Coefficients for Tests from a Decision Theoretic Point of View

Wim J. van der Linden
Twente University of Technology

Gideon J. Mellenbergh
University of Amsterdam

From a decision theoretic point of view a general coefficient for tests, δ, is derived. The coefficient is applied to three kinds of decision situations. First, the situation is considered in which a true score is estimated by a function of the observed score of a subject on a test (point estimation). Using the squared error loss function and Kelley's formula for estimating the true score, it is shown that δ equals the reliability coefficient from classical test theory. Second, the situation is considered in which the observed scores are split into more than two categories and different decisions are made for the categories (multiple decision). The general form of the coefficient is derived, and two loss functions suited to multiple decision situations are described. It is shown that for the loss function specifying constant losses for the various combinations of categories on the true and on the observed scores, the coefficient can be computed under the assumptions of the beta-binomial model. Third, the situation is considered in which the observed scores are split into only two categories and different decisions are made for each category (dichotomous decisions). Using a loss function that specifies constant losses for combinations of categories on the true and observed score and the assumption of an increasing regression function of t on x, it is shown that coefficient δ equals Loevinger's coefficient H between true and observed scores. The coefficient can be computed under the assumption of the beta-binomial model. Finally, it is shown that for a linear loss function and Kelley's formula for the regression of the true score on the observed score, the coefficient equals the reliability coefficient of classical test theory.

Decision Situations in Testing

Psychological and educational tests can be considered as instruments for making decisions. In research the decision problem is primarily the estimation of a true score, t, using a function of the observed score, x, of a subject. The problem can be described as *point estimation* of the parameter t of a person; the possible actions of the decision maker are all possible estimated values for the true score (Ferguson, 1967, p. 11). The decision rule is, in this case, the procedure used for estimating the true score. For example, in classical test theory the decision rules used are either estimating the true score by the value of the observed score or using Kelley's formula for estimating the true score (Lord & Novick, 1968, p. 63).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 119–134
© Copyright 1978 West Publishing Co.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

In applied fields the decision problem is the classification of persons into different categories; the possible actions of the decision maker are the different categories into which persons are classified. If there are more than two categories, the problem is called a *multiple decision* problem (Ferguson, 1967, p. 10). A common decision problem is classification of people into only two categories: accepted or rejected, for example, pass-fail decisions in education or acceptance-rejection decisions for applicants for jobs or for special treatments, such as psychotherapy or remedial teaching. Such problems will be called *dichotomous decision problems* (Mellenbergh, Koppelaar, & Van der Linden, 1976).

For solving decision problems, a loss function $L(t,x)$ is needed that depends on both the true score and the observed score; this function specifies the loss of the decision maker using a certain decision rule. For instance, in dichotomous decision problems the loss function specifies the loss of the decision maker for "accepted, suitable," "accepted, not suitable," "rejected, suitable," and "rejected, not suitable" subjects. Given t, the risk of a decision rule is the expected value of the loss with respect to the distribution of x (Ferguson, 1967, p. 7). The Bayes risk is the expected value of the risk with respect to the distribution of t (Ferguson, 1967, p. 31). The result is that the Bayes risk can be considered as the expected value of the loss with respect to the joint distribution of the random variables t and x in a given population of subjects: $R = E\,L(t,x)$. Throughout the remainder of the paper, the Bayes risk will be referred to as "the risk."

Much attention has been paid recently in psychometrics to dichotomous decision situations, particularly for mastery decisions in criterion-referenced measurement (Hambleton & Novick, 1973; Huyhn, 1976b; Meskauskas, 1976). Meskauskas (1976) distinguished between State and Continuum models for mastery decisions. In State models the true score is considered an all or none variable, representing either mastery or non-mastery of the subject matter. In Continuum models the true score is considered a continuous variable which represents the degree of mastery of the subject matter. Macready and Dayton (1977) have developed State models from a decision theoretic point of view; Huyhn (1976b) has considered mastery decisions with a continuous true score. Huyhn (1976a, 1976c) has also described coefficients for tests in dichotomous decision situations. In this article a general strategy for constructing coefficients for tests from a decision theoretic point of view is described, with the true score considered a continuous variable.

Coefficients for Tests

In decision theory the risk can be considered as an index of the quality of the decision: the smaller the risk, the better the decision rule. Therefore, risk is an appropriate basis for deriving a decision-oriented coefficient for tests. Two disadvantages must be removed, however. First, it is conventional in test theory to define indices so that the scale has a direction opposite to that in which the risk is represented. Second, although in test theory indices are nearly always defined on the standard interval [0,1], the range of possible values for the risk can be different. Both disadvantages can be removed by defining two reference points, such as R_c and R_n, which have a decision theoretic interpretation: R_c and R_n are the risks in the situation in which the test contains, respectively, complete and no information about the true score. Using these two reference points, an interpretable index for decision situations is

$$\delta = 1 - (R - R_c)/(R_n - R_c) = (R_n - R)/(R_n - R_c). \qquad [1]$$

Owing to this linear transformation, the scale of δ has a direction opposite to the direction of the scale in which R is represented. Furthermore, δ will be in the conventional interval from 0 to 1 when-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

