The Internal and External Optimality of Decisions Based on Tests

Gideon J. Mellenbergh
University of Amsterdam, Amsterdam, The Netherlands

Wim J. van der Linden
Twente University of Technology, Enschede, The Netherlands

In applied measurement, test scores are usually transformed to decisions. Analogous to classical test theory, the reliability of decisions has been defined as the consistency of decisions on a test and a retest or on two parallel tests. Coefficient kappa (Cohen, 1960) is used for assessing the consistency of decisions. This coefficient has been developed for assessing agreement between nominal scales. It is argued that the coefficient is not suited for assessing consistency of decisions. Moreover, it is argued that the concept consistency of decisions is not appropriate for assessing the quality of a decision procedure. It is proposed that the concept consistency of decisions be replaced by the concept optimality of the decision procedure. Two types of optimality are distinguished. The internal optimality is the risk of the decision procedure with respect to the true score the test is measuring. The external optimality is the risk of the decision procedure with respect to an external criterion. For assessing the optimality of a decision procedure, coefficient delta (van der Linden & Mellenbergh, 1978), which can be considered a standardization of the Bayes risk or expected loss, can be used. Two loss functions are dealt with: the threshold and the linear loss functions. Assuming psychometric theory, coefficient delta for internal optimality can be computed from empirical data for both the threshold and the linear loss functions. The computation of coefficient delta for external optimality needs no assumption of psychometric theory. For six tests coefficient delta as an index for internal optimality is computed for both loss functions; the results are compared with coefficient kappa for assessing the consistency of decisions with the same tests.

In applied psychological and educational measurement, test scores are used for making decisions. A very common decision problem is to classify subjects into two categories, 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. These problems will be called dichotomous-decision problems (Mellenbergh, Koppelaar, & van der Linden, 1977). Dichotomous-decision problems are a special case of multiple-decision problems: In multiple-decision problems, subjects are classified into more than two categories (Ferguson, 1967, p. 10). An example is the assignment of grades to the scores on an achievement test.

In classical test theory the reliability of a test is defined as the squared correlation coefficient between observed score and true score (Lord & Novick, 1968, p. 61). It can be shown that this squared correlation equals the correlation between parallel tests (Lord & Novick, 1968, p. 63). Therefore, an

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 2 Spring 1979 pp. 257–273

appropriate procedure for assessing the reliability of a test is correlating two parallel forms or two administrations of the test. In classical test theory the (predictive) validity of a test is defined as the correlation between the observed score and a criterion score. Consequently, the validity of a test can be assessed by the usual procedure for estimating the correlation between two variables. Note how the definitions of reliability and validity contain the concept of correlation; it follows that the reliability of a test is identical to its validity with respect to a parallel test.

For dichotomous decisions it has been suggested that the concept reliability of decisions, analogous to the reliability of test scores, be used and that the reliability of decisions be assessed by determining their consistency. Carver (1970) and Jackson (1970) suggested the consistency of decisions on parallel forms of a test, while Hambleton and Novick (1973) suggested that the consistency of decisions can be assessed on two administrations of the same test. According to the latter authors, the definition and determination of the validity of decisions would take a comparable form, with a new test providing a dichotomy against which the decisions would be validated. Swaminathan, Hambleton, and Algina (1974) proposed the use of coefficient kappa (Cohen, 1960) for assessing the consistency of decisions. In this paper the above reasoning is questioned.

In the first section the use of coefficient kappa for assessing decision consistency is investigated. Both the dual administration procedure of Swaminathan, Hambleton, and Algina (1974) and the single administration procedures of Huynh (1976b) and Subkoviak (1976, undated) are presented and reviewed. Coefficient kappa was introduced by Cohen (1960) as a coefficient of agreement between nominal scales, especially suited to assessing agreement between clinical judgments. Assessing the consistency of decisions is, however, a different and less appropriate application of kappa.

In the second section the question is raised whether the concept of consistency of decisions is an admissible substitute for decision reliability. In classical test theory reliability is defined as the squared correlation between observed score and true score and theoretically equals the consistency or correlation between parallel measurements. But decision consistency is not related in an analogous way to the association between categories to which persons are assigned and their true states. In a fully decision-oriented approach to test theory, optimality and not consistency of decisions should be considered. Decisions can be optimal with respect to categories on the true score that the test measures or with respect to categories on a criterion that the test predicts. Therefore, the concepts internal and external optimality are introduced as the decision theoretic counterparts of reliability and (predictive) validity in classical test theory. For deriving optimal decision procedures, it is necessary to use a loss function; and in order to assess the optimality of the resulting decision procedure, a coefficient is needed that is based on the loss function that is actually used. Coefficient delta (van der Linden & Mellenbergh, 1978) has this property. The coefficient is defined for a broad class of decision situations and loss functions; results for some special cases are known. In the remainder of the paper the decision theoretic concepts of loss and risk, the coefficient delta, and some applications of delta to the assessment of internal and external optimality are considered.

Kappa for Assessing Consistency

Following the suggestion made by Hambleton and Novick (1973), Swaminathan, Hambleton, and Algina (1974) defined reliability of a decision procedure as the measure of agreement between the decisions in repeated test administrations. Instead of choosing the known coefficient of agreement

$$\sum_{i=0}^{k} p_{ii} \,, \qquad [1]$$

where p_{ii} is the proportion of persons placed in the i^{th} category on both test administrations and k is the number of categories, Swaminathan et al. (1974) recommended coefficient kappa as a measure of agreement.

