beta-binomial model, the proportions p_{ij} in Equation 39 can be estimated using the computer program by Koppelaar, van der Linden, and Mellenbergh (1977). If the scores do not fit, a less restrictive model such as Lord's (1969) Method 20 can be tried. When an external criterion is present, the proportions in Equation 39 can be estimated from an empirical distribution of (X,Ξ). Mokken (1971, sect. 4.3) gives approximate sampling distributions of Loevinger's H for the null as well as the nonnull case. These can be used for testing hypotheses regarding Equation 39 or establishing confidence intervals.

Equation 25 gives the Bayes risk for the linear loss function considered earlier in this paper. Assume that a model with linear regression holds; $E(T|x)$ in Equation 25 can therefore be considered to be Kelley's regression line (see Equation 10). Since stochastic independence implies linear stochastic independence, Kelley's line is, under condition Equation 35, equal to

$$E(T|x) = E(X). \qquad [40]$$

From van der Linden and Mellenbergh (1978) it can be seen that under functional dependency between X and T, Equation 10 reduces to

$$E(T|x) = x. \qquad [41]$$

Substituting Equations 40 and 41 into Equation 25 gives B_n and B_c, respectively, and substituting these, in turn, into Equation 34 shows that

$$\delta = \rho_{XX'}. \qquad [42]$$

Similarly, it can be verified that for an external criterion

$$\delta = \rho_{X\Xi} \qquad [43]$$

(Mellenbergh & van der Linden, 1979). Thus, when test models with linear regression functions and loss functions in the form given by Equation 24 are an appropriate choice, the well-known reliability and validity coefficient from classical test theory can be viewed as a standardization of the Bayes risk incurred in the decision procedure and can serve as suitable coefficients for the optimality of the decision rule.

The standardization in Equation 38 guarantees that δ will be in the interval [0,1] whenever B is between B_c and B_n. The above results show that this condition holds for threshold and linear loss functions. Wilcox (1978) has given a standardization using the greatest lower and least upper bound of the Bayes risk for which this condition generally holds. De Gruijter (1978) has used a standardization which, for a linear loss function, resulted in a coefficient to be interpreted as the point-biserial analogue of Livingston's (1972) criterion-referenced reliability coefficient.

New Lines of Research

There are two new lines of psychometric research arising from the application of decision theory to criterion-referenced test data. The first is the extension of the decision-theoretic point of view to other aspects of criterion-referenced testing than optimizing mastery decisions. The second is research into appropriate loss functions.

The preceding section, which dealt with the assessment of decision optimality, is an example of the former line of research. What is new about coefficient δ is that it gives the Bayes risk a central place in the analysis of test-based decisions, "elbowing out" the reliability and consistency coefficients that have taken this place so far. Reliability coefficients are of restricted meaning, only applicable when measurements are considered and squared error loss is the appropriate choice. When decisions are considered, however, and other loss functions seem more realistic, the analyses ought to be based on the Bayes risk. This point of view is being developed for several other technical problems associated with criterion-referenced tests, such as test score equating, item analysis, and test length determination.

Another example can be found in the issue of optimizing treatment assignment in view of future mastery decisions. Many programs applying criterion-referenced tests are organized according to principles of individualized instruction. These programs typically consist of a series of small units or modules in which the students are allowed to take different routes but all are expected to pass the same end-of-unit test. If the assignment of students to different instructional routes or treatments is based on aptitude testing, decision theory can be used to optimize treatment assignment as well and to incorporate it within the framework of criterion-referenced testing technology (van der Linden, 1981).

The second line is research into appropriate loss functions for criterion-referenced testing. Several techniques for scaling losses are available. Most texts on decision theory contain a chapter on utility theory in which lottery methods are proposed for this purpose (see Luce & Raiffa, 1957, chap. 2); but, in principle, any psychological scaling method can be used. The point is, however, that these techniques do not automatically lead to elegant loss functions and optimal cutoff scores. It may be wise, therefore, to use these techniques not without a prior chosen mathematical form of the loss function. This model should be realistic and fit not only the utilities of the decision-maker but also the psychometric model needed to explain the test scores. In this respect, Novick and Lindley's (1978) plea for *c.d.f.*'s as loss functions that are the "natural conjugate" of the psychometric model is a most important development.