ever R is in the interval from R_c to R_n. The true score, t, is defined on the closed interval from 0 to 1; the distribution of t is $g(t)$. It is assumed that the test is composed of n items scored either 0 or 1, and the sum of the item scores is the observed score, x. In this case, x is defined on the set of integers: 0, 1, 2, . . . n; the distribution of x is $h(x)$. If $k(t,x)$ is the bivariate distribution of x and t, the risk is

$$R = E\ L(t,x) = \sum_{x=0}^{n} \int_{0}^{1} L(t,x)\ k(t,x)\ dt. \qquad [2]$$

For fixed distributions $h(x)$ and $g(t)$, R_n is defined as the risk in the situation in which the observed score contains no information about the true score. This notion is formalized by the statement that x and t are distributed independently:

$$k(t,x) = g(t)\ h(x). \qquad [3]$$

For a fixed distribution $h(x)$, R_c is defined as the risk in the situation in which the observed score contains complete information about the true score. This notion is formalized by the statement that t is an increasing function of x which maps the set of integers 0, 1, 2, . . . n into the closed interval from 0 to 1.

Note that throughout this paper $h(x)$ is considered to be fixed. The reason for the assumption that the distribution of x under conditions of stochastic independence and functional dependence has the same form as the observed distribution $h(x)$ is that coefficient δ is an evaluation of the given decision procedure for a specified test, test administration, $h(x)$, and cutting score c. Coefficient δ is derived by a thought-experiment in which the conditions of stochastic independence and functional dependence are merely hypothetical conditions introduced to standardize the risk R, which is incurred by using the given decision procedure. Therefore, the assumption of a fixed $h(x)$ is not empirical but hypothetical and introduced for standardization purposes only.

In general, R is not necessarily in the interval from R_c to R_n,; and consequently, δ is not in the interval from 0 to 1. The reason is that R, R_c, and R_n depend on the loss function and on the bivariate distribution of t and x. However, R_c and R_n can be considered as rather interpretable points for the risk. When the test contains complete information about the true score, R_c is the risk. When the test contains no information about the true score, R_n is the risk. In the remainder of this article, δ is derived for some cases. In some very important cases the risk is in the interval from R_c to R_n; therefore, δ has a clear interpretation: a value of 0 signifies that the test is worthless, and a value of 1 signifies that the test is perfect for the decision situation.

Point Estimation

When the decision problem is estimating the parameter t, one must seek a function $p(x)$ that is a suitable point estimator for t. The function $p(x)$ is the decision rule: it gives for each possible value of x an estimated value $p(x)$ for t. Here, interest is not in finding point estimators with recognized statistical properties, but in the assessment of the decision theoretic qualities of given point estimators.

For evaluating a point estimator, it is necessary to specify a loss function that provides a weighing of possible errors of estimation. Defining an error of estimation as the difference between the parameter t and the function $p(x)$ used for estimating t, the loss is often a function of $p(x)$ and t:

$$L(t,x) = L(p(x) - t). \qquad [4]$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Substitution of this loss function into Equation 2 yields a general form for the risk of a point estimator:

$$R = \sum_{x=0}^{n} \int_{0}^{1} L(p(x) - t)\, k(t,x)\, dt. \qquad [5]$$

This risk will be the starting point for the evaluation of a given point estimator with the decision-oriented coefficient δ given in Equation 1. First, the procedure is described for the regression function of t on x; this regression function, $E(t|x)$, can be used as a point estimator for the parameter t. Second, the result of the procedure is shown for the linear regression function of t on x (Kelley's formula for estimating the true score). In both cases, the loss function is the squared error loss function

$$L(t,x) = (p(x) - t)^2 . \qquad [6]$$

which is well-known from its use in statistics (Novick & Jackson, 1974, chap. 1).

For the regression function, $E(t|x)$, and loss function in Equation 6 the risk is

$$R = \sum_{x=0}^{n} \int_{0}^{1} (E(t|x) - t)^2\, k(t,x)\, dt. \qquad [7]$$

In the situation in which x and t are stochastically independent, Equation 3 can be substituted into Equation 7; and the risk is

$$R_n = \sum_{x=0}^{n} \int_{0}^{1} (E(t|x) - t)^2\, g(t)\, h(x)\, dt. \qquad [8]$$

In the situation in which x and t are functionally dependent, such as $t = \phi(x)$, the conditional distribution of t, given x, degenerates to a single value of t, given x. Therefore, when t is functionally dependent on x, $E(t|x)$ will be equal to $t = \phi(x)$; and

$$R_c = \sum_{x=0}^{n} \int_{0}^{1} (\phi(x) - t)^2\, k(t,x)\, dt = 0. \qquad [9]$$

Substituting Equations 7, 8, and 9 into Equation 1 yields the coefficient δ for the regression function, $E(t|x)$. Computation of δ requires that the regression function, $E(t|x)$, and the probability densities $k(t,x)$, $g(t)$, and $h(x)$ can be estimated; there is no psychometric theory for estimating these functions simultaneously without assumptions about the form of the regression function.