Coefficient kappa, introduced by Cohen (1960), is defined as

$$\kappa = (p_0 - p_c)/(1 - p_c) \ , \qquad [2]$$

where p_o and p_c are, respectively, the proportion of agreement and the chance-expected proportion of agreement defined by

$$p_0 = \sum_{i=0}^{k} p_{ii} \qquad [3]$$

and

$$p_c = \sum_{i=0}^{k} p_{i.} p_{.i} \ , \qquad [4]$$

where $p_{i.}$ and $p_{.i}$ are the marginal proportions. In the case of dichotomous decisions both proportions can be computed from a test-retest or parallel test design as displayed in Table 1a. Since in the numerator and denominator of Equation 2, the chance-expected proportion of agreement, p_c, is subtracted, kappa can be interpreted as the proportion of agreement between decisions after chance agreement is removed. Discussions of the properties of kappa, comparisons with other measures of agreement, extensions, sampling distributions, and tests for statistical significance can be found in

Table 1
Parallel (Re)Test Design (a) And Decision Table (b) For Dichotomous Test Decisions

(a)

Observed Test Score

Parallel (Re)Test	$X < c$	$X \geq c$	$p_{i.}$
	c		
$X' < c$	p_{00}	p_{01}	$p_{0.}$
$X' \geq c$	p_{10}	p_{11}	$p_{1.}$
$p_{.j}$	$p_{.0}$	$p_{.1}$	

(b)

True State	$X < c$	$X \geq c$	$p_{i.}$
	c		
$Z < d$	p_{00}	p_{01}	$p_{0.}$
$Z \geq d$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	

Bishop, Fienberg, and Holland (1976, chap. 11); Cohen (1960, 1968, 1972); Everitt (1968); Fleiss (1971); Fleiss and Cohen (1973); Fleiss, Cohen, and Everitt (1969); Hubert (1977); Landis and Koch (1975a, 1975b); Light (1971); and Spitzer, Cohen, Fleiss, and Endicott (1967).

Of special interest is weighted kappa, an extension of coefficient kappa introduced by Cohen (1968) to meet the requirements of those situations in which some off-diagonal cells in the $k \times k$ matrix are of greater importance than others. With this version of kappa, the availability of disagreement weights v_{ij} for the k^2 cells of the $k \times k$ table is assumed; weighted kappa can then be obtained from Equation 2 by replacing p_o and p_c by p_o' and p_c', respectively, from the proportions of weighted disagreement

$$1 - p_o' = \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} p_{ij} \qquad [5]$$

and

$$1 - p_c' = \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} p_{i.} p_{.j} \qquad [6]$$

Like kappa, coefficient weighted kappa is chance corrected: It can be interpreted as the proportion of weighted agreement after weighted chance agreement is removed. It should be noted that weighted kappa reduces to kappa when same weights are given to all diagonal cells and zero weights to all off-diagonal cells of the $k \times k$ matrix. Thus kappa is a special case of weighted kappa, but unlike weighted kappa, it is unaffected by discrepancies that exist between off-diagonal entries in the $k \times k$ table.

One of the disadvantages of coefficients based on test-retest designs is that because of practice, memory, or forgetting effects, it is hardly possible to obtain retest data that completely fulfill the requirements of a statistical replication. This does not apply to parallel test designs in which, however, the construction of a second test that meets these requirements is often a difficult task. For these reasons a single administration coefficient is desirable. Huynh (1976b) proposed such a coefficient of decision consistency across repeated test administrations on the basis of a single test administration. Using the beta-binomial model and assuming local independence between two parallel tests X and X', both of length n, the bivariate distribution $m(X,X')$ can be derived as a bivariate negative hypergeometric density with parameters α and β (Keats & Lord, 1962):

$$m(X,X') = \binom{n}{X}\binom{n}{X'} B(\alpha + X + X', \beta + 2n - X - X')/B(\alpha,\beta) \qquad [7]$$

Using estimates from a single test administration for the parameters α and β, Huynh (1976b) proposed to estimate the proportions in the cells of Table 1a with Equation 7 and, after that, to apply the usual procedure for computing coefficient kappa. From a slightly different angle Subkoviak (1976, undated) proposed a method like Huynh's (1976b) for estimating coefficient kappa from a single test administration. Assuming that X and X', given true score T, follow the binomial density, and assuming local independence with respect to true score T, the proportion of agreement and the chance-expected proportions of agreement can be written as

$$p_o'' = \sum_{j=1}^{N} \{[p(X_j \geq c)]^2 + [1 - p(X_j \geq c)]^2\}/N , \qquad [8]$$

and

$$p_c'' = [p(X \geq c)]^2 + [1 - p(X \geq c)]^2, \quad [9]$$

respectively, where

$$p(X_j \geq c) = \sum_{X_j=c}^{n} \binom{n}{X_j} T_j^{X_j}(1 - T_j)^{n-X_j}, \quad [10]$$

$$p(X \geq c) = \sum_{j=1}^{N} p(X_j \geq c)/N, \quad [11]$$

c is the cutting score on the test, and N the number of persons (Subkoviak, 1976, undated). After estimating T_j for all persons by the linear regression function,

$$\hat{T}_j = \alpha_{21} X_j/n + (1 - \alpha_{21})\bar{X}/n . \quad [12]$$

p_o'' and p_c'' can be computed and substituted in Equation 2 to obtain a single administration estimate of coefficient kappa (Subkoviak, undated). It is interesting to note that the estimator of Equation 12, which is recommended by Subkoviak for unimodal distributions of X, together with the assumption of X given T, is equivalent to assuming the beta-binomial model as has been done by Huynh (Lord & Novick, 1968, chap. 23). But though the method of Huynh and Subkoviak draws upon the same model, there is a difference. In Subkoviak's method the proportions of agreement in Equations 8 and 9 are weighted averages over the sample of N persons, with the sample frequency of X on which Estimator 12 of the binomial parameter T is based as weights. In Huynh's method, however, the proportions of agreement follow from Equation 7; and in deriving this equation a weighting with the density of T is involved. Because of the unreliability of test scores X, a weighting with the density of T and, after that, estimating the parameters in Equation 7 should be preferred to a weighting with the frequency of the estimates in Equation 12. For this reason Huynh's method is better suited than Subkoviak's method to a single test administration estimation of kappa.