Loss functions must also be as robust as possible with respect to the mastery decisions. Preferably, large differences in specifying the loss function should lead to small differences in the cutoff score (or in the Bayes risk associated with the procedure). As indicated earlier, the linear loss function (Equation 24) shows an interesting result when both parameters a_0 and a_1 can be considered equal to each other. Then, Equation 28 is the optimal cutoff score. The practical meaning is that under the restriction $a_0 = a_1$, the linear loss function is maximally robust with respect to c^*: For all possible values of b_0, b_1, and a, c^* assumes the same value. A comparable situation arises for normal ogive loss with the restriction $\mu_0 = \mu_1 = d$.

It should be noted that robustness of a loss function with respect to the optimal cutoff score is not an "absolute" property, i.e., a property of the loss function alone, but something that arises by the way loss function and test model combine into the Bayes risk. The same loss function may lead to robust results when combined with one model but may lose its properties when this model is replaced by a model with different structure or parameter values. This is illustrated in Table 2 where the same threshold loss function, with loss ratio $\lambda = l_{01}/l_{10}$ and $l_{00} = l_{11} = 0$, was used in combination with the

Table 2
Optimal Cutoff Scores for Varying Threshold Loss with the Beta-Binomial Model and Emrick's Model

	Loss Ratio				
$\lambda = l_{01}/l_{10}$	.33	.50	1	2	3
Beta-Binomial Model (v=14.47, w=21.31)	14	14	15	16	17
Emrick's Model (α=.25, β=.75, μ=.70)	10	10	10	10	11

beta-binomial model and with Emrick's mastery testing model (Emrick, 1971). The first row shows a result from the case study by Mellenbergh, Koppelaar, and van der Linden (1977) mentioned earlier in this paper. The test was composed of 19 three-choice items that yielded parameter estimates for the beta-binomial model equal to v = 14,47 and w = 21.31. For the loss ratio values in Table 2, the optimal cutoff score varies between 14 and 17. The second row is from a monte carlo experiment with Emrick's (1971) latent class model in which it was noted that from a certain test length, the optimal cutoff score is rather insensitive to differences in loss ratio values (van der Linden, 1980). A 20-item test with parameters α =.25, β = .75, and μ = .70, yields, for the same loss ratio values, an optimal cutoff score that is always equal to 10 with the exception of $\lambda = 3$, where it is equal to 11.

Thus, loss functions must not only fit the decision-maker's utilities but must also be easy to combine with the psychometric model and must give rise to results that are maximally robust. More research is needed to find functions and models meeting these requirements as simultaneously and as satisfactorily as possible.

Conclusion

In this paper approaches to criterion-referenced testing based on a Neyman-Pearson approach (e.g., Fhanér, 1974; Kriewall, 1972; Millman, 1973; Wilcox, 1976) have been disregarded. In these approaches an indifference zone instead of a true cutoff score is specified, and the cutoff score is found testing the hypothesis that the student is at the lower bound against the hypothesis that he/she is at the upper bound of this zone. Although these approaches do not involve loss functions and prior probabilities, it can be shown that from a Bayesian point of view, they are suboptimal unless the decision-maker is willing to accept certain losses and prior probabilities (Lindgren, 1976, sect. 8.4.4).

Apart from the example in Table 2, latent class models for criterion-referenced testing (e.g., Besel, 1973; Emrick, 1971; Macready & Dayton, 1977) have also been disregarded. These models follow a decision-theoretic approach but assume a latent class instead of a continuum conception of mastery. An advantage of latent class models over the models in this paper may seem that there is no need for setting a true cutoff score. By fitting a latent class model, nature indicates who is a master and who is not, and all that is necessary to find an optimal cutoff score on the test is an appropriate loss function.

It should be noted that, strictly speaking, there is no need for setting a true cutoff score in the models in this paper. The threshold loss function Equation 18 may be set at other points than $T = d$

or even at different points for $X < c$ and $X \geqslant c$, and an optimal cutoff score can still be derived. The linear loss function can be reparameterized into a function without d as a parameter. The normal ogive function does not even contain d as a parameter though choosing $\mu_0 = \mu_1 = d$ has been proposed. The only reason not to do this is that it seems unreasonable when decision models are applied to improve mastery decisions. Nevertheless, criterion-referenced testing without a true cutoff score is also possible for the models in this paper, and in this respect there is no difference between state and continuum models for criterion-referenced testing.

Reference

Alf, E. F., & Dorfman, D. D. The classification of individuals into two criterion groups on the basis of a discontinuous pay-off function. *Psychometrika*, 1967, *32*, 115-123.

Besel, R. *Using group performance to interpret individual responses to criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, February-March 1973. (EDRS No. ED 076 658.)

Cox, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. *Journal of Educational Measurement*, 1966, *3*, 147-150.