Using Kelley's linear regression function,

$$E_\ell(t|x) = \rho_{xx'}(x/n) + (1 - \rho_{xx'})\, E(x/n). \qquad [10]$$

as a point estimator for t implies that the classical test model and the linearity of $E(t|x)$ are taken for granted (Lord & Novick, pp. 64–65). The risk that is connected with the estimator in Equation 10 and the loss function in Equation 6 is

$$E\,[E_\ell(t|x) - t]^2. \qquad [11]$$

This is the definition of the squared standard error of estimation

$$\sigma_e^2 = \sigma_t^2 (1 - \rho_{xx'}) \qquad [12]$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

(Lord & Novick, p. 67). The general form of the risk is thus

$$R = \sigma_t^2 \, (1 - \rho_{xx'}). \qquad [13]$$

Stochastic independence of t and x implies linear stochastic independence:

$$\rho_{xt} = \rho_{xx'} = 0. \qquad [14]$$

Therefore, in the situation wherein t and x are stochastically independent Equation 13 has the form

$$R_n = \sigma_t^2. \qquad [15]$$

It was stated earlier that for an unspecified regression function Equation 9 applies when t and x are functionally dependent. In the case of the linear regression function (Equation 10), the same result is thus valid for R_c:

$$R_c = 0. \qquad [16]$$

Substituting Equations 13, 15, and 16 into Equation 1 yields

$$\delta = \{\sigma_t^2 - \sigma_t^2 \, (1 - \rho_{xx'})\}/\sigma_t^2 = \rho_{xx'}, \qquad [17]$$

Therefore, in estimating the true score with Kelley's formula, the well-known reliability coefficient can be interpreted as a standardized risk; its standardization is given by Equation 1.

Multiple Decisions

The true score is defined on the closed interval from 0 to 1. It is assumed that this interval is divided into $(k+1)$ disjoint intervals: $[0, d_1]$, $[d_1, d_2]$, . . . $[d_k, 1]$; these intervals are denoted as $D_0, D_1, \ldots, D_k$. The observed score x is defined on the set of integers: $0, 1, \ldots n$. It is assumed that this set is divided into the disjoint sets: $\{0, 1, \ldots c_1-1\}$, $\{c_1, c_1+1, \ldots c_2-1\}$, . . . $\{c_k, c_k+1, \ldots n\}$; these sets are denoted as $C_0, C_1, \ldots C_k$.

In multiple decision situations, the problem is to find cutting scores that, for a given loss function, optimally divide x into the sets C_i. For multiple decision situations, it seems to make sense to split the loss function $L(t,x)$ into $(k+1)^2$ separate continuous functions for each combination of an interval D_i and a set C_j:

$$L(t,x) = \begin{cases} L_{00} & t \in D_0, \; x \in C_0 \\ \vdots & \\ L_{0k} & t \in D_0, \; x \in C_k \\ L_{10} & t \in D_1, \; x \in C_0 \\ \vdots & \\ L_{1k} & t \in D_1, \; x \in C_k \\ L_{20} & t \in D_2, \; x \in C_0 \\ \vdots & \\ L_{kk} & t \in D_k, \; x \in C_k \end{cases} \qquad [18]$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Here, it will be required that the function L_{ij} is a continuous function depending only on t for each combination of an interval D_i and a set C_j. The meaning is that the loss function for a true score from D_i and an observed score from C_j only depends on the true score. From a decision theoretic point of view this makes sense: the same decision is made for all observed scores from C_j. Therefore, within C_j the loss function L_{ij} should not depend on the observed score, but only on the true score. Using the notation Σ_{C_j} for summation over the values of the observed score in the set C_j and $\int_{D_i}$ for the integral over the true score in the interval D_i, the general form of the risk for multiple decision is

$$R = E\, L(t,x) = \sum_{i=0}^{k} \sum_{j=0}^{k} \sum_{C_j} \int_{D_i} L_{ij}(t)\, k(t,x)\, dt. \qquad [19]$$

In applications the loss functions $L_{ij}(t)$ should be further specified. An example of a possible set of functions that seems appropriate for multiple decisions is

$$\begin{aligned} L_{ij}(t) = \ell_{ij} \qquad & \ell_{00} \leq \ell_{01} \leq \cdots\cdots\cdots\cdots \leq \ell_{0k} \\ & \ell_{10} \geq \ell_{11};\ \ell_{11} \leq \ell_{12} \cdots\cdots\cdots \leq \ell_{1k} \\ & \ell_{20} \geq \ell_{21} \geq \ell_{22};\ \ell_{22} \leq \ell_{23} \leq \cdots \leq \ell_{2k} \\ & \ell_{k0} \geq \ell_{k1} \geq \cdots\cdots\cdots\cdots \geq \ell_{kk}. \end{aligned} \qquad [20]$$

The loss is constant for each combination of D_i and C_j. It is smallest for the combinations D_i and C_i and increases when C_j is "further away" from D_i. Using the notation P_{ij} for the probability that the true score is in the interval D_i and that the observed score is in the set C_j, the risk for this loss functions is

$$R = \sum_{i=0}^{k} \sum_{j=0}^{k} \ell_{ij}\, P_{ij}. \qquad [21]$$

Using the notation $P_{j\cdot} = \sum_{i=0}^{k} P_{ij}$ and $P_{i\cdot} = \sum_{j=0}^{k} P_{ij}$, it follows from Equation 3 that

$$R_n = \sum_{i=0}^{k} \sum_{j=0}^{k} \ell_{ij}\, P_{j\cdot}\, P_{i\cdot}. \qquad [22]$$