For the relation of kappa to test score reliability, cutting score and test length, analyses of simulated data, and a generalization to multiple decisions the reader is referred to Huynh (1976b, 1976c, 1977), Algina and Noe (1978), and Subkoviak (1976, 1978, undated).

Although the formal aspects of the studies dealt with above are interesting, the meaning of coefficient kappa for determining decision consistency is to be criticized. Coefficient kappa was introduced by Cohen (1960) as a coefficient of *agreement* between nominal scales. An area of application mentioned by Cohen is that of determining agreement between clinical judgments, which are nearly always nominal. In that area many investigators of interjudge agreement were accustomed to compute the contingency coefficient C over the table with interjudge data as a measure of agreement or the Pearson χ^2 for use as a test of the statistical hypothesis of no agreement. Both statistics are, however, improper, since they point to *association* and not to agreement.

In order to assess the meaning of coefficient kappa regarding decision consistency, which is an application totally different from interjudge agreement, it is important to note the difference between

association and agreement. Generally, two responses or measurements agree when they fall into the same category, whereas they are associated when they are not distributed independently. From this distinction it follows that for agreement the main diagonal of the table has a special meaning: Each response to be classified into the main diagonal indicates agreement. For association, attention is not focused on the main diagonal but on the conditional distributions in the rows or columns of the table; differing conditional distributions indicate predictability of the one category from the other and, therefore, association. Agreement can be considered as a special case of association; it exists to the extent to which each of the conditional distributions has frequency in its own diagonal cell. The distinction between agreement and association can be stated from the point of view of weighting as well. With association all frequencies in the table are considered as unweighted or, equivalently, as having equal weights. The definition of agreement implies, however, a 0–1 weighting: All frequencies in the main diagonal cells are weighted by 1 and in the off-diagonal cells by 0. Therefore, agreement is that special case of association wherein the frequencies of all off-diagonal cells are weighted by zero. The 0–1 weighting procedure implied by the definition of agreement has a close relation to the nominal character of the categories. Nominal categories are the same or are not the same; there is no ordering or metric from which it may be derived that some pairs of categories are more alike than others. For this reason there is nothing wrong in stating that agreement is that case of association that is specially suited to nominal categories or in defining agreement as association between nominal scales. It should be clear that Cohen's (1960) proposal to use the coefficient of agreement kappa for determining the reliability of nominal interjudgment data is mainly a defense of the peculiar character that association between nominal scales has.

From the above paragraph it follows that coefficient kappa is less appropriate for determining the consistency of decisions based on tests. In the case of decisions the categories of the test-retest or parallel test table are not nominal but ordered with respect to each other. There is always an underlying variable or external criterion—for instance, mastery regions in criterion-referenced testing or suitability categories in selecting applicants for jobs—which can be demarcated and by which an ordering of the decisions is given. For the generalization of dichotomous to multiple decisions, it therefore means that a difference of one category points to more consistency than a difference of two categories, and the simple 0–1 weighting implied by coefficient kappa is far from appropriate. At first glance this conclusion can be explained as a plea for using coefficient weighted kappa instead of coefficient kappa. It is unclear, however, what weights should be used: the distance from the cells to the diagonal? or the square of this distance? Should all diagonal cells be weighted equally? or are consistent decisions of one kind more important than consistent decisions of another kind? No procedure seems available which provides a rational way to specify these weights. The reason is that evaluating the quality of decision outcomes can only be based on a comparison between decisions and true states (Table 1b) and not on the comparison between decisions across two test administrations (Table 1a). The loss function used in decision theory is based on the former, and in the next section it will be seen that this loss function should be chosen for evaluating the quality of decision outcomes. In this connection "consistency" has to be replaced by "optimality" as the decision theoretic counterpart of reliability and validity from classical test theory.

In addition to a loss function that is unsuited to the use of kappa as a coefficient for decision consistency, kappa has another flaw. It can only reach its maximum value of +1.00 when the two marginal distributions are identical. A low kappa value calculated from Table 1a can, therefore, have two sources: (1) the two test administrations give rise to decisions that are not consistent, and (2) the two test administrations do not meet the requirement of statistical parallelism. The latter merely points to the fact that one did not succeed in constructing a parallel test form or readministering the test under

fully identical circumstances, and this may not be confounded with the quality of the decisions based on one of these test administrations. A possible solution to this problem is the use of κ/κ_M instead of κ, where κ_M is the maximum value of coefficient kappa permitted by the marginals (Cohen, 1960, p. 42) or by Huynh's (1976b) single administration method for estimating kappa. Note that this problem, just as the problem of the improper loss function, does not exist for the area of application for which Cohen introduced kappa. In assessing interjudge agreement, differing marginal distributions point to differences in the frequencies with which judges classify persons into categories; this is, of course, disagreement and should result in a lower value for the coefficient.

Optimality versus Consistency

The outcome of the previous section is that coefficient kappa is not suited to assessing the consistency of decisions. A more fundamental question, however, is whether the concept consistency of decisions is adequate from a decision theoretic point of view. In classical test theory the reliability of a test is defined as the squared correlation between observed and true scores: ϱ^2_{XT} (Lord & Novick, 1968, p. 61). One of the first theorems that can be derived from the classical test model is that for parallel measurements X and X' (Lord & Novick, 1968, p. 63),

$$\rho^2_{XT} = \rho_{XX'} \qquad [13]$$

An appropriate procedure for assessing the reliability of measurements is, therefore, correlating parallel measurements X and X' and substituting the estimated correlation coefficient for the reliability coefficient ϱ^2_{xT}. Practical methods for obtaining estimated correlations between parallel measurements are the well-known test-retest, parallel test, and internal consistency methods. These methods are based on Equation 13 and their results are only valid when this equation holds. Note that in each of these three methods consistency of measurements across test administrations, across parallel forms, or across test parts plays a role and that Equation 13 permits the identification of reliability and consistency of measurements.