Davis, C. E., Hickman, J., & Novick, M. R. *A primer on decision analysis for individually prescribed instruction* (ACT Technical Bulletin No. 17). Iowa City, IA: The American College Testing Program, 1973.

DeGroot, M. H. *Optimal statistical decisions*. New York: McGraw-Hill, 1970.

de Gruijter, D.N.M. A criterion-referenced point biserial correlation coefficient. *Tijdschrift voor Onderwijsresearch*, 1978, *3*, 257-261.

Emrick, J. A. An evaluation model for mastery testing. *Journal of Educational Measurement*, 1971, *8*, 321-326.

Ferguson, T. S. *Mathematical statistics: A decision theoretic approach.* New York: Academic Press, 1967.

Fhanér, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 172-175.

Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, *15*, 237-261.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. *Journal of Experimental Education*, 1976, *45*, 57-64.

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

Huynh, H. Statistical considerations of mastery scores. *Psychometrika*, 1976, *42*, 65-79.

Johnson, N. L., & Kotz, S. *Distributions in statistics: Discrete distributions.* Boston: Houghton Mifflin, 1969.

Johnson, N. L., & Kotz, S. *Distributions in statistics: Continuous univariate distributions — 2.* Boston: Houghton Mifflin, 1970.

Kalin, S., & Rubin, H. Distributions possessing a monotone likelihood ratio. *Journal of the American Statistical Association*, 1956, *1*, 637-643.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika*, 1962, *27*, 59-72.

Koppelaar, H., van der Linden, W. J., & Mellenbergh, G. J. A computer program for classifications in dichotomous decisions based on dichotomously scored items. *Tijdschrift voor Onderwijsresearch*, 1977, *2*, 32-37.

Lewis, C., Wang, M., & Novick, M. R. Marginal distributions for the estimation of proportions in *m* groups. *Psychometrika*, 1975, *40*, 63-75.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. *Illinois School Research*, 1972, *9*, 5-18.

Lindgren, B. W. *Statistical theory* (3rd ed.). New York: Macmillan, 1976.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika*, 1969, *34*, 259-299.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Luce, R. D., & Raiffa, H. *Games and decisions*. New York: John Wiley & Sons, 1957.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 1977, *2*, 99-120.

Mellenbergh, G. J., Koppelaar, H., & van der Linden, W. J. Dichotomous decisions based on dichotomously scored items: A case study. *Statistica Neerlandica*, 1977, *31*, 161-169.

Mellenbergh, G. J., & van der Linden, W. J. The internal and external optimality of decisions based on tests. *Applied Psychological Measurement*, 1979, *3*, 259-273.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Mokken, R. J. *A theory and procedure of scale analysis*. The Hague: Mouton, 1971.

Novick, M. R., Lewis C., & Jackson, P. H. The estimation of proportions in *m* groups. *Psychometrika*, 1972, *38*, 19-46.

Novick, M. R., & Lindley, D. V. The use of more realistic utility functions in educational applications. *Journal of Educational Measurement*, 1978, *15*, 181-191.

Raiffa, H., & Schlaifer, R. *Applied statistical decision theory*. Boston: Division of Research, Harvard Business School, 1961.

Robbins, H. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley: University of California Press, 1956.

Robbins, H. The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 1964, *35*, 1-20.

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

van der Linden, W. J. Binomial test models and item difficulty. *Applied Psychological Measurement*, 1979, *3*, 401-411.

van der Linden, W. J. Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery score. *Psychometrika*, 1981, *45*, in press.

van der Linden, W. J. *Estimating the parameters of Emrick's mastery testing model*. Submitted for publication, 1980.

van der Linden, W. J. & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, 1977, *1*, 593-599.

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

Wilcox, R. R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, *1*, 359-364.

Wilcox, R. R. A note on decision-theoretic coefficients for tests. *Applied Psychological Measurement*, 1978, *2*, 609-613.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago, IL: MESA Press, 1979.

Yamane, T. *Statistics: An introductory analysis* (3rd ed.). New York: Harper & Row, 1973.

Acknowledgment

Much of the research reported here has been carried out with Gideon J. Mellenbergh. The author gratefully acknowledges this cooperation as well as his comments on an earlier draft of this paper. Thanks are also due to Dato N.M. de Gruijter, Ivo W. Molenaar, and Robert F. van Nearssen for their comments, and to Paula Achterberg for typing the manuscript.

Author's Address

Wim J. van der Linden, Afdeling Toegepaste Onderwijskunde, Technische Hogeschool Twente, Postbus 217, 7500 AE Enschede, The Netherlands.