The expression for R_c is more complicated. Whenever t has a functional relation to x, the bivariate distribution of t and x is degenerate with no scatter about the regression of t on x. Graphically, the functional relation $t = \phi(x)$ is a graph containing points of the original space, $t \times x$, from which the multiple decision table has been constructed by simultaneous partitioning along the t-axis and x-axis. A subset $D_i \times C_j = [d_i, d_{i+1}] \times \{c_j, c_{j+1}-1\}$ of the original space corresponds to every cell (i,j) of this table. Dependent on the functional form of $t=\phi(x)$ and on the cutting scores d_i and c_j, a cell (i,j) may or may not contain points of $t=\phi(x)$. Only when a cell (i,j) contains points of $t=\phi(x)$ can it have a non-zero probability of occurrence and a non-zero contribution to the risk. In the case of functional dependence between t and x, the following is true for the bivariate density:

$$k(t,x) = h(x). \qquad [23]$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Therefore, the contribution of cell *(i,j)* to the risk R_c can be written as

$$r_{ij} = \sum_{S_{ij}} \ell_{ij} h(x). \qquad [24]$$

where S_{ij} is the set of all values of x for which $[x, \phi(x)]$ is a point of $D_i \times C_j$. Summation over all cells of the decision table yields

$$R_c = \sum_{i=0}^{k} \sum_{j=0}^{k} r_{ij} \qquad [25]$$

and substituting Equations 21, 22, and 25 into Equation 1 gives the decision coefficient δ for multiple decisions with the loss function of Equation 20 and a regression function as yet unspecified.

In situations in which the linear regression function in Equation 10 can be assumed, $t = \phi(x)$ equals x/n in the case of functional dependence; and the sets S_{ij} are easily obtained. S_{ij} is the set of all values of x for which $(x,x/n)$ is an element of $D_i \times C_j$, signifying that

$$S_{ij} = [nd_i, nd_{i+1}) \cap \{c_j, c_{j+1}-1\}. \qquad [26]$$

(The interval $[d_i, d_{i+1}]$ has been multiplied by n to give t and x the same scale.) Figure 1 clarifies Equation 26. All probability mass is concentrated on the points representing the linear relation $t = x/n$. The sets S_{ij}, containing all values of x for which $(x,x/n)$ is a point of cell (i,j), is exactly the intersection of the sets $[nd_i, nd_{i+1}]$ and $\{c_j, c_{j+1}\}$.

Assuming the beta-binominal model, which implies that $h(x)$ has the negative hypergeometric form and that the regression of t on x is linear (Lord & Novick, 1968, chap. 23), $h(x)$ and the probabilities P_{ij}, P_i, and P_j, which are needed for the computation of δ, can be estimated in the same way as for dichotomous decisions (Koppelaar, Van der Linden, & Mellenbergh, 1977). Although δ can be computed under these assumptions, the coefficient is not necessarily between 0 and 1. The reason is that it depends on the loss function in Equation 20 and on the bivariate distribution $k(t,x)$ whether or not the risk, R, is in the interval from R_c to R_n.

Van der Linden and Mellenbergh (1977) described a linear loss function for dichotomous decisions. A generalization of this function seems suitable for multiple decisions; the loss is a linear function of t for each cell of the decision table:

$$\begin{aligned}
L_{i0}(t) &= b_0(t - d_1) + a_0 && i = 0,1, \ldots\ldots\ldots k \\
L_{ij}(t) &= \begin{cases} b_j(d_j - t) + a_j & t \leq t_j \\ b'_j(t - d_{j+1}) + a'_j & t > t_j \end{cases} && \begin{array}{l} i = 0,1, \ldots\ldots\ldots k \\ j = 1,2, \ldots (k-1) \\ t_j \in (d_j, d_{j+i}) \end{array} \qquad [27] \\
L_{ik}(t) &= b_k(d_k - t) + a_k && i = 0,1, \ldots\ldots\ldots k
\end{aligned}$$