In criterion-referenced measurement and mastery testing, the conceptualization of reliability and consistency of decisions, as in the classical theory of reliability, has been proposed. This proposal is implicit in Carver's (1970) and Jackson's (1970) suggestion to assess the reliability of decisions by determining the consistency of decisions on parallel forms of a test. Hambleton and Novick (1973) suggested that this be done by determining the consistency of decisions on two administrations of the same test. Marshall and Haertel's (1975) coefficient beta for the consistency across decisions based on all possible test halves is a similar idea. Note how each of these three proposals resembles one of the classical methods for determining measurement reliability; the only thing that differs is that decisions are substituted for measurements.

The idea that classical test theory can be used for decisions by simply substituting dichotomies that result from cutting scores on the true and observed scores variable for the original metric is, however, incorrect. In classical test theory the theorem of Equation 13 is valid, but an analogous theorem does not hold for dichotomous decisions. The consistency of decisions is not related in the same way to the association between decisions and true states as consistency of measurements is related to the reliability coefficient. This can be shown rather trivially by the following counterexample. Suppose that the cutting score on the true score variable is somewhere on the middle of the scale, so that for a given population about the half the number of persons is suitable and the other half is not suitable. Also suppose that the cutting score on the test has, for some reason, the minimum possible value,

which means that all persons are accepted. This decision procedure is far from "reliable," because for about half the number of persons a decision is made which does not agree with their true state. Nevertheless, the consistency of the decisions will be perfect in the sense that the procedure will yield the same decisions for all persons at a second occasion. Other examples making clear that for decisions all possible combinations of consistency and "reliability" can occur can easily be constructed. *The reason is that consistency of decisions only depends on the cutting score on the test, whereas "reliability" of decisions depends on both the cutting score on the test and the true score variable.*

It can be shown that decision consistency may be interpreted as reliability but that this only holds under the condition that reliability with respect to a transformation of the original true score variable is meant and not reliability with respect to the true states (suitable and not suitable). Suppose that the decisions rejected and accepted are coded as 0 and 1 and that $f(X|T)$ denotes the conditional density of X given T. Using the classical test theory true score definition, which says that the true score for a fixed person is equal to the expected observed score, the true score T′ pertaining to the observed dichotomy {0,1} can be defined:

$$T' = 0.\sum_{X=0}^{c-1} f(X|T) + 1.\sum_{X=c}^{n} f(X|T) = \sum_{X=c}^{n} f(X|T) . \qquad [14]$$

Defining this true score, classical test theory applies to the observed dichotomy {0,1} and T′, and within this context it is correct to posit that the consistency of dichotomous decisions reflects their reliability with respect to T′. But, obviously, we are not interested in reliability with respect to T′, which for the conditional densities $f(X|T)$ mostly encountered in psychometrics is merely an increasing monotonic transformation of T, but in the reliability of the decisions rejected vs. accepted with respect to the true states suitable vs. not suitable.

In a decision theoretic approach to test theory the entire decision situation should be considered, including the cutting scores on the test and the variable with respect to which decisions are made. The situation is displayed in Table 1b. For the sake of generality a general variable Z is used to denote the variable on which the intervals suitable and not suitable are demarcated. In most applications Z will be either the underlying true score variable T or an external criterion Y with respect to which the decisions are to be made. Note that Table 1b can be derived from the bivariate density $k(X, Z)$ and differs from Table 1a used in the methods that identify decision consistency with reliability. Table 1b is the decision table that plays a central role in determining optimal cutting scores on the test for a given bivariate density $k(X, Z)$ and cutting score d on Z (Hambleton & Novick, 1973; Mellenbergh, Koppelaar, & van der Linden, 1977; Subkoviak & Wilcox, 1978; van der Linden & Mellenbergh, 1977). In addition to the decision table a loss function, which specifies the "costs" of all possible decision outcomes for the decision maker, is required for a fully decision theoretic approach.

The assessment of the quality of the decision procedure should be based on this decision table and loss function. In decision theory the quality of decisions is frequently measured by the optimality of the expected loss or Bayes risk. This measure is based both on information from the decision table and the loss function defined in this table. According to the authors of this paper, the term reliability should be deleted for decisions and replaced by *internal optimality*. Reliability is a concept with a definite meaning within the framework of classical test theory and should only be used for measurements. Internal optimality refers to the expected loss or risk incurred in making decisions with respect to categories of the true score variable T underlying the test. Note how both concepts are each other's counterparts: The former refers to the relation between true and observed scores and is based on the squared error loss function, whereas the latter refers to the relation between categories on the

true and observed score variables and is based on a loss function defined on the relation between these categories. The *external optimality* of a decision procedure could also be considered as the decision theoretic counterpart of (predictive) validity from classical test theory. In that case the variable Z in Table 1b is not the true score T, but an independent criterion Y, which the test is supposed to be predicting or correlating with. The external optimality of a decision procedure indicates the expected loss or risk incurred in making decisions with respect to categories of this independent criterion Y. Again, it is not based on the squared error loss function, as the validity coefficient is, but makes use of a loss function defined on the relation between the categories on the external criterion and on the test.

Loss, Risk, and Coefficient Delta

For evaluating decisions it is necessary to use a loss function which specifies the total costs or value of all possible decision outcomes. These costs may concern all relevant psychological, social, and economic consequences which the decision brings along. The general form of a loss function is $L(X,Z)$; the loss resulting from a decision is both a function of the observed score X on the basis of which the decision is made and the position on Z of the person for which the decision is made. For a given value of the variable Z, 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 Z (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 $k(X,Z)$ of the random variables X and Z in a given population of subjects:

$$R = E\,L(X,Z) = \sum_{X=0}^{n} \int_{-\infty}^{+\infty} L(X,Z)\,k(X,Z)\,dZ \quad . \qquad [15]$$

The variable Z is considered a continuous variable bounded by its relation to X. However, it may be considered a discrete variable: Substituting the summation sign for the integral sign in Equation 15 will not alter the derivations given below. In the remainder of this paper the integral sign will be used, and the Bayes risk will shortly be called the risk.