Example

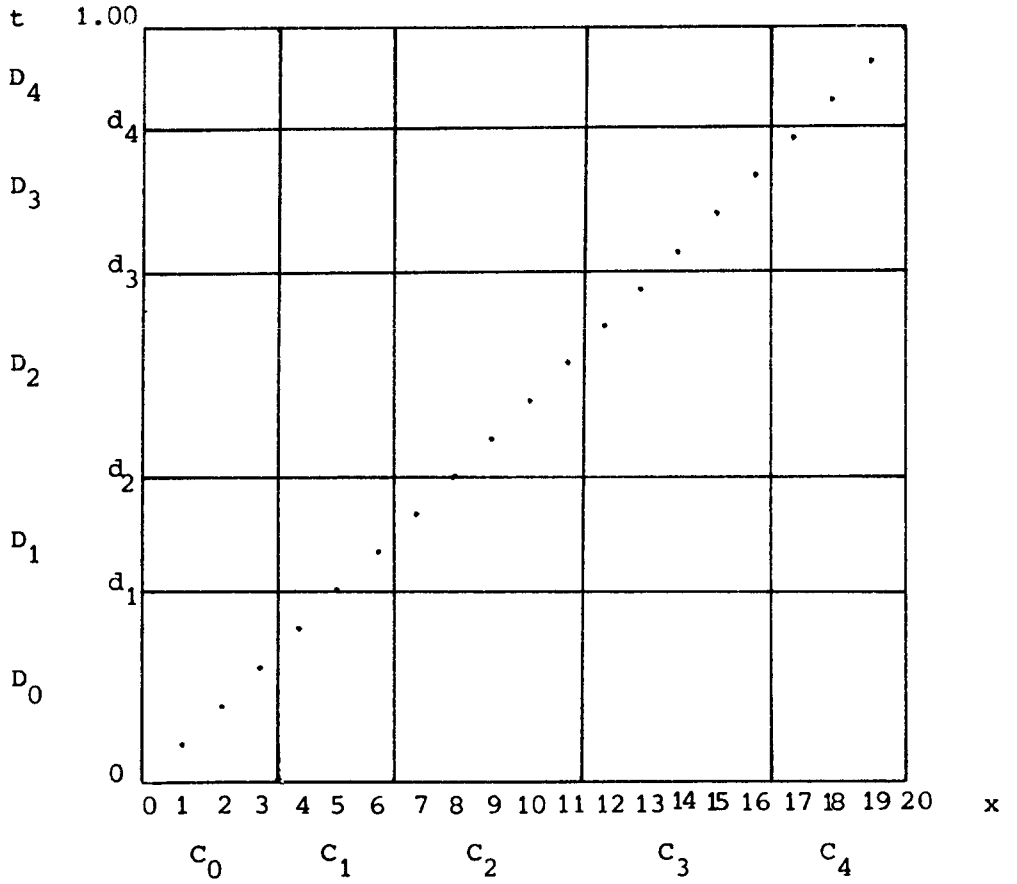
An example will clarify some of the statements made above. Suppose that in achievement measurement the true score interval is divided into five categories:

D_0: [0, 0,40], "very bad";

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 1

A Linear Functional Relation between True and Observed Score For Multiple Decisions



D_1: [0.40, 0.60], "insufficient";
D_2: [0.60, 0.75], "sufficient";
D_3: [0.75, 0.90], "good";
D_4: [0.90, 1.00], "excellent".

A twenty-five item four-choice test is administered. The observed scores are divided into five categories and grades are assigned as follows:

C_0: {0, 1, . . . 14} , grade E;
C_1: {15, 16, 17} , grade D;
C_2: {18, 19, 20} , grade C;
C_3: {21, 22, 23} , grade B;
C_4: {24, 25} , grade A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

A specification of the loss function in Equation 20 is reported in Table 1, as well as hypothetical proportions $P_{i.}$ and $P_{.j}$. Using Equation 22, R_n can be calculated for these hypothetical data. A specification of the loss function in Equation 27 is illustrated in Table 2.

Table 1
A specification of the Loss Function Equation 20
with Hypothetical Proportions $P_{i.}$ and $P_{.j}$

	Observed Score					
True Score	C_0	C_1	C_2	C_3	C_4	$P_{i.}$
D_0 .40	0	1	2	3	4	.10
D_1 .60	1	0	1	2	3	.20
D_2 .75	2	1	0	1	2	.40
D_3 .90	3	2	1	0	1	.20
D_4	4	3	2	1	0	.10
$P_{.j}$	.20	.10	.50	.15	.05	

In Table 2 the following specification is chosen for a_j : $a_2 = a_3 = a_4 = a'_2 = a'_3 = 0.02$ representing, for instance, the cost for testing, and $a_0 = a_1 = a'_1 = 0.10$ representing, for instance, the cost for both testing and remedial teaching. For b_i the specification is chosen: $b_0 = b_4 = 1$, $b_j = b'_j = 1$ $(j = 1,2,3)$.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2

A specification of the Loss Function Equation 27 with $a_2 = a_3 = a_4 = a'_2 = a'_3 = .02$, $a_0 = a_1 = a'_1 = .10$, $b_0 = b_4 = 1$, $b_j = b'_j = 1$ $(j=1,2,3)$

True Score		Observed Score C_0	C_1	C_2	C_3	C_4
D_0	.40	t-.30	.50-t	.62-t	.77-t	.92-t
D_1	.60	t-.30	.50-t / t-.50	.62-t	.77-t	.92-t
D_2	.75	t-.30	t-.50	.62-t / t-.73	.77-t	.92-t
D_3	.90	t-.30	t-.50	t-.73	.77-t / t-.88	.92-t
D_4		t-.30	t-.50	t-.73	t-.88	.92-t