Hambleton and Novick (1973), Gross and Su (1975), Huynh (1976a), Petersen (1976), and Mellenbergh, Koppelaar, and van der Linden (1977) used threshold loss functions. For a cutting score c on the observed score X and a cutting score d on the variable Z, the threshold loss function is

$$L(Z) = \begin{cases} \ell_{00} & \text{for } X < c,\ Z < d \\ \ell_{10} & \text{for } X < c,\ Z \geq d \\ \ell_{01} & \text{for } X \geq c,\ Z < d \\ \ell_{11} & \text{for } X \geq c,\ Z \geq d \end{cases} \qquad [16]$$

Petersen (1976) has pointed out that the threshold loss function is rather unrealistic. For example, in the threshold loss function it is supposed that for all accepted, not suitable subjects the amount of loss is equal. It will be more realistic to suppose that for accepted, not suitable subjects the loss is a mon-

otonically increasing function of the variable Z. Van der Linden and Mellenbergh (1977) used a linear loss function:

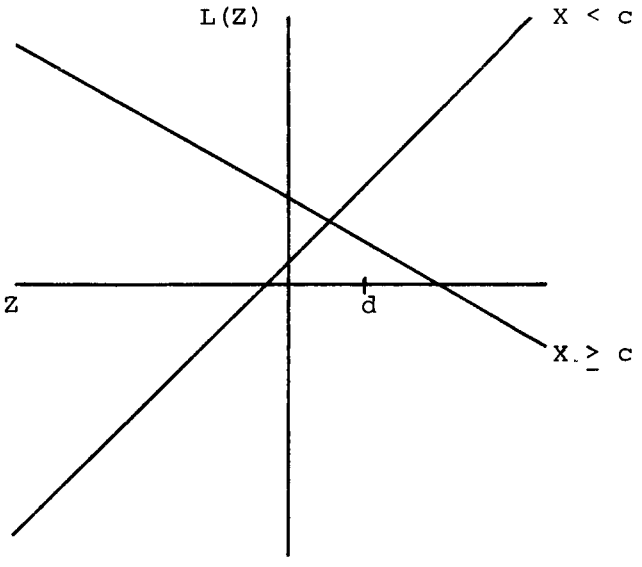
$$L(Z) = \begin{cases} b_0(Z - d) + a_0 & \text{for } X < c \\ b_1(d - Z) + a_1 & \text{for } X \geq c \end{cases} \quad [17]$$

An example of this function is given in Figure 1. The parameter a_0 is a constant for all rejected subjects; it can, for example, represent the costs for testing. The parameter a_1 is a constant for all accepted subjects; it can, for example, represent the costs for testing and for an educational program. Both $b_0(Z-d)$ and $b_1(d-Z)$ represent amounts of loss dependent on the score on the variable Z. These terms are proportional to the difference between the score on Z and the cutting score on Z; the parameters b_0 and b_1 are the constants of proportionality. The value of a_0 and a_1 should be chosen relative to each other and to $b_0(Z-d)$ and $b_1(d-Z)$, in such a manner that the resulting loss function represents as adequately as possible the total costs of all relevant consequences of the decisions. In general, the need to select values for the parameters of a loss function limits the practical utility of the decision theoretic approach. The linear loss function given here has, however, the advantage that under the condition $a_0 = a_1$ the parameter values may remain unspecified for some situations (van der Linden & Mellenbergh, 1977).

In decision theory the risk may be taken as an index of the quality of the decision procedure: The smaller the risk, the better the decision rule. Therefore, the risk is a proper starting-point for deriving decision-oriented coefficients for tests. Two disadvantages of the risk have to be removed. First, in test theory it is more conventional to define indices in which the scale has a direction opposite to that

Figure 1

An Example of the Linear Loss Function, $a_0 \neq a_1$, $b_0 \neq b_1$



in which the risk is represented. Second, the range of possible values for the risk can be different from the standard interval (0,1), whereas in test theory indices are nearly always defined on this interval. For removing both disadvantages van der Linden and Mellenbergh (1978) introduced coefficient delta:

$$\delta = (R_n - R)/(R_n - R_c) \qquad [18]$$

$R_n \neq R_c$, where R_n and R_c are hypothetical reference points with a decision theoretic interpretation. R_n is the risk in the situation in which the given test score distribution $h(x)$ would have no information about the true scores. This notion is formalized by the statement that X and Z are stochastically independent:

$$k(X,Z) = h(X)\,g(Z) \quad . \qquad [19]$$

where $h(X)$ and $g(Z)$ are, respectively, the densities of X and Z. R_c can be interpreted as the risk in the situation in which the given test score distribution $h(X)$ would have complete information about the true scores. Formally, this means that X and Z are functionally dependent, with Z as an increasing function of X. It should be noted that the conditions of stochastic independence and functional dependence are hypothetical borderline cases for a fixed distribution $h(X)$ and cutting score c. No empirical assumptions are involved; $h(X)$ and c are considered fixed merely as a consequence of considering the same test, test administration, and cutting score under different, nonexisting conditions.

Coefficient delta is defined over the correct table, contains the proper loss function, and represents the association between decisions and true states. It is a standardization of the decision theoretic risk which is used as a criterion in optimizing the decision procedure. For this reason coefficient delta should be interpreted as a coefficient for the optimality of a decision procedure: The closer delta is to 1, the closer the risk incurred by the decision procedure is to its minimal value. In order to obtain an optimal decision procedure and, thereby, a high value for coefficient delta, two conditions should be fulfilled. First, the observed variable X should contain as complete as possible information about the variable Z. This can be attained by using a test that yields a bivariate distribution of (X,Z) with minimal scatter around its regression function $E(Z|X)$. Second, the cutting scores used on the test should have optimal values. These values are to be found as the values that minimize the risk in Equation 15. Dependent upon whether T or Y is substituted for Z, delta is a coefficient for the internal or external optimality of the decision procedure, and different practical methods should be used to fulfill both conditions as well as possible.

Assessing Internal and External Optimality

Equation 18 is the general form of coefficient delta. Special cases for dichotomous decisions have been described by van der Linden and Mellenbergh (1978). From their derivations it immediately follows that the threshold loss function of Equation 16, with the restrictions $\ell_{10} = \ell_{01} = \ell$ and $\ell_{00} = \ell_{11} = 0$ and with the assumption of an increasing regression function of Z on X, yields a form of coefficient delta which is identical to Loevinger's coefficient H (Loevinger, 1947) between the dichotomized variables X and Z. When the test is used for the placement of persons into categories of the true score variable T underlying the test, coefficient delta takes the form of Loevinger's H computed over the proportions p_{ij} in Table 1b with T instead of Z. Assuming that the beta-binomial model holds, these proportions can be estimated by using a computer program described by Koppelaar, van der Linden, and Mellenbergh (1977). In this computer program the method of the moments is used to estimate the

beta parameters; for a discussion of maximum likelihood estimators the reader is referred to Griffiths (1973).

Instead of the beta-binomial model, an empirical Bayes approach such as Method 20 of Lord (1969) can be used; assuming a compound binomial model of X, given T and a smooth distribution of X, this method provides the bivariate distribution of X and T from which the proportions of Table 1b can be derived. For a computer program the reader is referred to Wingersky, Lees, Lennon, and Lord (1969). When the test is used for placement of persons into categories of an external criterion Y, the proportions p_{ij} have to be estimated from Table 1b, with Y instead of Z. An empirical bivariate frequency distribution may be used to obtain estimates of these proportions. It should be noted that the assumption of an increasing regression function of Y on X in this case can be checked by inspecting this frequency distribution.

For the linear loss function of Equation 17 the risk is:

$$R = E\,L(X,Z) = \sum_{X=0}^{c-1} \int_{-\infty}^{+\infty} \{b_0(Z - d) + a_0\}\, k(X,Z)\,dZ + \qquad [20]$$

$$\sum_{X=c}^{n} \int_{-\infty}^{+\infty} \{b_1(d - Z) + a_1\}\, k(X,Z)\,dZ$$

(van der Linden & Mellenbergh, 1977). Using $k(X,Z) = p(Z|X)h(X)$, $\int_{-\infty}^{+\infty} p(Z|X)\,dZ = 1$, and $\int_{-\infty}^{+\infty} Z p(Z|X)\,dZ = E(Z|X)$, where $p(Z|X)$ and $E(Z|X)$ are the distribution and the expected value of Z given X, it follows that

$$R = \sum_{X=0}^{c-1} [b_0\{E(Z|X) - d\} + a_0]\, h(x) - \qquad [21]$$

$$\sum_{X=c}^{n} [b_1\{E(Z|X) - d\} + a_1]\, h(x) \;.$$

Considering decisions with respect to T and assuming the regression function $E(Z|X)$ to be Kelley's linear regression function, (Lord & Novick, 1968, p. 65),

$$E_\ell(T|X) = \rho_{XX'}\, X + (1 - \rho_{XX'})\, E(X) \qquad [22]$$

coefficient delta for the internal optimality appears to be equal to the classical reliability coefficient $\rho_{xx'}$. This can be shown by noting that in case of no and complete information about T, Kelley's regression line equals

$$E_\ell(T|X) = E(X) \qquad [23]$$

and

$$E_\ell(T|X) = X \;, \qquad [24]$$

respectively, substituting Equations 22, 23, and 24 into Equation 21 and substituting the results in turn into the definition of coefficient delta, Equation 18. In this case coefficient delta can be estimated according to the theory of estimating the classical reliability coefficient (Lord & Novick, 1968, chap. 9).

Considering decisions with respect to an external criterion Y and assuming the regression function $E(Z|X)$ to be the regression line,

$$E_\ell(Y|X) = E(Y) + \{X - E(X)\}\ \rho_{XY}\ \sigma_Y/\sigma_X \qquad [25]$$

coefficient delta for the external optimality appears to be equal to the classical validity coefficient ϱ_{XY}. Noting that in case of no and complete information from X about Y the regression line will be equal to

$$E_\ell(Y|X) = E(Y) \qquad [26]$$

and

$$E_\ell(Y|X) = E(Y) + \{X - E(X)\}\ \sigma_Y/\sigma_X\ . \qquad [27]$$

respectively, this can be shown by substituting in the same way as before. In this case delta can be estimated according to the normal procedure of estimating a correlation coefficient.

Finally, it should be realized that the equality of coefficient delta to the reliability and validity coefficient relies heavily upon the assumption of linearity for $E(T|X)$ and $E(Y|X)$. Generally, the regression of Z on X will be curvilinear rather than linear, and delta will differ from the reliability or validity coefficient to some extent. The seriousness of this departure may, however, be checked by inspecting the estimated bivariate frequency distributions of (T, X) and (Y, X) and assessing the departures from linearity shown by the estimated conditional means. The first distribution can be obtained by using the aforementioned Method 20 of Lord, the latter by simply collecting pairs of observations (Y, X) from a random sample of persons.

Numerical Example

The coefficients reviewed in this paper were evaluated using data from end-of-unit criterion-referenced tests in a mastery learning course on elementary statistics. The teachers of this course considered a student as having mastered the contents of a unit of instruction if he/she could respond correctly to at least 80% of the total domain of multiple-choice items. Therefore, the cutting score d on the true score variable of these tests was fixed at .80. De Bruyne (1976, p. 97) tested the beta-binomial model for eight tests used in the course. For six of these tests the right tail probabilities of the chi-square values indicate that the fit of the model to the data was acceptable; classification proportions were computed using the computer program described by Koppelaar, van der Linden, and Mellenbergh (1977).