Dichotomous Decisions

If $k = 1$, multiple decisions reduce to dichotomous decisions. The true score t is divided into two disjoint intervals: $D_0 = [0,d]$ and $D_1 = [d,1]$; and the observed score x is divided into the two sets $C_0 = \{0,1, \ldots c-1\}$ and $C_1 = \{c,c+1, \ldots n\}$ (see Table 3). Given a cutting score, c, that is optimal for a defined loss function, the best decision for an observed score from the set C_j is to assume that the true score is in the interval D_j.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3

Dichotomous Decisions with Proportions

True Score	Observed Score: $x < c$	Observed Score: $x \geq c$ (c)	$P_{i.}$
$t < d$ (d)	P_{00}	P_{01}	$P_{0.}$
$t \geq d$	P_{10}	P_{11}	$P_{1.}$
$P_{.j}$	$P_{.0}$	$P_{.1}$	

The loss function of Equation 20 has the following form for dichotomous decisions:

$$L_{ij}(t) = \begin{cases} \ell_{00} & \text{for} \quad t < d, \; x < c \\ \ell_{10} & \text{for} \quad t \geq d, \; x < c \\ \ell_{01} & \text{for} \quad t < d, \; x \geq c \\ \ell_{11} & \text{for} \quad t \geq d, \; x \geq c \end{cases} \qquad [28]$$

where $\ell_{00} \leqslant \ell_{01}$ and $\ell_{10} \geqslant \ell_{11}$. The coefficient δ for dichotomous decisions and this loss function results from Equations 28, 25, 24, 22, 21, and 1 and the specification $k = 1$. The restrictions, $\ell_{00} = \ell_{11} = 0$ and $\ell_{10} = \ell_{01} = \ell$, for the loss function in Equation 28, together with the assumption of an increasing regression function of t on x, give rise to an interesting result. The loss function represents the situation in which the loss is equal for both correct classifications, as well as for both misclassifications, and in which a rescaling has occurred in order to give both correct decisions a zero loss. The risk is

$$R = (P_{01} + P_{10})\, \ell. \qquad [29]$$

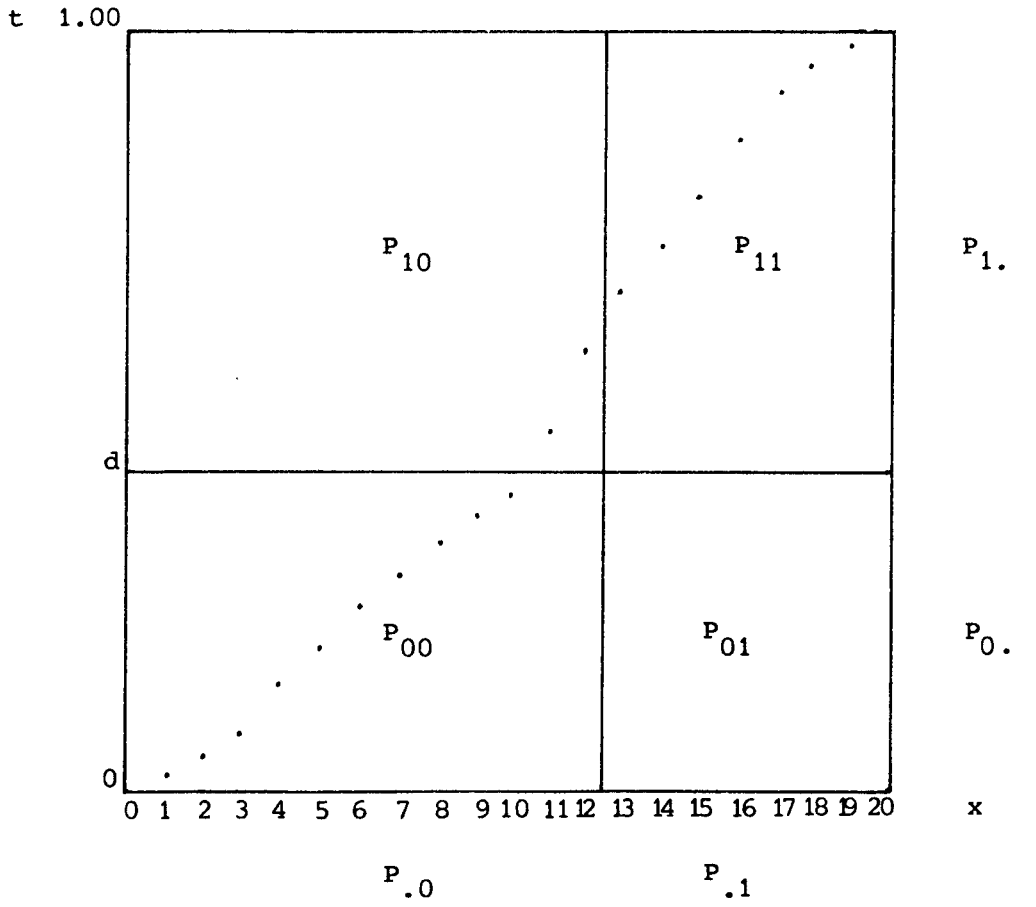
Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

In case of stochastic independence between t and x, the risk is

$$R_n = (P_{0.} P_{.1} + P_{1.} P_{.0}) \ell. \qquad [30]$$

For an increasing regression function of t on x and for functional dependence of t and x, the true score, t, is an increasing function of the observed score, x. It follows that P_{11} is the smallest of the probabilities P_1 and $P_{.1}$, and that either $P_{01} = 0$ (for $P_{.1} \leqslant P_1$) or $P_{10} = 0$ (for $P_{.1} > P_1$). For instance, when $P_1 \geqslant P_{.1}$, then $P_{11} = P_{.1}$ and $P_{01} = 0$; also $P_{10} = P_1 - P_{.1}$, and $P_{00} = P_0$. This is demonstrated in Figure 2; all probability mass is concentrated on the points representing the increasing functional relation between t and x. The probability P_{01} is equal to 0; the probability P_{11} is equal to the probability $P_{.1}$, the smallest of the probabilities P_1 and $P_{.1}$.
Therefore, R_c can be written as

Figure 2
An Increasing Functional Relation
Between True and Observed Score



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

$$R_c = \begin{cases} (P_{1.} - P_{.1})\,\ell & \text{for } P_{1.} \geq P_{.1} \\ (P_{.1} - P_{1.})\,\ell & \text{for } P_{1.} < P_{.1} \end{cases} \quad [31]$$

Substituting Equation 29, 30, and 31 into Equation 1 and simplifying yields

$$\delta = \begin{cases} (P_{11} - P_{1.}\,P_{.1})/\{P_{.1}(1 - P_{1.})\} & \text{for } P_{1.} \geq P_{.1} \\ (P_{11} - P_{1.}\,P_{.1})/\{P_{1.}(1 - P_{.1})\} & \text{for } P_{1.} < P_{.1} \end{cases} \quad [32]$$

which is the well-known coefficient H of Loevinger. The coefficient can be estimated if the proportions P_{ij} can be estimated. This can be done, assuming the aforementioned beta-binomial model. Of course, the same derivation applies for a linear regression function of t on x, because the linear regression function is an increasing function of x. In this case, however, the same result can be obtained from the general result using Equation 26.

For dichotomous decisions, the loss function in Equation 27 takes the following form:

$$L(t) = \begin{cases} b_0(t - d) + a_0 & \text{for } x < c \\ b_1(d - t) + a_1 & \text{for } x \geq c \end{cases} \quad [33]$$

Given this linear function, a simple formula for computing the optimal value of the cutting score, c, has been derived; the risk is, in this case (Van der Linden & Mellenbergh, 1977),

$$R = \sum_{x=0}^{c-1} [b_0\{E(t|(x/n)) - d\} + a_0]h(x) - \sum_{x=c}^{n} [b_1\{E(t|(x/n)) - d\} + a_1]h(x). \quad [34]$$

The assumption of the linear regression function of t on x (Equation 10) also gives rise to an interesting result. Substituting Equation 10 into Equation 34 gives the risk, R. In the situation in which t and x are stochastically independent, there is also linear stochastic independence; and Equation 10 has the form:

$$E_\ell(t|x) = E(x/n). \quad [35]$$

Substituting Equation 35 into Equation 34 gives the risk R_n for stochastic independence between t and x. It was shown earlier that for the situation in which t and x are functionally dependent, they are also linearly dependent; and Equation 10 has the form:

$$E_\ell(t|x) = x/n. \quad [36]$$

Substituting Equation 36 into Equation 34 gives the risk R_c for functional dependence between t and x. Since $h(x)$ was assumed to be fixed, the result of substituting R, R_n, and R_c into Equation 1 is simply

$$\delta = \rho_{xx'}. \quad [37]$$

Therefore, in the case of a linear regression function of t on x and the linear loss function in Equation 33, the well-known reliability coefficient is interpretable as a standardized risk. Its standardization is given by Equation 1.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Discussion

Attempts have recently been made to construct coefficients for dichotomous decision situations (Livingston, 1972; Harris, 1974). These attempts are not based on a theoretical foundation, such as the decision theoretic point of view presented here. The results of this development were rather surprising. Two old friends of psychometrics have shown up again: (1) Loevinger's coefficient H between the dichotomized true and observed score for the loss function in Equation 28 and (2) the reliability coefficient in the case of the loss function in Equation 33 and Kelley's formula for estimating the true score.

An obvious disadvantage of Equation 28 is that the loss is constant for each cell of the twofold table. For instance, a non-accepted subject with a true score just above the cutting scores gives the same loss as a not-accepted subject with a true score far above the cutting score. Therefore, in most applications, Equation 33 will be more realistic than Equation 28 for expressing the losses in dichotomous decision situations. There does not seem to be much need for a new coefficient specially adapted to dichotomous decision situations. In many practical applications, the reliability coefficient will be sufficient from a decision theoretic point of view.

Using loss function Equation 28, Huyhn (1976a) has described a coefficient ε for dichotomous decision situations. The coefficient is derived in a way that is consistent with our approach: it is also constructed from a decision theoretic point of view. The coefficient is defined as

$$\varepsilon = (R^* - R_0)/R^*. \qquad [38]$$

Comparing Equation 38 with Equation 1, it appears that Huyhn used R_0 instead of R, and R^* instead of R_n. For the situation in which an optimal cutting score is used R_0 is the risk; when an arbitrary cutting score is used, R is the risk.