The results are given in the lines of Table 2 denoted by $\hat{p}_{00}$, $\hat{p}_{01}$, $\hat{p}_{10}$, and $\hat{p}_{11}$. From these proportions Loevinger's H was estimated, to which coefficient delta is identical in case of the loss function of Equation 16 under the restrictions $\ell_{10} = \ell_{01} = 1$ and $\ell_{00} = \ell_{11} = 0$. When the linear loss function of Equation 17 is used, coefficient delta is identical to the reliability coefficient; the former can then be estimated using an estimate of the latter. For this purpose estimates of the Kuder-Richardson Formula 20 were used; the estimates were obtained by substituting statistics computed from the sample for the parameters in the definition of KR20. KR20 is a lower bound on the reliability coefficient, and it may thus be expected that the figures in Table 2 give a value for coefficient delta which is too low.

The estimated values of coefficient kappa in the last line of Table 2 were computed with Huynh's (1976b) method. As the computational work is rather tedious for tests of the given lengths, the normal approximation for kappa was used, together with a linear interpolation of the tables for the univariate and bivariate standard normal distributions given by Gupta (1963). The most conspicuous aspect of

Table 2
Goodness of Fit, Classification Proportions, Coefficient Delta, and Coefficient Kappa for Six Multiple Choice Tests Elementary Statistics

Test	A	B	C	D	E	F
Number of students	184	127	106	163	147	150
Number of items	19	18	20	20	19	20
Cutting score c	15	14	16	16	15	16
Chi-square	4.00	4.474	2.832	7.888	4.942	10.640
Degrees of freedom	7	6	7	10	8	7
Right tail probability	.77	.61	.90	.64	.76	.16
$\hat{p}_{00}$	.236	.410	.526	.914	.607	.418
$\hat{p}_{01}$	.156	.217	.170	.067	.202	.138
$\hat{p}_{10}$	.070	.043	.052	.005	.033	.057
$\hat{p}_{11}$	.573	.330	.252	.015	.158	.387
$\hat{\delta} \equiv \hat{H}$	.53	.75	.71	.72	.73	.73
$\hat{\delta} \equiv \hat{\rho}_{xx'}$ (KR-20)	.59	.63	.68	.66	.65	.72
$\hat{\kappa}_{xx'}$	.33	.35	.40	.35	.36	.46

the results represented in Table 2 is the lack of agreement between the estimated values of coefficient delta (H and $\varrho_{xx'}$) and kappa. The latter clearly points to other properties of the decision procedure than the former does; it is conspicuous that the estimated values for kappa are systematically lower. From coefficient kappa a more pessimistic impression of the decision qualities of these tests is obtained than from coefficient delta.

Discussion

In the previous sections two main points were stressed several times. The first is that coefficients for decision consistency necessarily suffer from an improper loss function. It is unclear what weights should be used for differences between decisions across repeated use of the same decision procedure. This is due to the fact that weights defined for inconsistencies in decisions can never represent the loss in fact incurred, since this loss is a function of the discrepancies between the decisions actually made and the decisions that should be made on the basis of the position of the person on the variable Z.

The second point is that decision consistency and reliability are not equivalent concepts. A theorem comparable to Equation 13 does not hold for decisions. Therefore, the body of classical test theory does not apply, and "decision reliability" cannot be determined with procedures for determining measurement reliability. Instead, a theory and procedures should be adopted which take the

peculiarities of decisions, the presence of cutting scores, and their positions on the variables Z and X into account.

According to a decision theoretic approach, not a cross-partitioned $X \times X'$ table like Table 1a, but the decision table, which results from a cross-partitioning of the (X, Z) space, is considered. As far as a psychometric model for measurements is involved, this model is only used for analyzing the bivariate distribution of (X, Z) underlying the decision table. The risk or expected loss is a good starting point for assessing the quality of decisions. Coefficient delta shows on a standardized scale how far the risk is away from its optimum: the optimal value pertains to the situation wherein the test contains complete information about Z. For this reason it may be interpreted as a coefficient for the optimality of the decision procedure, the internal optimality when $Z = T$ and the external optimality when $Z = Y$ is considered.

The finding that for decisions the concepts consistency and reliability are not equivalent has consequences for other research results, mainly in the area of criterion-referenced measurement or mastery testing. Coefficient kappamax was introduced and analyzed by Huynh (1977) as the maximum value of kappa over all possible cutting scores on the test. Although this coefficient may have some meaning for standardizing kappa, it is less appropriate for evaluating decisions than kappa itself. In addition to the cutting score on the variable Z with respect to which the decisions are made, this coefficient even ignores the cutting score c on X. The value of c for which kappa is maximal (this is not the value of c for which the decisions are optimal according to a decision theoretic approach) can differ from the cutting score actually used. Kappamax ought, therefore, to be interpreted as a coefficient for the consistency of decisions which may be other than the ones made by the decision procedure under consideration.

In Emrick's evaluation model for mastery testing, an important step is the substitution of the average interitem correlation for the correlation between mastery state and item response (Emrick & Adams, 1969; Emrick, 1971). This is an application of the theorem in Equation 13 to decisions at the level of an individual item, and the fact that this theorem does not apply to decisions invalidates the model for its purpose.

Analogous to Cronbach's alpha, which can be considered the average correlation between all possible split halves of a test corrected for test length, Marshall and Haertel (1975) introduced a mean split-half coefficient of agreement (see also Marshall, 1975). This coefficient, denoted by beta, is the average of Equation 1 computed over all possible pairs of test halves. The analogy of coefficient beta to Cronbach's alpha is, however, improper. The latter is a lower bound for the reliability coefficient within the framework of classical test theory, of which Equation 13 is a part. For decisions an analogous theorem does not hold, and this makes beta inappropriate as a coefficient for decision reliability.