In practical applications, a cutting score is often fixed before the test is administered; in general, this fixed cutting score is not optimal. Hence the use of R instead of R_0 seems to be more adapted to practical situations; furthermore, R is more general because R equals R_0 if the optimal cutting score is applied. Deriving R^*, Huyhn started with the independence of t and x; thus far his derivation is identical to R_n (Equation 22) with loss function Equation 28. In this case, Huyhn also looked for an optimal cutting score. He observed that the minimum risk accompanying a dichotomous decision, the loss function of Equation 28, and stochastic independence between t and x is attained for either $c = c_f$ (the cutting score for which all subjects fail) or $c = c_p$ (the cutting score for which all subjects pass). Therefore, R^* is defined as the minimum value of the risk for either $c = c_f$ or $c = c_p$. Hence R^* is the risk for the situation in which the true and observed scores are distributed independently, and in which all subjects either pass or fail the test.

R_n is derived for the situation in which the true and observed scores are distributed independently and in which, on the contrary, the chosen cutting score has been maintained. Similar to R, this seems to be better adapted to practical situations. Another difference between the coefficients ε and δ is that ε is only defined for dichotomous decisions with the loss function in Equation 28, whereas δ is defined for a broad class of decision situations and loss functions. Therefore, δ seems to be a more realistic and more general coefficient than ε.

In the literature on criterion-referenced measurement or mastery testing, the use of coefficient kappa (Cohen, 1960) for determining the reliability of decisions is becoming popular (Swaminathan, Hambleton, & Algina, 1974). Coefficient kappa is computed for data collected according to a test-retest or test-parallel test design and reflects the chance corrected consistency of decisions. Here, in-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

terest is not in determining consistency of decisions, but in deriving a coefficient that represents the optimality of decisions. Therefore, instead of a test-retest or test-parallel test design, the decision table, which has the latent and observed scores along its axes, is considered. For this table a coefficient like Cohen's kappa is not only less natural than a coefficient based on decision theoretic risk, but it is also erroneous.

Coefficient kappa is introduced by Cohen (1960) as a coefficient that expresses the degree of agreement between two nominal scales corrected for chance. Moreover, it should be noted that computing kappa supposes equal marginal distributions for both nominal scales, since kappa can not reach its maximal value in case of differing marginal distributions. For the decision table, however, equal marginal distributions are unnecessary: the used cutting scores and the unreliability of the observed score will give rise to differing marginal distributions.

Furthermore, determining the optimality of decisions is not a matter of agreement between nominal scales, but of association between the ordered categories along both axes of the decision table. Not only should the for-chance-corrected proportion of cases in the diagonal be considered, but also all combinations of the true and observed score should be taken into account, each weighted according to the loss function used. Though weighted coefficient kappa (Cohen, 1968) makes weighting procedures possible, it is inappropriate as well. With this coefficient, only weighting by loss functions of the type in Equation 18 is possible. Loss functions which are more plausible, such as Equations 27 and 33, cannot be used for this purpose. On the contrary, an attractive feature of coefficient δ is that it can be derived for decision procedures with loss functions of any type.

Finally, three points should be noted. First, δ is not always in the interval from 0 to 1. It has been shown that the coefficient is in this interval for some important applications. For example, at least coefficient H and the reliability coefficient are in the interval. Second, at several places the beta-binomial model is mentioned. The statistical properties of estimates computed under the assumption of this model are not known, however; and whether or not this model fits the data should be investigated. Third, there seems to be a lack of interest in multiple decision situations; emphasis has mainly been on point estimation and dichotomous decisions. In fact, many situations are multiple decision situations. For instance, subjects with scores equal or above a score c_1 on a test pass the test, subjects with scores below c_0 fail the test, and subjects with scores between C_0 and c_1 are retested or get special treatments, such as remedial teaching. Therefore, multiple decisions require attention.

References

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20,* 37–46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* 1968, *70,* 213–220.

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

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.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement.* Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Huyhn, H. *On mastery scores and efficiency of criterion-referenced tests when losses are partially known.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976. (a)

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

Huyhn, H. Reliability of decisions in domain-referenced testing. *Journal of Educational Measurement,* 1976, *13,* 253–264. (c)

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Koppelaar, H., Van der Linden, W. J., & Mellenbergh, G. J. A computer-program for classification proportions in dichotomous decisions based on dichotomously scored items. *Tijdschrift voor Onderwijs Research*, 1977, *2*, 32–36.

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

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

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*, 1978, *32*, in press.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: views regarding mastery and standard-setting. *Review of Educational Research*, 1976, *46*, 133–158.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263–268.

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

Acknowledgements

We thank Fred N. Kerlinger, Robert F. van Naerssen, and Pieter Vijn for their comments. The order of the names of the authors is alphabetical; they are equally responsible for the content.

Authors' Adresses

G. J. Mellenbergh, Psychologisch Laboratorium, Universiteit van Amsterdam, Weesperplein 8, Amsterdam, The Netherlands; Wim J. van der Linden, Twente University of Technology, Enschede, The Netherlands.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/