Summarizing, the decision problem in testing consists of the following elements: an observed score X on a test, a variable Z, the relation between these variables, a decision table with possible outcomes, a loss function, and a coefficient for assessing the quality of the test for making decisions. In applications, specifications are made. The variable Z can be interpreted in different ways. The most important distinction is the one between an observed and an unobserved variable. For an observed variable Z the relation between Z and X is determined empirically. For an unobserved variable Z psychometric theory is necessary for determining the relation between Z and X. The unobserved variable Z has been interpreted as the true score of the observed score; use has been made of classical test theory, the beta-binomial error model, and the beta-compound binomial error model. Another possible interpretation of an unobserved Z is the latent trait that the test is measuring; for this interpretation latent-trait theory can be used for determining the relation between Z and X (Fischer, 1974; Lord & Novick, 1968).

The decision problems considered are dichotomous: one cutting score on both variables and four possible outcomes in the decision table (Table 1b). This can be generalized to multiple-decision problems: $(k-1)$ cutting scores on both variables and k^2 possible outcomes in the decision table (van der Linden & Mellenbergh, 1978). Furthermore, in this paper the threshold loss function and the linear loss function have been considered. It is also possible to investigate other loss functions. Finally, coefficient delta has been used for assessing the quality of a test for making decisions. It is possible that other indices also based on the risk can be constructed for assessing the quality of a test in decision making.

References

Algina, J., & Noe, M.J. A study of the accuracy of Subkoviak's single administration estimate of the coefficient of agreement using two true score estimates. *Journal of Educational Measurement*, 1978, *15*, 101–110.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: The MIT Press, 1976.

Carver, R.P. Special problems in measuring chance with psychometric devices. In *Evaluative research: Strategies and methods*. Pittsburgh: American Institutes for Research, 1970.

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.

Cohen, J. Weighted chi-square: An extension of the kappa method. *Educational and Psychological Measurement*, 1972, *32*, 61–74.

de Bruyne, H.C.D. *Blokken in het onderwijs*. Groningen: Tjeenk Willink, 1976.

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

Emrick, J.A., & Adams, F.N. *An evaluation model for individualized instruction* (Report RC2674). Yorktown Hts., NY: IBM, Thomas J. Watson Research Center, October 1969.

Everitt, B.S. Moments of the statistic kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97–103.

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

Fischer, G.H. *Einführung in die theorie psychologischer tests*. Bern: Verlag Hans Huber, 1974.

Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin, 1971, 76*, 378–382.

Fleiss, J.L., & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, *33*, 613–619.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323–327.

Griffiths, D.A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, 1973, *29*, 637–648.

Gross, A.L., & Su, W.H. Defining a "fair" or "unbaised" selection model: A question of utilities. *Journal of Applied Psychology*, 1975, *60*, 345–351.

Gupta, S.S., Probability integrals of multivariate and normal *t*. *Annals of Mathematical Statistics*, 1963, *34*, 792–828.

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.

Hubert, L. Kappa revisited. *Psychological Bulletin*, 1977, *84*, 289–297.

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

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976, *13*, 253–264. (b)

Huynh, H. *Reliability of multiple classifications*. Paper presented at the spring meeting of the Psychometric Society, Murray Hill, NJ, April 1976. (c)

Huynh, H. *The kappamax reliability index for decisions in domain-referenced testing*. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Jackson, R. *Developing criterion-referenced tests* (TM Report No. 1). Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1970.

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 classification proportions in dichotomous decisions based on dichotomously scored items. *Tijdschrift voor Onderwijsresearch*, 1977, *2*, 32–36.

Landis, J.R., & Koch, G.G. A review of statistical methods in the analysis of data arising from observer reliability studies (Part 1). *Statistica Neerlandica*, 1975, *29*, 101–123. (a)

Landis, J.R., & Koch, G.G. A review of statistical methods in the analysis of data arising from observer reliability studies (Part 2). *Statistica Neerlandica*, 1975, *29*, 151–161. (b)

Light, R.J. Measures of agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 1971, *76*, 365–377.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 1947, *61* (Whole No. 285).

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.

Marshall, J.L. *The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data* (Technical Report 350). Madison: University of Wisconsin, Research and Development Center for Cognitive Learning, 1975.

Marshall, J.L., & Haertel, E.H. *A single-administration reliability index of criterion-referenced test: The mean split-half coefficient of agreement.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1975.

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.

Petersen, N.S. An expected utility model for "optimal" selection. *Journal of Educational Statistics*, 1976, *1*, 333–358.

Spitzer, R.L., Cohen, J., Fleiss, J.L., & Endicott, J. Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry*, 1967, *17*, 83–87.

Subkoviak, M.J. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265–276.

Subkoviak, M.J. Empirical estimation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, *15*, 111–116.

Subkoviak, M.J. *Estimating reliability from a single administration of a mastery test* (Occasional Paper No. 15). Madison: University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design, undated.

Subkoviak, M.J., & Wilcox, R.R. *Estimating the probability of correct classification in mastery testing*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Swaminathan, H., Hambleton, R.K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263–267.

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.

Wingersky, M.S., Lees, D.M., Lennon, V., & Lord, F.M. A computer program for estimating true-score distributions and graduating observed-score distributions. *Educational and Psychological Measurement*, 1969, *29*, 689–692.

Acknowledgments

The authors thank Fred N. Kerlinger, Niels Veldhuyzen, and Michel Zwarts for their comments. The order of the names of the authors is immaterial; they are equally responsible for the content.

Author's Addresses

Send requests for reprints or further information to Gideon J. Mellenbergh, Psychologisch Laboratorium, Universiteit van Amsterdam, Weesperplein 8, Amsterdam, The Netherlands, or Wim J. van der Linden, Vakgroep Onderwijskunde, Technische Hogeschool Twente, Postbus 217, Enschede, The Netherlands.